Journal of Educational Measurement & Evaluation Studies فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
Vol. 6, No. 14, Summer 2016 سال ششم، شمارهٔ ۱۴، تابستان ۱۳۹۵، صص ۲۳۴-۲۰۳

تعیین مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد-کاغذی[۱]

نگار شریفی یگانه *
محمدرضا فلسفی‌نژاد**
علی دلاور***
نورعلی فرخی****
احسان جمالی*****

تاریخ دریافت مقاله: ۹۵/۰۲/۱۹
تاریخ پذیرش مقاله: ۹۵/۰۶/۱۰

چکیده

هدف مطالعه حاضر تعیین مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد – کاغذی و تعیین الگوریتم بهینه آزمون انطباقی کامپیوتری بر اساس روش‌های مختلف برآورد توانایی (بیشینه درست‌نمایی و پسین مورد انتظار) و ملاک خاتمه آزمون (خطای استاندارد ثابت و طول ثابت آزمون) در آزمون‌های خطیر بود. جامعه پژوهش شامل تمامی شرکت‌کنندگان آزمون سراسری گروه آزمایشی علوم ریاضی و فنی سال ۱۳۹۲ بود که از میان آنها تعداد ۱۰۰۰ آزمودنی با روش نمونه‌گیری تصادفی انتخاب شدند. تحلیل سؤال‌های آزمون ریاضی با استفاده از مدل لجستیک سه‌پارامتری صورت گرفت. ٤٠ مجموعه داده با حجمی برابر با داده‌های واقعی شبیه‌سازی شد و شبیه‌سازی پس‌تجربی آزمون انطباقی کامپیوتری انجام شد. یافته‌های تحلیل بیانگر همبستگی بالای برآورد توانایی اجرای انطباقی کامپیوتری و مداد-کاغذی خرده آزمون ریاضی بود. همچنین مقادیر سوگیری، میانگین قدر مطلق تفاوت برآوردهای توانایی آزمون انطباقی کامپیوتری و مداد- کاغذی و ریشه میانگین مجذور تفاوت بیانگر آن بود که برآوردهای توانایی آزمون انطباقی کامپیوتری در روش برآورد پسین مورد انتظار در راستای برآورد توانایی آزمون کامل است. نتایج نشان داد که آزمون انطباقی کامپیوتری قادر به بازیابی توانایی در خرده‌آزمون ریاضی است. به علاوه روش برآورد پسین مورد انتظار و ملاک خاتمه خطای استاندارد ثابت ۰/۳ الگوریتم بهینه دستیابی به اهداف پایایی مناسب، طول منطقی آزمون و بازیابی برآورد توانایی در اجرای انطباقی کامپیوتری خرده‌آزمون ریاضی است.

واژگان کلیدی: آزمون انطباقی کامپیوتری، شبیه‌سازی پس‌تجربی، برآورد بیشینه درست‌نمایی، برآورد پسین مورد انتظار، ملاک خاتمه آزمون

---

[۱] این مقاله برگرفته از رساله دکتری است.

* دانشجوی دکتری سنجش و اندازه‌گیری دانشگاه علامه طباطبائی
** دانشیار دانشکده روان‌شناسی و علوم تربیتی دانشگاه علامه طباطبائی (نویسنده مسئول falsafinejad@yahoo.co.uk)
*** استاد دانشکده روان‌شناسی و علوم تربیتی دانشگاه علامه طباطبائی
**** دانشیار دانشکده روان‌شناسی و علوم تربیتی دانشگاه علامه طباطبائی
***** استادیار سازمان سنجش آموزش کشور

مقدمه

کاربرد فناوری‌های نوین اطلاعاتی در جهان معاصر با سرعت فزاینده‌ای در حال گسترش است و تأثیر چشمگیری بر تمام جوانب زندگی ازجمله ابعاد اقتصادی، اجتماعی، صنعتی و آموزشی داشته است. در حوزه آموزش، فناوری‌های اطلاعاتی، فرایند تدریس و یادگیری را متحول کرده‌اند؛ به‌گونه‌ای که منجر به کاهش محدودیت‌های یادگیری، برابری فرصت‌ها و بهره‌وری در آموزش شده‌اند و مفاهیمی چون آموزش و یادگیری الکترونیکی، آموزش به کمک کامپیوتر، آموزش از راه دور و آموزش مجازی را در حوزه تعلیم و تربیت مطرح ساخته‌اند (بابایی، ۱۳۸۹). سنجش نیز به‌عنوان بخش جدایی‌ناپذیر فرایند آموزش بی‌تأثیر از ظهور و پیشرفت‌های فناوری‌های نوین ازجمله کامپیوتر نبوده است. به عقیده متخصصان آموزشی کاربرد فناوری کامپیوتر در سنجش به دلیل کاربرد روزافزون آن در حوزه آموزش، گریزناپذیر است. زمانی که فراگیران بیشترین فعالیت یادگیری خود را با استفاده از کامپیوتر انجام می‌دهند، سنجش یادگیری این افراد با روش‌های سنتی و مرسوم مداد – کاغذی غیر قابل توجیه است (بنت[1]، ۲۰۰۲).

