فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
سال نهم، شماره ۳۳، پاییز ۹۷، ص ۲۳ تا ۵۱

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 9, No. 33, Fall 2018

تعیین روش بهینه شناسایی کارکرد افتراقی در سنجش انطباقی کامپیوتری

نگار شریفی‌یگانه[۱]، محمدرضا فلسفی‌نژاد[۲]، نورعلی فرخی[۳]، احسان جمالی[۴]

تاریخ دریافت: ۹۷/۰۲/۲۶
تاریخ پذیرش: ۹۷/۰۹/۱۵

چکیده

یکی از چالش‌های اساسی گذر از آزمون‌های مداد – کاغذی به انطباقی کامپیوتری عادلانه بودن است که ارزیابی آن در چارچوب کارکرد افتراقی ضرورتی اجتناب‌ناپذیر می‌باشد. هدف مطالعه حاضر بررسی کارکرد افتراقی، ارزیابی عوامل مداخله‌گر در میزان آشکارسازی و معرفی روش بهینه مطالعه کارکرد افتراقی در سنجش انطباقی کامپیوتری بود. با توجه به مسئله پژوهش از روش تجربی استفاده شد. گردآوری داده‌ها و دستکاری متغیرها با استفاده از روش شبیه‌سازی صورت گرفت. پاسخ‌های گروه نمونه ۱۰۰۰ نفری (گروه مرجع و کانونی با حجم یکسان ۵۰۰ نفری) به بانک ۵۵ سؤالی دوارزشی براساس مدل لجستیک سه‌پارامتری در ۲۰ تکرار شبیه‌سازی شد. ۱۵ سؤال بانک از نظر نوع و اندازه کارکرد افتراقی دستکاری شدند و اثر آزمون براساس تفاوت میانگین توانایی گروه‌های مقایسه تعیین گردید. آزمون انطباقی کامپیوتری ۳۰ سؤالی با نرم‌افزار Firestar اجرا شد. تحلیل کارکرد افتراقی با روش رگرسیون لجستیک و آزمون نسبت درستنمایی صورت گرفت و روش‌ها براساس توان و خطای نوع اول مقایسه شدند. میزان خطای نوع اول روش آزمون نسبت درستنمایی کم‌تر از رگرسیون لجستیک بود. توان هر دو روش متاثر از نوع، مقدار کارکرد افتراقی و اثر آزمون بود. روش آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی یکنواخت در هر دو موقعیت اثر و بدون اثر نسبت به روش رگرسیون لجستیک توان

---

۱. دانشجوی دوره دکتری سنجش و اندازه‌گیری دانشگاه علامه‌طباطبائی، تهران، ایران
۲. دانشیار گروه سنجش و اندازه‌گیری دانشگاه علامه‌طباطبائی، تهران، ایران (نویسنده مسئول)
falsafinejad@yahoo.co.uk
۳. دانشیار گروه سنجش و اندازه‌گیری دانشگاه علامه‌طباطبائی، تهران، ایران
۴. استادیار سازمان سنجش آموزش کشور، تهران، ایران

بیشتری داشته است و با افزایش شدت کارکرد افتراقی توان نیز افزایش یافته است. در ارزیابی کارکرد افتراقی غیریکنواخت تفاوتی بین روش‌ها مشاهده نشد و هر دو روش توان کمی داشتند. با توجه به توان و میزان خطای نوع اول، روش آزمون نسبت درستنمایی رویکرد مطلوب در بررسی کارکرد افتراقی یکنواخت است، در حالی که ارزیابی کارکرد افتراقی غیریکنواخت مستلزم مطالعات تکمیلی می‌باشد.

واژگان کلیدی: آزمون انطباقی کامپیوتری، کارکرد افتراقی سؤال ، روش رگرسیون لجستیک، روش آزمون نسبت درستنمایی سؤال – پاسخ

مقدمه

عدالت آموزشی ایجاب می‌کند آزمون‌ها برای تمامی افراد جامعه از هرجنس، نژاد، سن و موقعیت اجتماعی و اقتصادی عادلانه[1] باشد و آزمودنی‌های متعلق به گروه‌های مختلف جامعه با توانایی‌های مشابه، امکان توفیق یکسانی در آزمون داشته باشند. هدف از اجرای آزمون‌ها تفکیک و تمایز افراد است، آزمون‌ها نبایستی منجر به تبعیض میان آزمودنی‌ها شود. دیدگاه‌های متفاوتی در خصوص عادلانه بودن آزمون وجود دارد. تاکنون تعریفی از عادلانه بودن آزمون که مورد توافق همگان باشد، ارائه نشده است. ویلینگهام و کول[2] آزمون عادلانه را به عنوان آزمونی روا و معتبر برای تمام آزمودنی‌ها تعریف کرده‌اند. به عقیده آن‌ها آزمون عادلانه می‌بایست فرصت‌های یکسانی را به منظور ارائه دانش و مهارت‌های مورد سنجش در اختیار تمامی آزمودنی‌ها قرار دهد (ناسستروم[3]،۲۰۰۳).

چالش اساسی عادلانه بودن آزمون و ضرورت بررسی آن به عنوان بخشی از فرایند آزمون‌سازی همزمان با آغاز عصر حقوق مدنی در راستای انتظارات مبنی بر یکسانی نتایج آزمون‌ها شکل گرفت (دانکن[4]، ۲۰۰۶). عادلانه بودن آزمون‌ها در چارچوب سوگیری سؤال‌ها[5] مورد بررسی قرار می‌گیرد. به عقیده انگاف و فورد[6] (۱۹۷۳) سؤال در صورتی دارای سوگیری است که شاخص دشواری آن برای یک گروه در مقایسه با گروه دیگر

1. fairness
2. Willingham & Cole
3. Näsström
4. Duncan
5. Item bias
6. Angoff & Ford

نسبتاً بیشتر یا کم‌تر باشد. هم‌چنین شونمن[1] سؤالی را سودار می‌داند که نسبت پاسخ درست افرادی که دارای توانایی یکسانی در آزمون هستند در هر زیر گروه از آزمودنی‌ها متفاوت باشد (کروکر و آلجینا[2]، ۲۰۰۶، ترجمه فرزاد و زارع، ۱۳۸۸). به طور کلی سوگیری زمانی مطرح می‌شود که احتمال پاسخ‌گویی درست یک گروه از آزمودنی‌ها به دلیل مشخصات سؤال‌ها و شرایط اجرای آزمون از گروه دیگر آزمودنی‌ها کم‌تر یا بیشتر باشد (زومبو[3]، ۱۹۹۹).

تعیین کارکرد افتراقی سؤال روش آماری بررسی وجود سوگیری در سؤال‌های آزمون است. سؤال زمانی دارای کارکرد افتراقی است که آزمودنی‌ها با توانایی یکسان متعلق به گروه‌های مختلف، احتمال پاسخ‌گویی درست متفاوت داشته باشند (راجو[4]، ۱۹۹۰؛ هرریرا و گومز[5]، ۲۰۰۸؛ همبلتون، سوامی‌ناتان و راجرز[6]، ۱۹۹۱، ترجمه فلسفی‌نژاد، ۱۳۸۹). در ادبیات کارکرد افتراقی این گروه‌ها مرجع[7] (گروه اکثریت آزمودنی‌ها که عملکرد آن‌ها به عنوان نقطه مرجع محسوب می‌شود) و کانونی[8] (گروه اقلیت آزمودنی‌ها که انتظار می‌رود سؤال‌های آزمون به نفع آنان نباشد) نامیده می‌شوند (سوامی‌ناتان و راجرز، ۱۹۹۰). وجود کارکرد افتراقی بیانگر آن است که عوامل مربوط به عضویت در گروه، احتمال پاسخ‌گویی درست را تحت تاثیر قرار می‌دهد. در حضور کارکرد افتراقی، وجود سوگیری سؤال باید به روش‌های مختلف مورد بررسی قرار گیرد (دانکن، ۲۰۰۶). به دلیل آن که سوگیری سؤال‌ها باعث ایجاد خطای سیستماتیک می‌شود، تحریف‌هایی در استنباط‌های مبتنی بر نمره آزمون به وجود می‌آورد و تهدیدی برای تفسیر روا نمره‌های آزمون محسوب می‌شود (کونولی[9]، ۲۰۰۳). لذا شناسایی و کنترل سوگیری سؤال‌های آزمون گام مهمی در سنجش علمی می‌باشد و باید به عنوان بخشی از فرایند تحلیل آزمون‌ها در نظر گرفته شود.

1. Scheuneman
2. Crocker & Algina
3. Zumbo
4. Raju
5. Herrera & Gómez
6. Hambleton, Swaminathan & Rogers
7. reference
8. focal
9. Conoley

مطالعه سوگیری و روش‌شناسی آن به طور قابل توجه‌ای تحت‌تاثیر سنجش و شرایط آن است. یکی از عوامل موثر بر سوگیری سؤال‌ها، فرمت اجرای آزمون است. با ظهور و پیشرفت فن‌آوری کامپیوتر و کاربرد آن در سنجش در چارچوب آزمون‌های کامپیوتری و انطباقی کامپیوتری[1] (CAT) مسئله عادلانه بودن این نوع از آزمون‌ها به طور جدی مطرح شد. در آزمون انطباقی کامپیوتری، سؤال‌ها متناسب با توانایی آزمودنی‌ها در جریان آزمون انتخاب و اجرا می‌شوند (ریکیسی[2]، ۱۹۸۹، به نقل از دیویدسون[3]، ۲۰۰۳). آزمون‌های کامپیوتری خصوصاً انطباقی کامپیوتری با توجه به توانمندی بالقوه‌ای که در بهینه‌سازی فرایند سنجش دارند، جایگزین جدی آزمون‌های مداد- کاغذی قلمداد می‌شوند. بسیاری از سازمان‌ها و مؤسسات آزمون‌سازی بین‌المللی نظام سنجش و ارزشیابی خود را از فرمت مداد – کاغذی به کامپیوتری و انطباقی کامپیوتری تغییر داده‌اند یا در حال مطالعه و برنامه‌ریزی جهت انجام این تغییر هستند. آزمون تحصیلات تکمیلی[4](GRE)، آزمون پذیرش تحصیلات تکمیلی دوره مدیریت[5](GMAT)، آزمون مجوز پرستاری[6] (NCLEX)، مجموعه آزمون‌های استعداد شغلی ارتش[7] و آزمون تافل[8] از جمله آزمون‌های خطیری هستند که به صورت انطباقی کامپیوتری اجرا می‌شوند (کلندر[9]، ۲۰۱۱).