آزمون‌های مداد- کاغذی اگرچه شکل غالب و پذیرفته شده اجرای آزمون‌ها هستند، اما ضرورتاً مؤثرترین و کارآمدترین روش محسوب نمی‌شوند. هنگامی که گروهی از آزمودنی‌ها با آزمون مداد- کاغذی واحد مورد سنجش قرار می‌گیرند، آزمون نمی‌تواند به‌طور هم‌زمان برای تمامی افراد بیشینه کارایی را داشته باشد و احتمال این‌که سؤال‌ها برای آزمودنی‌ها خیلی ساده یا دشوار باشد، زیاد است. چنین سؤال‌هایی اطلاعات اندکی درباره آزمودنی‌ها فراهم می‌کند و مانع تحقق سنجش کارآمد و پایای[2] آزمودنی ها می‌شود. در این موارد با استفاده از فناوری کامپیوتر می‌توان آزمون‌های مداد- کاغذی مرسوم را به‌گونه‌ای کارآمدتر به‌کار برد و اندازه‌ای دقیق‌تر از توانایی آزمودنی‌ها به دست آورد (کلندر[3]، ۲۰۱۱). انعطاف‌پذیری، تسهیل شرایط اجرا، افزایش کارایی، امکان سنجش دقیق و گسترده‌تر سازه‌ها، امکان نمره‌دهی ماشینی،

1. Bennett
2. reliable
3. Kalender

فراهم‌سازی امکان بازخورد سریع و افزایش امنیت آزمون[1] ازجمله دلایل اصلی کاربرد کامپیوتر در سنجش در موقعیت‌های مختلف ازجمله سنجش کلاسی و سنجش خطیر[2] محسوب می‌شود (ژوبرت و کریک[3]، ۲۰۰۹).

آزمون‌های کامپیوتری با داشتن فرم‌های مختلف امکان اجرای آزمون را در تمامی شرایط فراهم می‌سازند. در ساده‌ترین شکل که آزمون کامپیوتری خطی[4] است از کامپیوترها صرفاً به‌عنوان ابزاری برای اجرای آزمون‌ها استفاده می‌شود. در سطح پیچیده کامپیوترها با ارائه آزمون‌های متفاوت برای آزمودنی‌های مختلف سنجش را انفرادی می‌سازند که سنجش انطباقی کامپیوتری[5] (CAT) نامیده می‌شود (دوی و پیتونیاک[6]، ۲۰۰۶). در آزمون انطباقی کامپیوتری، سؤال‌ها متناسب با توانایی آزمودنی‌ها در جریان آزمون انتخاب و اجرا می‌شوند (ریکیسی[7]، ۱۹۸۹، به نقل از دیویدسون[8]، ۲۰۰۳). سنجش انطباقی کامپیوتری متأثر از مدل‌های نظریه سؤال - پاسخ[9] است. هدف اساسی سنجش انطباقی کامپیوتری استفاده از ویژگی نامتغیری[10] نظریه سؤال- پاسخ برای ایجاد الگوریتم اجرای انطباقی آزمون است. به این ترتیب آزمون برای هر آزمودنی خاص به‌گونه‌ای طراحی می‌شود که سؤال آزمون دقیقاً متناسب با سطح توانایی وی باشد و بیشترین آگاهی[11] را فراهم سازد، در عین حال نمره‌های آزمودنی‌ها روی مقیاس مشترک قرار گیرد. آزمون انطباقی کامپیوتری سعی در دستیابی به تعادلی ظریف میان چندین ویژگی جذاب اما در عین حال متضاد برآورد پایاتر توانایی آزمودنی‌ها، کاهش قابل‌توجه تعداد سؤال‌های آزمون، نمره‌دهی سریع، انعطاف‌پذیری در زمان‌بندی آزمون و افزایش امنیت آزمون را دارد (دوی و پیتونیاک، ۲۰۰۶؛ کلندر، ۲۰۱۱).

---

1. test security
2. high Stake
3. Joubert & Kriek
4. Linear Computer Tests
5. Computerized Adaptive Testing
6. Davey & Pitoniak
7. Reckase
8. Davidson
9. Item Response Theory
10. invariance
11. information

آزمون‌های انطباقی کامپیوتری توانمندی بالایی در تعیین جایگاه افراد در خصایص مکنون دارند و جایگزین جدی برای آزمون‌های مداد- کاغذی قلمداد می‌شوند. بسیاری از سازمان‌ها و مؤسسات آزمون‌سازی بین‌المللی، نظام سنجش خود را به این سمت تغییر داده‌اند یا در حال مطالعه و برنامه‌ریزی جهت انجام آن هستند. آزمون تحصیلات تکمیلی[۱] (GRE)، آزمون پذیرش تحصیلات تکمیلی دوره مدیریت[۲] (GMAT)، آزمون مجوز پرستاری[۳] (NCLEX)، مجموعه آزمون‌های استعداد شغلی ارتش[٤] و آزمون تافل[٥] نمونه‌هایی از آزمون‌های خطیری هستند که به‌صورت انطباقی کامپیوتری اجرا می‌شوند (کلندر، ۲۰۱۱). انجمن پذیرش دانشکده حقوق[٦] نیز ازجمله مؤسساتی است که در حال انجام پروژه‌های مطالعاتی گسترده به‌منظور اجرای انطباقی کامپیوتری آزمون ورودی دانشکده حقوق[۷] است (برینگسجرد[۸]، ۲۰۰۱).

فرایند گذر از آزمون‌های مداد- کاغذی به انطباقی کامپیوتری بسیار پرمخاطره است و با چالش‌های نظری و عملی فراوانی همراه است که می‌توان به ضرورت فراهم‌سازی تجهیزات و زیرساخت‌های لازم، تأمین منابع مالی مورد نیاز، آماده‌سازی افراد جامعه برای پذیرش آزمون‌های کامپیوتری، تهیه بانک سؤال[۹] مدرج شده، بررسی و انتخاب استراتژی‌های بهینه اجرای آزمون، مقایسه‌پذیری[۱۰] نتایج و ارزیابی عادلانه[۱۱] بودن آزمون اشاره کرد. مقایسه‌پذیری نتایج حاصل از دو روش اجرای آزمون یکی از مهم‌ترین دغدغه‌های فرایند تغییر روش اجرای آزمون‌ها است. تغییر شیوه آزمون ممکن است منجر به تغییر سازه مورد سنجش شود. از این رو بر انجام مطالعات مقایسه‌پذیری به‌منظور حصول اطمینان از قابل مقایسه بودن نمره‌های آزمون‌های مداد- کاغذی و کامپیوتری (خطی و انطباقی) تأکید جدی شده است، به‌ویژه زمانی که

---

1. Graduate Record Examination
2. Graduate Management Admission Test
3. Nursing Licensing Examination (NCLEX)
4. Armed Services Vocational Aptitude Battery (ASVAB)
5. TOEFL
6. The Law School Admission Council (LSAC)
7. Law School Admission Test (LSAT)
8. Bringsjord
9. item Bank
10. comparability study
11. fair

از نسخه کامپیوتری به همراه نسخه مداد- کاغذی آزمون استفاده می‌شود. در این موقعیت هدف این است که نمره‌های دو نسخه تا حد ممکن قابل مقایسه باشند به‌طوری که هیچ آزمودنی با شرکت در یک شکل آزمون، امتیاز غیرمنصفانه‌ای دریافت نکند (انجمن تحقیقات آموزشی آمریکا[۱]، انجمن روان‌شناسی آمریکا[۲] و انجمن ملی سنجش در آموزش[۳]، ۱۹۹۹؛ وانگ و کولن[٤]، ۲۰۰۱). مطالعات مقایسه‌پذیری، اثرات احتمالی حاصل از تغییر روش اجرای آزمون را ارزیابی و روایی تفسیر نمره‌های آزمون را تضمین می‌کند (پیک[٥]، ۲۰۰۵).