این در حالی است که گذر از آزمون‌های مداد- کاغذی به کامپیوتری و انطباقی کامپیوتری فرایندی پرمخاطره با چالش‌های نظری و عملی فراوان است. عادلانه بودن آزمون‌ها یکی از مهم‌ترین چالش‌های فرایند تغییر روش اجرای آزمون‌ها است که ضرورت ارزیابی آن در چارچوب کارکرد افتراقی سؤال‌ها مورد تاکید قرار گرفته است (انجمن تحقیقات آموزشی آمریکا[10]، انجمن روان‌شناسی آمریکا[11] و انجمن ملی سنجش

1. Computerized Adaptive Testing
2. Reckase
3. Davidson
4. Graduate Record Examination
5. Graduate Management Admission Test
6. Nursing Licensing Examination (NCLEX)
7. Armed Services Vocational Aptitude Battery (ASVAB)
8. TOEFL
9. Kalender
10. American Educational Research Association (AERA)
11. American Psychological Association (APA)

آموزشی[1]، ۱۹۹۹). اهمیت بررسی کارکرد افتراقی در آزمون‌های انطباقی کامپیوتری به دلیل وجود چندین منبع بالقوه کارکرد افتراقی و هم‌چنین کم بودن تعداد سؤال‌ها دوچندان می‌شود. کم بودن تعداد سؤال‌ها در آزمون انطباقی کامپیوتری باعث می‌شود پاسخ به هر سؤال نقش مهم‌تری در نمره آزمودنی داشته باشد و هرگونه عیب و کاستی در سؤال‌ها عواقب جدی‌تری برای آزمودنی به همراه داشته باشد.

اگرچه روش‌های مطالعه کارکرد افتراقی سؤال ادبیات نسبتاً وسیعی را در متون سنجش به خود اختصاص داده است اما روش‌های ارائه شده بیشتر در آزمون‌های مداد – کاغذی که به لحاظ اجرایی دارای فرم بسته هستند کاربرد دارد. بررسی کارکرد افتراقی سؤال‌ها در آزمون‌های انطباقی کامپیوتری با چالش دشواری تعیین متغیر تطبیق[2] به منظور جور کردن آزمودنی‌های گروه‌های مقایسه همراه است. در آزمون‌های انطباقی کامپیوتری تمامی آزمودنی‌ها به سؤال‌های یکسانی پاسخ نمی‌دهند، این مسئله منجر به نمره‌های کل متفاوتی می‌شود. بنابراین کاربرد نمره کل به عنوان متغیر تطبیق در سنجش انطباقی کامپیوتری بی‌معنا است (پیرومسومبات[3]، ۲۰۱۴؛ زویک[4]، ۲۰۱۰). به علاوه در سنجش انطباقی کامپیوتری ارزیابی کارکرد افتراقی مستلزم انتخاب روشی است که در نمونه‌های کوچک نتایج پایداری ارائه کند. اگر چه تعداد کل آزمودنی‌ها در آزمون انطباقی کامپیوتری زیاد است، تعداد پاسخ‌ها برای برخی از سؤال‌های آزمون ممکن است خیلی کم باشد (زویک، ۲۰۱۰). بر این اساس شناسایی کارکرد افتراقی سؤال در آزمون‌های انطباقی کامپیوتری مستلزم به‌کارگیری روش‌های جدید و مخصوص می‌باشد (پیرومسومبات، ۲۰۱۴؛ زویک، ۲۰۱۰).

رویکردهای فعلی تعیین کارکرد افتراقی در سنجش انطباقی کامپیوتری مبتنی بر روش‌های به کار رفته در آزمون‌های مداد – کاغذی است. به عنوان مثال زویک و همکارانش (زویک، تایر و وینگرسکی[5]، ۱۹۹۴ و زویک، تایر و لویس[6]، ۱۹۹۷) روش‌های زویک، تایر و وینگرسکی(ZTW) و زویک، تایر و لویس (ZTL) را براساس آماره

1. National Council on Measurement in Education (NCME
2. matching variable
3. Piromsombat
4. Zwick
5. Thayer & Wingersky
6. Lewis

مانتل – هنزل[1] (هالند[2] و تایر، ۱۹۸۸) و روش استانداردسازی[3] (دورانس و کولیک[4]، ۱۹۸۶) مطرح کردند. در این روش‌ها آزمودنی‌های گروه مرجع و کانونی براساس نمره واقعی مبتنی بر برآورد توانایی آزمون انطباقی کامپیوتری جور می‌شوند. تفاوت این دو روش مربوط به استفاده از برآورد بیزین تجربی[5] در روش ZTL برای ارائه نتایج با ثبات‌تر در نمونه‌های کوچک می‌باشد (پیرومسومبات، ۲۰۱۴؛ زویک، ۲۰۱۰). هم‌چنین ناناداکومار و روسوس[6] (۲۰۰۱) روش SIBTEST (شیلی و استوت[7]، ۱۹۹۳) را برای سنجش انطباقی کامپیوتری توسعه بخشیدند که منجر به ایجاد روش CATSIB شد. در روش CATSIB آزمودنی‌ها براساس برآورد توانایی آزمون انطباقی کامپیوتری جور می‌شوند، سپس روش SIBTEST برای ارزیابی کارکرد افتراقی به کار می‌رود. روش‌های ZTW، ZTL و CATSIB صرفاً برای شناسایی کارکرد افتراقی یکنواخت[8] (هماهنگ) طراحی شده‌اند. لی، چن و یو[9] (۲۰۰۶) نسخه اصلاح شده رگرسیون لجستیک[10] و آزمون نسبت درستنمایی سؤال– پاسخ[11] را به منظور بررسی کارکرد افتراقی غیریکنواخت[12] (غیرهماهنگ) در آزمون انطباقی کامپیوتری در سؤال‌های پیش‌آزمون معرفی کردند. در کارکرد افتراقی یکنواخت، تفاوت احتمال پاسخ‌گویی درست گروه کانونی و مرجع در تمام سطوح توانایی یکسان است و تعامل بین سطح توانایی و عضویت در گروه وجود ندارد. در حالی که در کارکرد افتراقی غیریکنواخت احتمال پاسخ‌گویی درست گروه کانونی و مرجع در تمام سطوح توانایی یکسان نمی‌باشد و بین سطح توانایی و عضویت در گروه تعامل وجود دارد (دریانا[13]، ۲۰۰۷).

1. Mantel - Haenszel
2. Holland & Thayer
3. Standardization
4. Dorans & Kulick
5. Empirical bayesian estimation
6. Nandakumar & Roussos
7. Shealy & Stout
8. uniform
9. Lei, Chen & Yu
10. Logistic Regression (LR)
11. Item Response Theory-Likelihood Ratio Tests (IRT-LRT)
12. non uniform
13. Driana

از آنجایی که بیشتر الگوریتم‌های آزمون انطباقی کامپیوتری مستلزم برآورد مبتنی بر نظریه سؤال – پاسخ است به نظر می‌رسد روش نسبت درستنمایی انتخاب منطقی برای تحلیل کارکرد افتراقی در چارچوب آزمون‌های انطباقی کامپیوتری باشد (مایلر[1]، ۱۹۹۲). کاربرد آزمون نسبت درستنمایی سؤال- پاسخ در ارزیابی کارکرد افتراقی توسط تیسن، استینبرگ و واینر[2] (۱۹۹۳) مطرح شد. در این روش تفاوت پارامتر سؤال‌ها بین گروه‌ها مشروط بر نامتغیر بودن سایر سؤال‌های آزمون مدل‌سازی می‌شود. برازش مدل افزایشی[3] و مدل فشرده[4] براساس تفاوت لگاریتم درستنمایی دو مدل مورد مقایسه قرار می‌گیرد. در مدل افزایشی پارامترهای سؤال‌های مورد بررسی امکان تغییر در گروه‌های مورد مقایسه را دارند، در حالی که در مدل فشرده پارامترهای سؤال محدود می‌شوند (کوهن، کیم و ولاک[5]، ۱۹۹۶؛ لی، چن و یو، ۲۰۰۶). اگر تفاوت لگاریتم درستنمایی دو مدل که دارای توزیع مجذور کای است معنادار باشد، آزمون‌های بعدی به منظور بررسی تفاوت‌های پارامتر سؤال‌ها انجام می‌شود (پیرومسومبات، ۲۰۱۴).

یک مزیت روش نسبت درستنمایی امکان کاربرد مستقیم آن در آزمون انطباقی کامپیوتری است. روش نسبت درستنمایی برخلاف سایر روش‌ها مبتنی بر نمره کل آزمون نیست (کیم، ۲۰۰۶). با وجود مطلوبیت‌های نظری این روش کاربرد آن به دلیل پیچیدگی‌های ناشی از داده‌های ناقص حاصل از اجرای انطباقی آزمون‌ها، با مشکلاتی مواجه شد. هم‌چنین برازش مدل‌های متعدد و محاسبات سنگین مورد نیاز روش نسبت درستنمایی استفاده از آن را خصوصاً در مطالعات شبیه‌سازی غیرعملی ساخت. تیسن (۲۰۰۱) نرم‌افزار IRTLRDIF را به منظور افزایش کاربرد روش نسبت درستنمایی در تحلیل کارکرد افتراقی آزمون‌های انطباقی کامپیوتری ارائه کرد. این نرم‌افزار با تسریع انجام محاسبات امکان استفاده از روش نسبت درستنمایی را در مطالعات شبیه‌سازی ممکن ساخته است (لی، چن و یو، ۲۰۰۶).

1. Miller
2. Thissen, Steinberg & Wainer
3. augmented
4. compact
5. Cohen, Kim & Wollack

روش رگرسیون لجستیک نیز با توجه به قابلیت شناسایی کارکرد افتراقی غیریکنواخت می‌تواند در چارچوب آزمون انطباقی کامپیوتری مورد استفاده قرار گیرد. رگرسیون لجستیک مبتنی بر مدل‌سازی آماری احتمال پاسخ صحیح به سؤال براساس عضویت در گروه و ملاک است. در این روش پاسخ سؤال‌ها به عنوان متغیر وابسته در نظر گرفته می‌شود، متغیر گروهی به عنوان متغیری که قصد بررسی پاسخ‌ها را در آن دارند به عنوان گروه‌های مرجع و کانونی در نظر گرفته می‌شود. نمره کل آزمون نیز به عنوان ملاک به کار می‌رود. در روش رگرسیون لجستیک وجود کارکرد افتراقی از طریق بررسی بهبود ایجاد شده در برازش مدل رگرسیون پس از اضافه کردن عضویت در گروه و تعامل بین نمره آزمون و عضویت در گروه تعیین می‌شود. تعامل به منظور تعیین کارکرد افتراقی غیریکنواخت استفاده می‌شود (ملیسپ[1]، ۲۰۱۱ به نقل از گرامی‌پور،۱۹۹۰). در نسخه اصلاح شده رگرسیون لجستیک به منظور بررسی کارکرد افتراقی در آزمون‌های انطباقی کامپیوتری، برآورد توانایی آزمون انطباقی کامپیوتری در معادله رگرسیون لجستیک جایگزین می‌شود (پیرومسومبات، ۲۰۱۴).

به طور کلی عملکرد روش‌های شناسایی کارکرد افتراقی سؤال‌ها مشابه نمی‌باشند و تحت‌تاثیر عوامل مختلفی از جمله شدت کارکرد افتراقی، توزیع توانایی گروه‌های مقایسه، طول آزمون، حجم نمونه و نسبت سؤال‌های دارای کارکرد افتراقی می‌باشند (کیم ، ۲۰۱۰؛ کاباساکال، گوک، آرسان و کلیسی‌گل ، ۲۰۱۴). کارایی روش‌های شناسایی کارکرد افتراقی در سنجش انطباقی کامپیوتری و عوامل موثر بر آن در مطالعات معدودی مورد بررسی قرار گرفته است. به عنوان مثال زویک، تایر و وینگرسکی (۱۹۹۴) ویژگی‌های روش ZTW را با شبیه‌سازی پاسخ‌ها به سه بانک ۷۵ سؤالی (بانک اول سؤال‌ها فاقد کارکرد افتراقی، بانک دوم سؤال‌ها دارای کارکرد افتراقی ناهمبسته با دشواری سؤال و بانک سوم سؤال‌ها دارای کارکرد افتراقی همبسته با دشواری) ارزشیابی کردند. به هر آزمودنی ۲۵ سؤال از هر یک از بانک‌ها ارائه شد. آزمودنی‌های شبیه‌سازی شده گروه‌های مرجع و کانونی براساس نمره واقعی مورد انتظار برای کل بانک جور شدند (زویک، ۲۰۱۰). حجم گروه‌های مرجع و کانونی، توزیع توانایی گروه‌ها و الگوی کارکرد افتراقی متغیرهای بودند که دستکاری شدند. عملکرد روش ZTW با آماره مانتل هنزل مرسوم

1. Millsap

مورد مقایسه قرار گرفت. نتایج بیانگر آن بود که برآوردهای کارکرد افتراقی روش ZTWهمبستگی بالایی با مقدار کارکرد افتراقی واقعی دارد. به علاوه میزان خطای نوع اول روش ZTW در اکثر موقعیت‌های شبیه‌سازی شده قابل قبول بود (پیرومسومبات، ۲۰۱۴؛ زویک، ۲۰۱۰، زویک، تایر و وینگرسکی،۱۹۹۴). زویک و تایر (۲۰۰۲) عملکرد روش ZTL را با شبیه‌سازی داده‌ها با استفاده از بانک سؤال، حجم نمونه، طول آزمون، اندازه کارکرد افتراقی یکنواخت و اندازه اثر آزمون (تفاوت میانگین توانایی گروه‌ها) مورد بررسی قرار دادند. نتایج بیانگر آن بود که روش ZTL در شناسایی کارکرد افتراقی یکنواخت در سؤال‌های پیش‌آزمون خصوصاً در نمونه‌های کوچک موثر است .

نانداکومار و روسوس (۲۰۰۴؛۲۰۰۱) عملکرد روش CATSIB را در شناسایی کارکرد افتراقی در سؤال‌های پیش‌آزمون با استفاده از روش شبیه‌سازی بررسی کردند. نتایج بیانگر رضایت‌بخش بودن میزان خطای نوع اول CATSIB با تصحیح رگرسیون بود. توان این روش دامنه گسترده‌ای (۰/۱۷ تا ۱) داشت و با افزایش حجم گروه نمونه افزایش یافته بود. حساسیت روش CATSIB در شرایط کوچک بودن حجم هر یک از گروه‌ها، نابرابر بودن حجم گروه‌ها و زیاد بودن اندازه اثر آزمون (تفاوت میانگین توزیع توانایی گروه مرجع و کانونی) کم بود

لی، چن و یو (۲۰۰۶) نیز عملکرد CATSIB، رگرسیون لجستیک و آزمون نسبت درستنمایی سؤال – پاسخ را به طور همزمان با استفاده از مطالعه شبیه‌سازی در سؤال‌های پیش آزمون مورد بررسی قرار دادند. داده‌ها تحت شرایط نسبت حجم نمونه و بزرگی اثر آزمون شبیه‌سازی شدند. نتایج مطالعه نشان داد که روش CATSIB، رگرسیون لجستیک و آزمون نسبت درستنمایی از نظر توان شناسایی کارکرد افتراقی سئوال یکنواخت قابل مقایسه هستند در حالی که در شناسایی کارکرد افتراقی غیریکنواخت روش‌های رگرسیون لجستیک و آزمون نسبت درستنمایی نسبت به CATSIB توانمندتر هستند. آزمون نسبت درستنمایی کنترل خوبی بر میزان خطای نوع اول در شرایط حجم نمونه نابرابر و اثر آزمون داشته است.

با توجه به مطالب مطرح شده می‌توان نتیجه گرفت که با وجود اهمیت عادلانه بودن آزمون‌های انطباقی کامپیوتری و ضرورت ارزیابی آن در چارچوب کارکرد افتراقی، مطالعات کمی در این حوزه انجام شده است. به علاوه در تمامی مطالعات، کارکرد افتراقی

در سؤال‌های پیش‌آزمون با جور کردن آزمودنی‌ها براساس توانایی مبتنی بر سنجش انطباقی کامپیوتری مورد بررسی قرار گرفته است. سؤال‌های پیش‌آزمون به صورت غیرانطباقی اجرا می‌شوند و بنابراین نمی‌تواند بیانگر شرایط و موقعیت کارکرد افتراقی در بافت آزمون‌های انطباقی کامپیوتری باشند. کم‌تر مطالعه‌ای به طور خاص کارکرد افتراقی را در سؤال‌های عملیاتی در جریان آزمون انطباقی کامپیوتری مورد بررسی قرار داده است. از این رو اهمیت مطالعه در این حوزه کاملاً مشهود است.

در جامعه ما نیز ضرورت بهینه‌سازی آزمون‌ها و کاربرد فن‌آوری کامپیوتر در سنجش با توجه به حجم بالای کاربرد آزمون‌ها در موقعیت‌های مختلف به منظور اخذ تصمیمات سرنوشت‌ساز کاملاً احساس می‌شود. آزمون‌های انطباقی کامپیوتری از جمله حوزه‌هایی آزمون‌سازی محسوب می‌شوند که با خلاء مطالعات نظامند روبه رو هستند. از این رو انجام مطالعه در خصوص چالش‌های تغییر روش اجرای آزمون‌ها از فرمت مداد- کاغذی به کامپیوتری و خصوصاً انطباقی کامپیوتری و بررسی راه‌حل‌های ممکن پیش از هر گونه تصمیم‌گیری برای تغییر فرمت آزمون‌ها بسیار حائز اهمیت است. در راستای این هدف، پژوهش حاضر به عنوان یکی از اولین مطالعات به بررسی کارکرد افتراقی در سؤال‌های عملیاتی در اجرای انطباقی کامپیوتری براساس روش آزمون نسبت درستنمایی سؤال – پاسخ و رگرسیون لجستیک پرداخته است. این دو روش به دلیل توانمندی در شناسایی کارکرد افتراقی یکنواخت و غیریکنواخت و هم‌چنین ویژگی‌های نظری و عملی که پیشتر به آن اشاره شد انتخاب شدند.

روش

به منظور ارزیابی کارکرد افتراقی سؤال‌های عملیاتی در اجرای انطباقی کامپیوتری و مقایسه دو روش رگرسیون لجستیک و آزمون نسبت درستنمایی از روش‌های شبیه‌سازی مونت‌کارلو[1] و پس‌تجربی[2] (واقعی) برای ایجاد داده‌هایی مشابه با داده‌های واقعی، دستکاری متغیرهای مورد بررسی و اجرای آزمون انطباقی کامپیوتری استفاده شد. براین اساس پژوهش حاضر از نظر روش، تحقیق تجربی است.کاربرد روش‌های شبیه‌سازی در

1. Monte Carlo Simulations
2. Post-hoc Simulations

تحقیقات سنجش آموزشی بسیار متداول شده است و نتایج آن‌ها معمولاً به عنوان شواهد روایی برای تصمیم‌گیری درباره برنامه‌های سنجش در نظر گرفته می‌شوند. سنجش انطباقی کامپیوتری یکی از حوزه‌هایی است که به طور فزاینده‌ای با استفاده از روش‌های شبیه‌سازی مورد بررسی قرار می‌گیرد (هارول، استون، هسو و کرسکی[1]، ۱۹۹۶).

داده‌ها با استفاده از روش شبیه‌سازی مونت‌کارلو تولید شدند. روش کار به این صورت است که پارامتر توانایی آزمودنی‌ها و پارامترهای سؤال‌ها براساس توزیع‌های تعیین شده و با توجه به مدل سؤال – پاسخ انتخابی ایجاد شدند. در ادامه ماتریس پاسخ سؤال‌ها برای هر ترکیب سؤال و آزمودنی با مقایسه احتمال پاسخ‌گویی درست به سؤال با اعداد تصادفی استخراج شده از توزیع یکنواخت تشکیل شد. در صورتی که احتمال پاسخ‌گویی محاسبه شده بزرگ‌تر از عدد تصادفی بود نمره ۱ و در صورتی که کوچک‌تر از عدد تصادفی بود نمره ۰ در نظر گرفته شد. (نیدیک و ویس ، ۲۰۰۹).