در بررسی ادبیات پژوهشی، مقایسه‌پذیری دو دسته مطالعه را می‌توان جدا کرد: ۱- مقایسه‌پذیری آزمون‌های مداد- کاغذی و کامپیوتری و ۲- مقایسه‌پذیری آزمون‌های مداد- کاغذی و انطباقی کامپیوتری. در حالی که مطالعات نوع اول صرفاً تغییر روش اجرای آزمون را ارزیابی می‌کنند، در مطالعات دوم تأثیر روش اجرا و اثرات الگو در عملکرد آزمودنی‌ها بررسی می‌شود. بیشتر مطالعات مقایسه‌پذیری انجام شده مربوط به بخش اول است. با وجود جدی‌تر بودن مقایسه‌پذیری آزمون‌های انطباقی کامپیوتری به دلیل دشواری و مشکلات اجرایی ناشی از تغییر روش اجرا و الگوریتم ارائه آزمون، آزمون‌های انطباقی کامپیوتری کمتر با آزمون‌های مداد- کاغذی مقایسه شده‌اند (وانگ و شین[٦]، ۲۰۱۰؛ وانگ و کولن، ۲۰۰۱). الگوریتم‌های اجرایی آزمون انطباقی کامپیوتری مجموعه‌ای از قواعد هستند که شیوه آغاز، تداوم و خاتمه آزمون انطباقی کامپیوتری را نشان می‌دهد (اگن[۷]، ۲۰۰۷) و مشتمل بر ٦ مؤلفه ۱- تعیین مدل سؤال – پاسخ مناسب ۲- بانک سؤال بزرگ مدرج شده ۳- تعیین نقطه شروع آزمون ٤ – روش گزینش سؤال ٥- روش نمره‌دهی توانایی و ٦- ملاک خاتمه آزمون[۸] است (پونسودا[۹]، ۲۰۰۰؛ ویز و کینگزبری[۱]، ۱۹۸٤ به نقل از بابکوک[۲] و ویز، ۲۰۰۹). البته در

---

1. American Educational Research Association (AERA)
2. American Psychological Association (APA)
3. National Council on Measurement in Education (NCME)
4. Wang & Kolen
5. Peak
6. Shin
7. Eggen
8. test termination criterion
9. Ponsoda

در آزمون‌های انطباقی کامپیوتری که در برنامه‌های سنجش خطیر استفاده می‌شوند علاوه بر ٦ مؤلفه مذکور دو مؤلفه تعادل محتوایی[3] و کنترل میزان افشاء سؤال‌های[4] آزمون به‌منظور تأمین امنیت آزمون نیز در نظر گرفته می‌شود (ریکیسی، ۲۰۱۰). مزایای آزمون‌های انطباقی کامپیوتری بدون طراحی دقیق این مؤلفه‌ها حاصل نمی‌شود. ارزشیابی مقایسه‌پذیری آزمون‌های انطباقی کامپیوتری به‌واسطه مسائل مربوط به فرایندهای اجرایی هر یک از مؤلفه‌های الگوریتم‌های اجرایی آزمون بسیار پیچیده می می‌شود (وانگ و کولن، ۲۰۰۱؛ وانگ و شین، ۲۰۱۰).

برای اینکه قابل مقایسه بودن آزمون‌های مداد- کاغذی و انطباقی کامپیوتری از دقت و اعتبار کافی برخوردار باشد و بتواند به نتایج تکرارپذیری منتج شود، لازم است این مقایسه‌ها بر ملاک‌های مشخص، دقیق و از قبل تعیین شده مبتنی باشد. مقایسه‌پذیری آزمون‌های مداد- کاغذی و انطباقی کامپیوتری معمولاً بر اساس ملاک‌های روایی، روان‌سنجی و مفروضه‌های آماری- اجرایی ارزیابی می‌شود. ملاک روایی مستلزم آن است که نسخه انطباقی کامپیوتری و مداد- کاغذی آزمون، سازه یکسانی را اندازه بگیرند. برقراری ملاک روایی در آزمون‌های انطباقی کامپیوتری خیلی چالش‌برانگیز است، به این دلیل که اجرای انطباقی سؤال‌ها احتمال تفاوت در سازه و محتوای مورد سنجش را افزایش می‌دهد. در این ملاک مشخصات محتوایی، ابعاد و روابط آزمون با سایر متغیرها ازجمله تفاوت‌های زیرگروه‌ها بررسی می‌شود. این ملاک اشاره به یکسانی ویژگی‌های روان‌سنجی ازجمله خطای استاندارد اندازه‌گیری و پایایی آزمون مداد – کاغذی و انطباقی کامپیوتری دارد. در پیشینه مطالعات مقایسه‌پذیری بیشتر ملاک روان‌سنجی بررسی شده است (وانگ و شین، ۲۰۱۰). ملاک مفروضه‌های آماری- اجرایی آزمون هم به برقراری مفروضه‌های مقایسه‌پذیری اشاره دارد. مفروضه‌ها ممکن است برای طرح جمع‌آوری داده‌ها یا تحلیل آماری نمره‌های آزمون لازم باشند (وانگ و کولن، ۲۰۰۱؛ انجمن آموزشی تگزاس[5]، ۲۰۰۸).