بر این اساس در مطالعه حاضر پاسخ به سؤال‌های بانک ۵۵ سؤالی دوارزشی براساس مدل لجستیک سه پارامتری با استفاده از برنامه WinGen (هان[2]، ۲۰۰۷) در ۱۸ شرایط آزمایشی مختلف با توجه به عوامل مداخله‌گر شبیه‌سازی شد. ویژگی‌های ۱۸ مجموعه داده حاصل در جدول شماره ۱ ارائه شده است. منطبق بر مطالعات صورت گرفته، پارامتر توانایی آزمودنی‌ها از توزیع نرمال با میانگین ۰ و انحراف استاندارد ۱ (۱، ۰)N ، پارامتر دشواری سؤال‌ها از توزیع یکنواخت (۳، ۳-)، پارامتر تشخیص از توزیع لگ نرمال با میانگین ۰/۱- و انحراف استاندارد ۰/۱ (۰/۱، ۰/۱-)LnN و پارامتر حدس از توزیع یکنواخت (۰،۰/۲۵) انتخاب شدند. توزیع پارامتر دشواری سؤال‌ها به منظور پوشش دامنه وسیعی از سطوح توانایی واقعی انتخاب شده است و منجر به تابع آگاهی قابل مقایسه‌ای در سطوح توانایی می‌شود. این توزیع‌ها دامنه وسیعی از پارامترهای سؤال‌ها را که در عمل و در شبیه‌سازی‌های قبلی آزمون انطباقی کامپیوتری به کار رفته‌اند را تولید می‌کنند

1. Harwell, Stone, Hsu & Kirisci
2. Han

(پیرومسومبات، ۲۰۱۴؛ چانگ و ینگ[1]، ۱۹۹۶؛ زویک، ۱۹۹۴؛ گیور[2]، ۲۰۰۸؛ وانگ و ویسپول[3]، ۱۹۹۸).

۱۵ سؤال بانک از نظر نوع کارکرد افتراقی در سه سطح بدون کارکرد افتراقی، دارای کارکرد افتراقی یکنواخت و کارکرد افتراقی غیریکنواخت دستکاری شدند. در کارکرد افتراقی یکنواخت سؤال‌ها از لحاظ سطح دشواری در دو گروه متفاوت بودند، اما از نظر قدرت تشخیص تفاوتی نداشتند. در کارکرد افتراقی غیریکنواخت سؤال‌ها صرفاً از نظر قدرت تشخیص در دو گروه متفاوت بودند. علاوه بر نوع کارکرد افتراقی، اندازه کارکرد افتراقی و اثر آزمون نیز دستکاری شد. اندازه کارکرد افتراقی در چهار سطح ۰/۴، ۰/۶، ۰/۸ و ۱ دستکاری شد. اثر آزمون نیز به صورت تفاوت میانگین توانایی گروه‌ها تعریف می‌شود. در شرایط داده‌های واقعی یافتن گروه‌های مرجع و کانونی با توزیع توانایی یکسان دشوار می‌باشد. تفاوت توزیع توانایی گروه‌ها در شناسایی کارکرد افتراقی سؤال‌ها اثر گذار است (کاباساکال و همکاران، ۲۰۱۴). وقتی که اثر آزمون وجود ندارد هر دو گروه مرجع و کانونی از جامعه‌ای با توزیع نرمال استاندارد با میانگین ۰ و انحراف استاندارد ۱ (۱،۰) N نمونه‌برداری شدند. هنگامی که اثر آزمون وجود دارد فرض می‌شود میانگین جامعه گروه کانونی یک انحراف استاندارد زیر میانگین جامعه گروه مرجع (۱،۱-)N است.

به منظور کاهش سوگیری بالقوه چندین مجموعه داده شبیه‌سازی شد. تعداد تکرارها با توجه به هدف مطالعه، میزان کاهش مورد نظر در واریانس نمونه‌گیری پارامتر و توان مورد نیاز آزمون‌های آماری انتخاب می‌شود. در مطالعات مقایسه روش‌شناسی رویکرد سؤال – پاسخ به عنوان مثال بررسی کارکرد افتراقی سؤال‌ها تعداد تکرارهای کم‌تری لازم است. هارول و همکاران (۱۹۹۶) حداقل ۱۰ تکرار را در این موارد توصیه کرده‌اند. براین اساس در مطالعه حاضر برای کسب اطمینان از حداقل بودن سوگیری نمونه و به منظور دستیابی به توان آماری مطلوب ۲۰ تکرار برای یک از شرایط آزمایشی صورت گرفت.

1.Chang & Ying
2. Guyer
3. Wang & Vispoel

جدول ۱. ویژگی‌های شرایط آزمایشی

شرایط	نوع کارکرد افتراقی	مقدار کارکرد افتراقی	توزیع گروه کانونی
۱	بدون کارکرد افتراقی	–	N(۰,۱)
۲	بدون کارکرد افتراقی	–	N(-۱,۱)
۳	کارکرد افتراقی یکنواخت	۰/۴	N(۰,۱)
۴	کارکرد افتراقی یکنواخت	۰/۶	N(۰,۱)
۵	کارکرد افتراقی یکنواخت	۰/۸	N(۰,۱)
۶	کارکرد افتراقی یکنواخت	۱	N(۰,۱)
۷	کارکرد افتراقی یکنواخت	۰/۴	N(-۱,۱)
۸	کارکرد افتراقی یکنواخت	۰/۶	N(-۱,۱)
۹	کارکرد افتراقی یکنواخت	۰/۸	N(-۱,۱)
۱۰	کارکرد افتراقی یکنواخت	۱	N(-۱,۱)
۱۱	کارکرد افتراقی غیر یکنواخت	۰/۴	N(۰,۱)
۱۲	کارکرد افتراقی غیر یکنواخت	۰/۶	N(۰,۱)
۱۳	کارکرد افتراقی غیر یکنواخت	۰/۸	N(۰,۱)
۱۴	کارکرد افتراقی غیر یکنواخت	۱	N(۰,۱)
۱۵	کارکرد افتراقی غیر یکنواخت	۰/۴	N(-۱,۱)
۱۶	کارکرد افتراقی غیر یکنواخت	۰/۶	N(-۱,۱)
۱۷	کارکرد افتراقی غیر یکنواخت	۰/۸	N(-۱,۱)
۱۸	کارکرد افتراقی غیر یکنواخت	۱	N(-۱,۱)

شبیه‌سازی پس‌تجربی آزمون‌های انطباقی کامپیوتری ۳۰ سؤالی (با طول ثابت) براساس بانک سؤال تولید شده و پارامتر توانایی آزمودنی‌ها با استفاده از نرم‌افزار) Firestar چویی و پودرابسکی و مکینی ، ۲۰۱۱) انجام شد. طول آزمون مشابه با آزمون‌های به کار رفته در مطالعات قبلی که در دامنه ۱۵ تا ۳۰ سؤال بودند، انتخاب شده است (زویک، تایر و وینگرسکی، ۱۹۹۴؛ لی، چن و یو، ۲۰۰۶؛ نامداکومار و روسوس، ۲۰۰۴). فرایند شبیه‌سازی به این ترتیب است که براساس قاعده شروع و ملاک گزینش تعیین شده، سؤالی از بانک انتخاب و به آزمودنی ارائه می‌شود. سپس پاسخ آزمودنی به سؤال که بر اساس پارامتر توانایی و سؤال‌ها تولید شده است مورد بررسی قرار می‌گیرد و با توجه به پاسخ و روش برآورد توانایی تعیین شده توانایی آزمودنی برآورد می‌شود. در ادامه براساس پاسخ آزمودنی و ملاک گزینش سؤال بعدی از بانک انتخاب می‌شود و مجدداً پاسخ جدید مورد بررسی قرار می‌گیرد و برآورد توانایی آزمودنی اصلاح می‌شود. این فرایند تا زمان

تحقق ملاک‌های خاتمه آزمون ادامه می‌یابد. در این مرحله آزمون خاتمه می‌یابد و برآورد نهایی توانایی آزمودنی ارائه می‌شود. ۷۲۰ شبیه‌سازی آزمون انطباقی کامپیوتری انجام شد. برای تمام آزمودنی‌ها نقطه شروع آزمون توانایی صفر در نظر گرفته شد و گزینش سؤال‌ها براساس بیشینه آگاهی انجام گرفت. در این مرحله توانایی آزمودنیها با استفاده از روش بیزین پسین مورد انتظار (EAP) برآورد شد.

از آن‌جایی که در آزمون‌های انطباقی کامپیوتری هر آزمودنی به مجموعه‌ای متفاوت از سؤال‌ها به تناسب توانایی خود پاسخ می‌دهد، پاسخ‌های جمع‌آوری شده هر آزمودنی بخش کوچکی از داده‌هایی است که در آزمون غیرانطباقی متناظر به دست می‌آید. به عبارت دیگر ماتریس دادهها ناکامل میباشد. پیش از تحلیل کارکرد افتراقی سؤالها میبایست ماتریس دادهها کامل شود. تکمیل ماتریس داده‌ها به روش جایگذاری براساس احتمال پاسخ‌گویی درست به سؤال‌ها مبتنی بر برآورد توانایی آزمون انطباقی کامپیوتری برای هر آزمودنی صورت گرفت. ابتدا با توجه به این که هر یک از آزمودنی‌ها به مجموعه سؤال‌های متفاوتی در شبیه‌سازی آزمون انطباقی کامپیوتری پاسخ داده‌اند، ۱۴۴۰ فایل"سؤال‌های ارائه شده" و "پاسخ سؤال‌ها" (۱۸ شرایط مطالعه 2 گروه مرجع و کانونی 20 تکرار 2 مجموعه فایل) با استفاده از سامانه مدیریت پایگاه داده‌ها متن‌باز MySQLو برنامه‌نویسی PHP مرتب شد و به این ترتیب ماتریس اولیه دادهها تشکیل شد. در ادامه احتمال پاسخ‌گویی صحیح به سؤال‌ها براساس مدل لجستیک سه‌پارامتری با استفاده از پارامتر سؤال‌ها و برآوردهای توانایی آزمون انطباقی کامپیوتری برای هر مجموعه آزمودنیهای شبیهسازی شده محاسبه شد. سپس مقادیر احتمال پاسخ‌گویی به دست آمده با اعداد تصادفی از توزیع یکنواخت (۰ و ۱ U(مورد مقایسه قرار گرفت. در صورتی که مقادیر احتمال محاسبه شده مساوی و یا بزرگ‌تر از عدد تصادفی بود، پاسخ ۱ و در صورتی که کوچک‌تر بود پاسخ ۰ لحاظ شد و به این ترتیب ماتریس داده‌ها جایگذاری و کامل شد.

پس از تشکیل ماتریس کامل داده‌ها کارکرد افتراقی سؤال‌های دستکاری شده با استفاده از روش رگرسیون لجستیک و آزمون نسبت درستنمایی سؤال – پاسخ تحلیل شدند. در روش رگرسیون لجستیک نمره واقعی مبتنی بر برآورد توانایی آزمون انطباقی کامپیوتری به عنوان متغیر تطبیق مورد استفاده قرار گرفت. در روش آزمون نسبت

درستنمایی نیز سایر سؤال‌های بانک به عنوان سؤال‌های لنگر در نظر گرفته شدند. در مطالعه حاضر مقایسه دو رویکرد رگرسیون لجستیک و آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی براساس توان (نسبت مثبت واقعی) و خطای نوع اول (نسبت مثبت کاذب) در ۲۰ تکرار برای هر موقعیت صورت گرفت. در صورتی که سؤال دارای کارکرد افتراقی باشد و روشهای شناسایی کارکرد افتراقی به درستی کارکرد افتراقی را مشخص سازند، در این حالت کارکرد افتراقی مثبت واقعی تشخیص داده شده است. حال در صورتی که سؤال فاقد کارکرد افتراقی باشد ولی روشهای شناسایی کارکرد افتراقی به اشتباه کارکرد افتراقی را در سؤال مشخص سازند در این حالت کارکرد افتراقی منفی کاذب تشخیص داده شده است.

یافته‌ها

توان آماری کارکرد افتراقی یکنواخت (نسبت مثبت واقعی): توان آماری شناسایی کارکرد افتراقی یکنواخت (نسبت مثبت واقعی) دو رویکرد رگرسیون لجستیک و آزمون نسبت درستنمایی در ۲۰ تکرار برای سطوح مختلف کارکرد افتراقی و تحت شرایط اثر و بدون اثر آزمون در جداول شماره ۲ و ۳ ارائه شده است. هر دو روش رگرسیون لجستیک و آزمون نسبت درستنمایی هنگامی که شدت کارکرد افتراقی بیشتر است و توزیع توانایی گروه مرجع و کانونی یکسان است (بدون اثر آزمون) توان بیشتری در شناسایی سؤال‌های دارای کارکرد افتراقی یکنواخت دارند. با افزایش شدت کارکرد افتراقی توان روش رگرسیون لجستیک و آزمون نسبت درستنمایی در هر دو موقعیت اثر و بدون اثر آزمون افزایش یافته است.

مقادیر توان مساوی و بزرگ‌تر از ۰/۸ بیانگر توانمندی روش در شناسایی کارکرد افتراقی است (کاباساکال و همکاران، ۲۰۱۴؛ لی، چن و یو، ۲۰۰۶). در روش رگرسیون لجستیک در شرایط بدون اثر هنگامی که اندازه کارکرد افتراقی ۰/۴ (کم) است نسبت مثبت واقعی (توان) برای هیچ یک از سؤال‌های آزمون مساوی و بزرگ‌تر از ۰/۸ نیست. هنگامی که اندازه کارکرد افتراقی برابر یک است نسبت مثبت واقعی ۷ سؤال مساوی و بزرگ‌تر از ۰/۸ است. این نتایج بیانگر عدم کفایت روش رگرسیون لجستیک در شرایط کارکرد افتراقی کم می‌باشد. در روش نسبت درستنمایی در شرایط بدون اثر هنگامی که

اندازه کارکرد افتراقی ۰/۴ است نسبت مثبت واقعی یک سؤال مساوی و بزرگ‌تر از ۰/۸ است. حال آن‌که با افزایش شدت کارکرد افتراقی تعداد سؤال‌هایی که نسبت مثبت واقعی مساوی و بزرگ‌تر از ۰/۸ دارند افزایش می‌یابد.

جدول ۲. توان کارکرد افتراقی یکنواخت شرایط بدون اثر

	کارکرد افتراقی یکنواخت							
سؤال	رگرسیون لجستیک				آزمون نسبت درستنمایی			
	۰/۴	۰/۶	۰/۸	۱	۰/۴	۰/۶	۰/۸	۱
۴۱	۰/۴	۰/۷۵	۰/۸۵	۱	۰/۴۵	۰/۹	۱	۰/۹۵
۴۲	۰/۴۵	۰/۷	۰/۹	۰/۹۵	۰/۶	۰/۸	۰/۹۵	۱
۴۳	۰/۳۵	۰/۴	۰/۷۵	۰/۸۵	۰/۳۵	۰/۷۵	۰/۸۵	۰/۸۵
۴۴	۰/۰۵	۰/۱	۰/۱	۰/۱۵	۰/۰۵	۰/۱	۰/۱	۰/۲
۴۵	۰/۲	۰/۲۵	۰/۴	۰/۵	۰/۲۵	۰/۵	۰/۶۵	۰/۸
۴۶	۰/۴۵	۰/۹	۰/۹۵	۱	۰/۷	۰/۹	۱	۰/۹
۴۷	۰/۲	۰/۳	۰/۳۵	۰/۳۵	۰/۳	۰/۳	۰/۶۵	۰/۷
۴۸	۰/۱	۰/۱۵	۰/۲	۰/۳۵	۰/۱۵	۰/۳۵	۰/۳۵	۰/۷
۴۹	۰/۳۵	۰/۵	۰/۸	۰/۹۵	۰/۵۵	۰/۸	۰/۸	۰/۸
۵۰	۰/۲	۰/۲۵	۰/۲۵	۰/۳۵	۰/۵	۰/۵	۰/۴۵	۰/۷
۵۱	۰/۱	۰/۲۵	۰/۳	۰/۳۵	۰/۱	۰/۴۵	۰/۵	۰/۷۵
۵۲	۰/۶۵	۰/۷۵	۱	۱	۰/۸	۱	۱	۱
۵۳	۰/۱	۰/۱	۰/۳	۰/۶۵	۰/۲	۰/۴	۰/۷۵	۰/۷۵
۵۴	۰/۱	۰/۲۵	۰/۵	۰/۷	۰/۲	۰/۴	۰/۷۵	۰/۶۵
۵۵	۰/۴۵	۰/۸۵	۰/۹۵	۱	۰/۶	۰/۹	۰/۹۵	۱
میانگین	۰/۲۸	۰/۴۳	۰/۵۷	۰/۶۸	۰/۳۸	۰/۶۱	۰/۷۲	۰/۷۸

در تمامی شرایط (برای تمامی مقادیر کارکرد افتراقی و شرایط اثر آزمون و بدون اثر آزمون) توان روش نسبت درستنمایی در شناسایی کارکرد افتراقی یکنواخت بیشتر از روش رگرسیون لجستیک می‌باشد. در موقعیت بدون اثر روش نسبت درستنمایی با نرخ متوسط آشکارسازی ۶۲٪ نسبت به روش رگرسیون لجستیک با نرخ متوسط آشکارسازی ۴۹٪ قدرت بیشتری در رد فرض صفر مبنی بر عدم کارکرد افتراقی یکنواخت سؤال و شناسایی کارکرد افتراقی یکنواخت دارد. تحت شرایط اثر آزمون نیز روش نسبت درستنمایی با نرخ

متوسط آشکارسازی ۵۲٪ نسبت به روش رگرسیون لجستیک با نرخ متوسط آشکارسازی ۴۲٪ توان بیشتری در شناسایی کارکرد افتراقی دارد. میزان شناسایی کارکرد افتراقی یکنواخت در هر دو روش در شرایط اثر آزمون کم‌تر است. با بررسی پارامتر سؤال‌ها مشخص می‌شود که روش نسبت درستنمایی معمولاً درشناسایی کارکرد افتراقی یکنواخت سؤال‌هایی که پارامتر تشخیص بالا و دشواری متوسط دارند، توانمندتر است.

جدول ۳. توان کارکرد افتراقی یکنواخت شرایط اثر

سؤال	کارکرد افتراقی یکنواخت							
	رگرسیون لجستیک				آزمون نسبت درستنمایی			
	۰/۴	۰/۶	۰/۸	۱	۰/۴	۰/۶	۰/۸	۱
۴۱	۰/۶	۰/۷۵	۰/۸	۰/۹۵	۰/۶۵	۰/۸۵	۱	۰/۹۵
۴۲	۰/۴	۰/۵۵	۰/۹۵	۱	۰/۶۵	۰/۷۵	۰/۹۵	۱
۴۳	۰/۱	۰/۱۵	۰/۲۵	۰/۵	۰/۱	۰/۱۵	۰/۳	۰/۳۵
۴۴	۰/۱۵	۰/۱	۰/۱۵	۰	۰/۰۵	۰/۱	۰/۰۵	۰/۰۵
۴۵	۰/۱۵	۰/۲	۰/۰۵	۰/۱۵	۰/۱۵	۰/۲۵	۰/۱	۰/۳
۴۶	۰/۴	۰/۶	۰/۸۵	۰/۸۵	۰/۵	۰/۷	۰/۹۵	۰/۹۵
۴۷	۰/۱	۰/۰۵	۰/۱	۰/۱	۰/۱	۰/۰۵	۰/۱	۰/۱
۴۸	۰/۱	۰/۲	۰/۲	۰/۵	۰/۳۵	۰/۴	۰/۶	۰/۸
۴۹	۰/۲۵	۰/۳	۰/۴۵	۰/۷	۰/۲۵	۰/۴	۰/۶	۰/۷۵
۵۰	۰/۳۵	۰/۴۵	۰/۵۵	۰/۶۵	۰/۵	۰/۶	۰/۷	۰/۸۵
۵۱	۰/۱۵	۰/۲	۰/۳	۰/۴۵	۰/۲۵	۰/۶	۰/۶	۰/۷۵
۵۲	۰/۳۵	۰/۸	۰/۹۵	۱	۰/۶۵	۰/۹۵	۰/۹۵	۰/۹۵
۵۳	۰/۱	۰/۴۵	۰/۵	۰/۸	۰/۲۵	۰/۹۵	۰/۸۸	۰/۸۵
۵۴	۰/۲	۰/۱	۰/۲۵	۰/۰۵	۰/۱۵	۰/۱۵	۰/۲۵	۰/۲
۵۵	۰/۵	۰/۴۵	۰/۲۵	۰/۸۵	۰/۶	۰/۶۵	۰/۹۵	۰/۹۵
میانگین	۰/۲۶	۰/۳۶	۰/۴۸	۰/۵۷	۰/۳۵	۰/۵	۰/۶	۰/۶۴

توان آماری کارکرد افتراقی غیریکنواخت: توان آماری شناسایی کارکرد افتراقی غیریکنواخت (نسبت مثبت واقعی) دو رویکرد رگرسیون لجستیک و آزمون نسبت درستنمایی در ۲۰ تکرار برای سطوح مختلف کارکرد افتراقی و تحت شرایط اثر آزمون و بدون اثر در جداول شماره ۴ و ۵ ارائه شده است توان آماری روش رگرسیون لجستیک و

آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی غیریکنواخت به شدت کارکرد افتراقی بستگی دارد. با افزایش شدت کارکرد افتراقی میانگین نسبت مثبت واقعی هر دو روش رگرسیون لجستیک و آزمون نسبت درستنمایی تحت شرایط یکسانی و عدم یکسانی توزیع توانایی گروه مرجع و کانونی افزایش یافته است. در شرایط بدون اثر زمانی که شدت کارکرد افتراقی متوسط بود (۰/۶) رویکرد نسبت درستنمایی در مقایسه با روش رگرسیون لجستیک در شناسایی کارکرد افتراقی غیریکنواخت توانمندتر بود. هنگامی که شدت کارکرد افتراقی ۰/۴، ۰/۸ و ۱ بود، توان دو روش در شناسایی کارکرد افتراقی سؤال‌ها تقریباً یکسان است. نرخ متوسط آشکارسازی کارکرد افتراقی غیریکنواخت روش رگرسیون لجستیک و نسبت درستنمایی به ترتیب برابر ۲۹ و ۳۱ درصد است که مقادیر کمی می‌باشد.

جدول ٤. توان کارکرد افتراقی غیریکنواخت شرایط بدون اثر

سؤال	کارکرد افتراقی غیریکنواخت							
	رگرسیون لجستیک				آزمون نسبت درستنمایی			
	۰/۴	۰/۶	۰/۸	۱	۰/۴	۰/۶	۰/۸	۱
۴۱	۰/۳	۰/۶۵	۰/۸	۰/۸۵	۰/۳۵	۰/۸۵	۰/۸۵	۰/۷۵
۴۲	۰/۱	۰/۱	۰/۳۵	۰/۳	۰/۰۵	۰/۲	۰/۳	۰/۴۵
۴۳	۰/۰۵	۰/۱	۰/۲۵	۰/۳۵	۰/۰۵	۰/۲۵	۰/۲	۰/۲۵
۴۴	۰/۰۵	۰/۲۵	۰/۲۵	۰/۲۵	۰/۰۵	۰/۲۵	۰/۱	۰/۱۵
۴۵	۰/۰۵	۰/۱	۰/۱	۰/۱۵	۰/۰۵	۰/۱	۰/۱	۰/۳
۴۶	۰/۵	۰/۸	۰/۹	۱	۰/۵	۰/۷	۰/۸۵	۰/۹۵
۴۷	۰/۰۵	۰/۱	۰/۱	۰/۰۵	۰/۰۵	۰/۱	۰/۱۵	۰/۱۵
۴۸	۰/۱۵	۰/۱	۰/۰۵	۰/۱	۰/۲۵	۰/۲	۰/۲	۰/۱۵
۴۹	۰/۱	۰/۲	۰/۵	۰/۴	۰/۱۵	۰/۲۵	۰/۵	۰/۳۵
۵۰	۰/۰۵	۰/۱	۰/۰۵	۰/۱	۰/۱۵	۰/۱۵	۰/۱	۰/۱۵
۵۱	۰/۲	۰/۰۵	۰/۱	۰/۱۵	۰/۰۵	۰/۱	۰/۱۵	۰/۲
۵۲	۰/۲۵	۰/۷	۰/۹	۱	۰/۳۵	۰/۸	۰/۹	۱
۵۳	۰/۱	۰/۱	۰/۰۵	۰/۰۵	۰/۰۵	۰/۱۵	۰/۰۵	۰/۱
۵۴	۰/۱	۰/۱	۰/۱۵	۰/۲۵	۰/۱۵	۰/۱	۰/۱۵	۰/۰۵
۵۵	۰/۲۵	۰/۶۵	۰/۷۵	۰/۸۵	۰/۳۵	۰/۸	۰/۸	۰/۸
میانگین	۰/۱۵	۰/۲۷	۰/۳۵	۰/۳۹	۰/۱۷	۰/۳۳	۰/۳۶	۰/۳۹

در شرایطی که توزیع توانایی دو گروه مرجع و کانونی متفاوت بود (اثر آزمون) به استثناء زمانی که شدت کارکرد افتراقی متوسط (۰/۶) بود در سایر شرایط توان دو روش در شناسایی کارکرد افتراقی سؤال‌ها تقریباً مشابه بود. نرخ متوسط آشکارسازی کارکرد افتراقی غیریکنواخت روش رگرسیون لجستیک و نسبت درستنمایی در حالت اثر به ترتیب برابر ۳۴ و ۳۱ درصد است. میزان شناسایی کارکرد افتراقی غیریکنواخت روش رگرسیون لجستیک در موقعیت اثر آزمون بیشتر از بدون اثر است. در حالی که میزان شناسایی کارکرد افتراقی غیریکنواخت روش نسبت درستنمایی هنگامی که شدت کارکرد افتراقی ۰/۴ (کم) است در موقعیت اثر آزمون بیشتر از بدون اثر است و برای سایر مقادیر کارکرد افتراقی تفاوتی دو موقعیت مشاهده نمی‌شود. با بررسی پارامتر سؤال‌هایی که نسبت مثبت واقعی بالایی دارند مشخص می‌شود این سؤال‌ها پارامتر تشخیص بالا و دشواری متوسطی دارند.

جدول ۵. توان کارکرد افتراقی غیریکنواخت شرایط اثر

سؤال	کارکرد افتراقی غیریکنواخت							
	رگرسیون لجستیک				آزمون نسبت درستنمایی			
	۰/۴	۰/۶	۰/۸	۱	۰/۴	۰/۶	۰/۸	۱
۴۱	۰/۳۵	۰/۶۵	۰/۷	۰/۸	۰/۷	۰/۸۵	۰/۹	۱
۴۲	۰/۱۵	۰/۴۵	۰/۶۵	۰/۵	۰/۰۵	۰/۴	۰/۸	۰/۸
۴۳	۰/۱۵	۰/۲۵	۰/۲۵	۰/۵	۰/۱۵	۰/۰۵	۰/۰۵	۰/۱
۴۴	۰/۵۵	۰/۴۵	۰/۴۵	۰/۴۵	۰/۳	۰/۳	۰/۳	۰/۲۵
۴۵	۰/۱۵	۰/۳	۰/۳	۰/۴۵	۰/۱۵	۰/۱	۰/۱	۰/۲
۴۶	۰/۱۵	۰/۵	۰/۱۵	۰/۱۵	۰/۳۵	۰/۴۵	۰/۴۵	۰/۶
۴۷	۰/۶	۰/۴۵	۰/۹	۰/۸۵	۰/۳۵	۰/۴	۰/۳	۰/۳۵
۴۸	۰/۱	۰/۲	۰/۱	۰/۲	۰/۰۵	۰/۱۵	۰/۱۵	۰/۱
۴۹	۰/۰۵	۰/۱۵	۰/۲۵	۰/۲۵	۰/۱	۰/۱	۰/۱	۰/۰۵
۵۰	۰/۱	۰/۰۵	۰/۰۵	۰/۱۵	۰/۱۵	۰/۱۵	۰/۱	۰/۲۵
۵۱	۰/۰۵	۰/۰۵	۰/۰۵	۰/۱	۰/۱	۰/۰۵	۰/۱۵	۰/۲۵
۵۲	۰/۲	۰/۹۵	۰/۹۵	۰/۹۵	۰/۶	۰/۹	۰/۹۵	۰/۹۵
۵۳	۰/۰۵	۰/۰۵	۰/۱	۰/۰۵	۰/۰۵	۰/۱۵	۰/۱۵	۰/۱۵
۵۴	۰/۵	۰/۵۵	۰/۶	۰/۵	۰/۱۵	۰/۱	۰/۱۵	۰/۱۵
۵۵	۰/۰۵	۰/۲	۰/۲	۰/۲	۰/۱۵	۰/۲۵	۰/۴۵	۰/۴

میانگین	۰/۲۰	۰/۳۵	۰/۳۸	۰/۴۱	۰/۲۳	۰/۲۹	۰/۳۴	۰/۳۷

با مقایسه نتایج جداول ۲ الی ۵ مشخص می‌شود که به طور کلی توانمندی هر دو روش نسبت درستنمایی سؤال – پاسخ و رگرسیون لجستیک در شناسایی کارکرد افتراقی یکنواخت بیشتر از کارکرد افتراقی غیریکنواخت می‌باشد.

خطای نوع اول (نسبت مثبت کاذب): خطای نوع اول (نسبت مثبت کاذب) دو رویکرد رگرسیون لجستیک و آزمون نسبت درستنمایی در ۲۰ تکرار برای موقعیت اثر آزمون و بدون اثر آزمون در جدول شماره ۶ ارائه شده است. همان‌گونه که در جدول مشخص شده است کنترل خطای نوع اول در رگرسیون لجستیک تحت تاثیر اثر آزمون (توزیع متفاوت گروه‌های کانونی و مرجع) است. در شرایط بدون اثر روش رگرسیون لجستیک کنترل مناسبی بر خطای نوع اول داشته است، میانگین نسبت مثبت کاذب روش رگرسیون لجستیک در شرایط بدون اثر برابر با ۰/۰۵ با بیشینه ۰/۲ برای سؤال شماره ۴۴ است. در شرایط بدون اثر میزان خطای نوع اول روش رگرسیون لجستیک برای ۹ سؤال از ۱۵سؤال مساوی یا کم‌تر از ۰/۰۵ است. در شرایط اثر آزمون میانگین نسبت مثبت کاذب روش رگرسیون لجستیک برابر ۰/۱۶ با بیشینه ۰/۵ برای سؤال شماره ۴۴ است. سؤال ۴۴ با داشتن پارامتر دشواری ۲/۳۱۷ یکی از دشوارترین سؤال‌های بانک می‌باشد به این ترتیب مشخص می‌شود که تورم خطای نوع اول در روش رگرسیون لجستیک در شرایط اثر آزمون به پارامتر سؤال‌ها بستگی داشته است. در شرایط اثر میزان خطای نوع اول روش رگرسیون لجستیک ۵ سؤال مساوی یا کم‌تر از ۰/۰۵ است.

جدول ٦. نرخ خطای نوع اول (مثبت کاذب) در بانک سؤال اول

سؤال	نرخ خطای نوع اول			
	بدون اثر		با اثر	
	رگرسیون لجستیک	آزمون نسبت درستنمایی	رگرسیون لجستیک	آزمون نسبت درستنمایی
۴۱	۰/۱	۰/۰۵	۰/۱	۰/۱
۴۲	۰/۱	۰	۰/۱	۰/۰۵
۴۳	۰/۱	۰/۱۵	۰/۱۵	۰/۱
۴۴	۰/۲	۰/۱۵	۰/۵	۰/۱۵
۴۵	۰	۰	۰/۱	۰/۱۵

۴۶	۰/۱	۰	۰/۲	۰/۰۵
۴۷	۰	۰/۰۵	۰/۴	۰
۴۸	۰	۰/۰۵	۰/۰۵	۰/۱۵
۴۹	۰	۰	۰/۳	۰/۱۵
۵۰	۰/۰۵	۰/۰۵	۰/۰۵	۰
۵۱	۰/۰۵	۰	۰	۰
۵۲	۰	۰	۰/۰۵	۰/۰۵
۵۳	۰	۰/۰۵	۰/۰۵	۰
۵۴	۰/۱	۰/۰۵	۰/۲	۰/۱
۵۵	۰	۰/۰۵	۰/۱۵	۰/۱۵
میانگین	۰/۰۵	۰/۰۴	۰/۱۶	۰/۰۸

در روش آزمون نسبت درستنمایی نیز کنترل خطای نوع اول تحت تاثیر اثر آزمون است. میزان خطای نوع اول در شرایط بدون اثر کم‌تر از موقعیت اثر آزمون است. آزمون نسبت درستنمایی نیز در شرایط بدون اثر کنترل خوبی بر میزان خطای نوع اول داشته است میانگین نسبت مثبت کاذب روش نسبت درستنمایی در شرایط بدون اثر برابر با ۰/۰۴ است. در شرایط اثر آزمون میانگین نسبت مثبت کاذب برابر ۰/۰۸ است. همان‌گونه که مشاهده می‌شود روش نسبت درستنمایی خطای نوع اول را بهتر کنترل می‌کند در این روش تحت شرایط بدون اثر میزان خطای نوع اول ۱۳ مساوی یا کم‌تر از ۰/۰۵ است. حال آن که در شرایط اثر آزمون میزان خطای نوع اول ۷ سؤال در دامه مطلوب قرار گرفته است. در مقایسه دو روش مشخص می‌شود که در موقعیت بدون اثر دو روش رگرسیون و آزمون نسبت درستنمایی از نظر کنترل خطای نوع اول تقریباً مشابه هستند و در شرایط اثر نیز روش آزمون درستنمایی کنترل نسبتاً خوبی بر خطای نوع اول دارد.

بحث و نتیجه‌گیری

نظر به تمایل فزاینده کاربرد کامپیوترها در سنجش در قالب آزمون‌های کامپیوتری و انطباقی کامپیوتری، بررسی عادلانه بودن این نوع آزمون‌ها در چارچوب کارکرد افتراقی اهمیت بسزایی یافته است. تدوین و به کارگیری روش‌های شناسایی کارکرد افتراقی سؤال، پاسخی به ضرورت سنجش بدون سوگیری آزمودنی‌ها می‌باشد. ضرورت اجرا سؤال‌های فاقد کارکرد افتراقی در آزمون‌های انطباقی کامپیوتری از ماهیت انطباقی آزمون نشات

می‌گیرد (کیم، ۲۰۱۰). در آزمون‌های انطباقی کامپیوتری سؤال‌های کم‌تری اجرا می‌شود در نتیجه هر سؤال نقش برجسته‌تری در برآورد توانایی آزمودنی ایفا می‌کند و از این رو عدم سوگیری سؤال‌ها به منظور برآورد بهینه توانایی آزمودنی‌ها ضرورتی اجتناب‌ناپذیر می‌باشد. علی‌الرغم اهمیت بررسی عملکرد متفاوت سؤال‌ها در آزمون‌های انطباقی کامپیوتری، انجام آن در مقایسه با آزمون‌های مداد – کاغذی دشوارتر می‌باشد و با چالش‌های نظری و عملی همراه است که می‌توان به دشواری تعیین متغیر تطبیق به منظور جور کردن آزمودنی‌های گروه‌های مورد مقایسه اشاره کرد (زویک، ۲۰۱۰).

در ادبیات تحقیق شاخص‌هایی مبتنی بر روش‌های به کار رفته در آزمون‌های مداد – کاغذی به منظور بررسی کارکرد افتراقی در چارچوب آزمون‌های انطباقی کامپیوتری توسعه یافته‌اند. این روش‌ها کارکرد افتراقی را در سؤال‌های پیش‌آزمون که معمولاً به صورت غیرانطباقی اجرا می‌شوند، مورد بررسی قرار می‌دهند و تلاشی برای ارزیابی کارکرد افتراقی در سؤال‌های عملیاتی در جریان آزمون انطباقی کامپیوتری صورت نگرفته است. کارکرد افتراقی و روش‌های شناسایی آن در بافت آزمون‌های انطباقی کامپیوتری از جمله حوزه‌های آزمون‌سازی محسوب می‌شوند که علی‌الرغم اهمیت با کمبود مطالعات نظامند مواجه هستند.

در مطالعه حاضر کارکرد افتراقی سؤال‌های عملیاتی در اجرای انطباقی کامپیوتری براساس روش آزمون نسبت درستنمایی سؤال – پاسخ و رگرسیون مورد بررسی قرار گرفت. به طور کلی این دو روش از جمله روش‌های انتخابی مطالعه کارکرد افتراقی در آزمون‌های انطباقی کامپیوتری می‌باشند. اثربخشی هر یک از روش‌ها تابعی از شرایط آزمون و نوع سوگیری مندرج در سؤال‌ها است. از روش‌های شبیه‌سازی به منظور تولید داده‌ها و اجرای آزمون انطباقی کامپیوتری استفاده شد. نوع کارکرد افتراقی ، مقدار کارکرد افتراقی و توزیع توانایی گروه مرجع و کانونی عواملی بودند که در این مطالعه دستکاری شدند. تکمیل ماتریس داده‌های ناقص حاصل از اجرای مجموعه متفاوتی از سؤال‌ها در اجرای انطباقی کامپیوتری به روش جایگذاری براساس احتمال پاسخ‌گویی درست به سؤال‌ها مبتنی بر برآورد توانایی آزمون انطباقی کامپیوتری برای هر آزمودنی صورت گرفت. مقایسه دو رویکرد رگرسیون لجستیک و آزمون نسبت درستنمایی براساس ملاک نسبت

مثبت کاذب (خطای نوع اول) و مثبت واقعی (توان) در سطح آلفا ۰/۰۵ در ۲۰ تکرار برای هر موقعیت انجام شد.

بررسی ادبیات پژوهشی بیانگر آن است که میزان توان و خطای نوع اول روش‌های شناسایی کارکرد افتراقی تحت‌تاثیر عوامل مختلفی از جمله طول آزمون، حجم نمونه، شدت کارکرد افتراقی، نسبت سؤال‌های دارای کارکرد افتراقی و تفاوت میانگین گروه‌های مقایسه (اثر آزمون) می‌باشد (کاباساکال و همکاران، ۲۰۱۴؛ کیم، ۲۰۱۰). نتایج مطالعه حاضر نشان داد که میزان کنترل خطای نوع اول روش رگرسیون لجستیک و آزمون نسبت درستنمایی سؤال – پاسخ متاثر از اثر آزمون است. در شرایطی که توزیع توانایی دو گروه مرجع و کانونی یکسان است (بدون اثر) میزان خطای نوع اول هر دو روش مطلوب است. همان‌گونه که لی، چن و یو (۲۰۰۶) گزارش کردند در این مطالعه نیز میزان خطای نوع اول روش رگرسیون لجستیک تحت شرایط بدون اثر به گونه کافی کنترل شده است اما تحت شرایط اثر آزمون خطای نوع اول متورم شده است. کفایت کنترل خطای نوع اول روش رگرسیون لجستیک در شرایط یکسانی میانگین گروه‌های مقایسه مرجع و کانونی و تورم خطای نوع اول در شرایط تفاوت میانگین گروه‌ها در چارچوب آزمون‌های مداد – کاغذی نیز توسط نارایانان[1] و سوامی‌ناتان (۱۹۹۴)، لی و استوت[2] (۱۹۹۶) نیز گزارش شده است.

با افزایش شدت کارکرد افتراقی، توان هر دو روش رگرسیون لجستیک و آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی یکنواخت و غیریکنواخت افزایش یافته است. این روند در هر دو موقعیت یکسانی و عدم یکسانی توزیع توانایی گروه‌های مقایسه مشاهده شده است. این نتایج همسو با یافته‌های مستند شده برای آزمون‌های انطباقی کامپیوتری و مداد- کاغذی در پیشینه تحقیق است (لی، چن و یو ، ۲۰۰۶؛ نانداکومار و روسوس،۲۰۰۴).

توان هر دو روش نسبت درستنمایی سؤال – پاسخ و رگرسیون لجستیک در شناسایی کارکرد افتراقی یکنواخت در شرایط یکسانی میانگین گروه‌های مرجع و کانونی بیشتر است. هم‌چنین توان روش آزمون نسبت درستنمایی سؤال – پاسخ در تشخیص سؤال‌های

1. Narayanan
2. Li & Stout

دارای کارکرد افتراقی یکنواخت در هر دو موقعیت یکسانی و عدم یکسانی توزیع توانایی گروه‌ها از روش رگرسیون لجستیک بیشتر می‌باشد. در تمامی شرایط نرخ متوسط آشکارسازی کارکرد افتراقی یکنواخت روش نسبت درستنمایی بیشتر از روش رگرسیون لجستیک است. این یافته با نتایج مطالعه لی، چن و یو (۲۰۰۶) که بیانگر مشابهت دو روش نسبت درستنمایی سؤال – پاسخ و رگرسیون لجستیک در شناسایی کارکرد افتراقی یکنواخت است، هم‌خوانی ندارد. یافته‌های مطالعات کارکرد افتراقی در چارچوب آزمون‌های مداد - کاغذی نیز موید حساسیت بیشتر روش نسبت درستنمایی در مقایسه با روش رگرسیون لجستیک می‌باشد (کیم، ۲۰۱۰).

هر دو روش رگرسیون لجستیک و آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی غیریکنواخت عملکرد ضعیفی داشتند. توان هر دو روش در شناسایی کارکرد افتراقی غیریکنواخت به شدت کارکرد افتراقی بستگی داشته است و با افزایش شدت کارکرد افتراقی افزایش یافته است. در شرایط بدون اثر زمانی که شدت کارکرد افتراقی ۰/۴، ۰/۸ و ۱ می‌باشد توان دو روش در شناسایی کارکرد افتراقی سؤال‌ها تقریباً یکسان بوده است، در حالی که تحت شرایط کارکرد افتراقی متوسط (۰/۶) رویکرد نسبت درستنمایی توانمندتر است. هنگامی که توزیع توانایی دو گروه مرجع و کانونی متفاوت است، به استثناء زمانی که شدت کارکرد افتراقی متوسط (۰/۶) بود در سایر شرایط توان دو روش در شناسایی کارکرد افتراقی سؤال‌ها تقریباً مشابه است. این یافته با نتایج مطالعه لی، چن و یو (۲۰۰۶) هم‌خوانی ندارد.

با توجه به توان و میزان خطای نوع اول دوریکرد رگرسیون لجستیک و آزمون نسبت درستنمایی در شناسایی کارکرد افتراقی یکنواخت در چارچوب آزمون انطباقی کامپیوتری می‌توان نتیجه گرفت که در شناسایی کارکرد افتراقی یکنواخت، روش آزمون نسبت درستنمایی از کارایی بیشتری برخوردار است در حالی‌که در شناسایی کارکرد افتراقی غیریکنواخت تفاوتی میان دو روش مشاهده نشد و هر دو روش به طور مشابهی ضعیف عمل کرده‌اند.

به نظر می‌رسد نتایج مطالعه حاضر و عدم هم‌خوانی آن با یافته‌ها پژوهش لی، چن و یو (۲۰۰۶) به دلیل بررسی ارزیابی کارکرد افتراقی در سؤال‌های عملیاتی ، فرایند جایگذاری داده‌های ناقص و هم‌چنین تغییرپذیری پارامتر تشخیص سؤال‌های مورد بررسی باشد. در

مطالعات قبلی کارکرد افتراقی در سؤال‌های پیش آزمون مورد بررسی قرار گرفته‌اند. سؤال‌های عملیاتی برخلاف سؤال‌های پیش آزمون که برای تمامی آزمودنی‌ها اجرا می‌شود، به طور انطباقی برای آزمودنی‌ها اجرا می‌شود. فرایند جایگذاری پاسخ‌ها در سؤال‌های عملیاتی براساس روش احتمال پاسخ‌گویی درست به سؤال‌ها و ماهیت انطباقی اجرای سؤال‌ها ممکن است توان و خطای روش‌های ارزیابی رگرسیون لجستیک و آزمون نسبت درستنمایی را تحت تاثیر قرار دهد و منجر به کاهش توان دو روش مذکور شده باشد. ترکیب مختلف پارامتر تشخیص سؤال‌ها مورد بررسی نیز ممکن است بر نتایج به دست آمده تاثیرگذاشته باشد.

به علاوه به نظر می‌رسد نتایج مطالعه حاضر به دلیل اثر تفاوت مرحله اجرای سؤال‌های دارای کارکرد افتراقی در اجرای انطباقی کامپیوتری آزمون هم باشد. از آن‌جایی که در آزمون انطباقی کامپیوتری سؤال‌ها متناسب با توانایی آزمودنی‌ها اجرا می‌شود، ترتیب اجرای سؤال‌ها برای آزمودنی‌های مختلف، متفاوت است در نتیجه سؤال دارای کارکرد افتراقی در جریان آزمون انطباقی کامپیوتری در مراحل مختلفی برای آزمودنی‌ها اجرا می‌شود. به عنوان مثال سؤال ممکن است برای یک آزمودنی در ابتدا آزمون برای آزمودنی دیگر در اواسط و یا در مراحل پایانی اجرا شود. تفاوت در زمان اجرا سؤال ممکن است عملکرد افتراقی سؤال را تحت‌تاثیر قرار دهد. به نظر می‌رسد وقتی سؤال دارای کارکرد افتراقی در ابتدا آزمون اجرا می‌شود، عملکرد الگوریتم آزمون انطباقی کامپیوتری قادر به اصلاح اثرات کارکرد افتراقی سؤال باشد، حال اگر این سؤال در انتهای آزمون اجرا شود اثرات سوگیرانه آن نمود بیشتری پیدا کند. تعداد سؤال‌های دارای کارکرد افتراقی و شدت کارکرد افتراقی نیز ممکن است عملکرد اصلاحی الگوریتم آزمون انطباقی کامپیوتری را تحت تاثیر قرار دهد. بنابراین تاثیر مرحله اجرای سؤال‌های دارای کارکرد افتراقی و عوامل موثر برآن می‌بایست در مطالعات بعدی مورد بررسی قرارگیرد.

به منظور بررسی دقیق‌تر پیشنهاد می‌شود عملکرد سایر روش‌های ارزیابی کارکرد افتراقی از جمله روش مانتل هنزل و CATSIB نیز در شناسایی کارکرد افتراقی سؤال‌های عملیاتی آزمون‌های انطباقی کامپیوتری به شیوه جایگذاری پاسخ‌ها در سؤال‌های عملیاتی براساس روش احتمال پاسخ‌گویی درست به سؤال‌ها مورد بررسی قرار گیرد. به علاوه برای کسب تصویری روشن‌تر از روش بهینه ارزیابی کارکرد افتراقی در آزمون‌های

انطباقی کامپیوتری تاثیر سایر عوامل از جمله درصد تعداد سؤال‌های دارای کارکرد افتراقی و نسبت حجم نمونه در شناسایی کارکرد افتراقی در مطالعات دیگر مورد بررسی قرار گیرد. هم‌چنین روش نسبت درستنمایی سؤال – پاسخ همانند هر روش مبتنی بر مدل ممکن است در صورت انتخاب مدل نامناسب عملکرد خوبی نداشته باشد. بنابراین لازم است در آینده میزان مقاوم بودن این روش در صورت کاربرد مدل نادرست نیز مورد ارزیابی قرار گیرد.

منابع

کروکر، لیندا و آلجینا، جیمز. (۲۰۰۶).مباحث نو در روان سنجی. ترجمه و گردآوری ولی اله فرزاد، حسین زارع (۱۳۸۸). تهران: آییژ

گرامی پور، مسعود. (۱۳۹۰). مقایسه قدرت آزمون نسبت درستنمایی مبتنی بر مدل سؤال – پاسخ با روش‌های تحلیل عاملی تاییدی و رگرسیون لجستیک در شناسایی کنش افتراقی سؤال به منظور اطمینان از عادلانه بودن سنجش آزمون‌های سرنوشت ساز. پایان نامه دکتری سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی

همبلتون، رونالد ک.، سوامیناتان، اچ. و راجرز، جین (۱۹۹۱). مبانی نظریه پرسش- پاسخ. ترجمه محمدرضا فلسفی نژاد، (۱۳۸۹). انتشارات دانشگاه علامه طباطبائی

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999).Standards for educational and psychological testing. Amer Educational Research Assn.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20(3), 213-229.

Choi, S. W., Podrabsky, T., & McKinney, N. (2011). Firestar-D: Computerized Adaptive Testing Simulation Program for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 0146621611406107.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. Applied Psychological Measurement, 20(1), 15-26.

Conoley, C. A. (2003). *Differential item functioning in the Peabody Picture Vocabulary Test–Third Edition: Partial correlation versus expert judgment* (Doctoral dissertation, Texas A&M University).

Davidson, P. (2003). Why technology had had only a minimal impact on testing in education? Proceedings from the 2nd Education Technology Conference and Exhibition. Oman: Sultan Qaboos University

Dorans, N. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368

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).

Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. Applied Psychological Measurement, 31(5), 457-459.

Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied psychological measurement, 20(2), 101-125.

Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques.Quality & Quantity, 42(6), 739-755.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel- Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), Test validity (pp. 129- 145). Hillsdale, NJ: Lawrence Erlbaum.

Kabasakal, K. A., Gök, B., Arsan, N., & Kelecioglu, H. (2014). Comparing Performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. Kuram ve Uygulamada Egitim Bilimleri, 14(6), 2186.

Kalender, Ilker. (2011). *Effect of Different Computerized Adaptive Testing Strategies on Recovery of Ability*.(Doctoral dissertation, Middle East Technical University).

Kim, J. (2010). Controlling type 1 error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. Journal of Educational Measurement, 43(4), 355-381.

Lei, P.-W., Chen, S.-Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. Journal of Educational Measurement, 43, 254-264.

Li, H.-H., & Stout,W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*(4), 647–677

Miller, T. R. (1992). Practical considerations for conducting studies of differential item functioning (DIF) in a CAT environment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Nandakumar, R., & Roussos, L. (2001). CATSIB: A Modified SIBTEST Procedure To Detect Differential Item Functioning in Computerized Adaptive Tests. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. Journal of Educational and Behavioral Statistics, 29, 177-199.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. Applied Psychological Measurement, 18(4), 315-328.

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

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www. psych. umn. edu/psylabs/CATCentral*

Piromsombat, C. (2014). *Differential Item Functioning in Computerized Adaptive Testing: Can CAT Self-Adjust Enough?* (Doctoral dissertation, University of Minnesota).

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

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational measurement, 361-370.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. Journal of Educational Measurement, 35(2), 109-135

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, R. (2010). The investigation of differential item functioning in adaptive tests. In Elements of adaptive testing (pp. 331-352). Springer New York.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. Applied Psychological Measurement, 26, 57-76.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis (ETS Research Report RR-97-21). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-97-21.pdf

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. Applied Psychological Measurement, 18(2), 121-140.