---

1. Weiss & Kingsbury
2. Babcock
3. content balancing
4. item exposure
5. Texas Education Agency

مطالعات مقایسه‌پذیری نسخه انطباقی کامپیوتری و مداد– کاغذی آزمون با استفاده از روش‌های شبیه‌سازی[1] و داده‌های واقعی انجام می‌شود. روش‌های شبیه‌سازی معمولاً در مراحل آزمایشی و اولیه ارائه نسخه انطباقی کامپیوتری استفاده می‌شود (وانگ و کولن، ۲۰۰۱). مطالعات شبیه‌سازی امکان بررسی و قضاوت در خصوص چالش‌های تغییر روش اجرای آزمون را فراهم می‌سازد. با استفاده از روش‌های شبیه‌سازی می‌توان مؤلفه‌های مختلف آزمون انطباقی کامپیوتری را بررسی و استراتژی بهینه اجرایی آزمون انطباقی کامپیوتری را تعیین کرد (هارول و همکاران[2]، ۱۹۹۶).

اثرات تبدیل روش اجرای آزمون در مطالعات مختلف بررسی شده است و با نتایج متفاوت و بعضاً متناقضی همراه بوده است. یافته‌های برخی از مطالعات (القادر، کلارک و اندرسون[3]، ۱۹۹۸؛ برگستروم[4]، ۱۹۹۲؛ بولوت و کان[5]،۲۰۱۲؛ کلندر، ۲۰۱۱؛ هنلی کلب، مک‌براید و کودک[6]، ۱۹۸۹) نشان‌دهنده هم‌ارزی آزمون‌های انطباقی کامپیوتری و مداد- کاغذی است در حالی که نتایج مطالعات دیگر (شفر و همکاران[7]، ۱۹۹۸؛ وانگ، پان و هریس[8]، ۱۹۹۹) بیانگر عدم هم‌ارزی نمره‌های دو روش اجرا است. زمانی که نتایج مطالعات بیانگر قابل مقایسه بودن عملکرد آزمودنی‌ها در دو نسخه مداد- کاغذی و انطباقی کامپیوتری آزمون است، این به آن معنا است که آزمودنی‌ها از روش اجرای آزمون متضرر یا متنفع نمی‌شوند (انجمن آموزشی تگزاس، ۲۰۰۸).

القادر، کلارک و اندرسون (۱۹۹۸) در مطالعه‌ای، هم‌ارزی شکل مداد- کاغذی و انطباقی کامپیوتری سه خرده آزمون توانایی عددی[9]، استدلال انتزاعی[10] و استدلال مکانیکی[11] آزمون استعداد افتراقی[1] را ارزیابی کردند. نتایج مطالعه بیانگر هم‌ارزی دو

---

1. simulation
2. Harwell et al
3. Alkhadher, Clarke & Anderson
4. Bergstrom
5. Bulut & Kan
6. Henly, Klebe, McBride& Cudeck
7. Schaeffer et al
8. Pan & Harris
9. Numerical Ability (NA)
10. Abstract Reasoning (AR)
11. Mechanical Reasoning (M)

روش اجرای آزمون در دو خرده آزمون استدلال انتزاعی و استدلال مکانیکی بود اما هم‌ارزی در خرده آزمون توانایی عددی برقرار نبود.

شفر و همکاران (۱۹۹۳) مقایسه‌پذیری آزمون تحصیلات تکمیلی (GRE) را در دو مرحله بررسی کردند. در مرحله اول نسخه مداد – کاغذی با نسخه کامپیوتری آزمون مقایسه شد. نتایج نشان داد که نمره‌های دو نسخه آزمون قابل مقایسه هستند. در مرحله دوم نسخه انطباقی کامپیوتری و کامپیوتری آزمون مقایسه شد. نتایج مرحله دوم بیانگر مشابهت نمره‌های بخش کلامی و کمی در دو نسخه آزمون بود اما نمره‌های بخش تحلیلی نسخه انطباقی کامپیوتری به‌طور معناداری بیشتر از نسخه کامپیوتری آزمون بود (شفر و همکاران، ۱۹۹۵). شفر و همکاران (۱۹۹۸) مطالعه‌ای دیگری را برای مقایسه نسخه انطباقی کامپیوتری و مداد – کاغذی آزمون تحصیلات تکمیلی (GRE) طراحی و اجرا کردند. مقایسه‌پذیری به‌صورت مشابهت توزیع نمره‌های نسخه انطباقی کامپیوتری و مداد – کاغذی تعریف شد. نتایج نشان داد که میانگین نمره‌های نسخه انطباقی کامپیوتری آزمون بیشتر از نسخه مداد – کاغذی آزمون بود.

وانگ، پان و هریس (۱۹۹۹) شبیه‌سازی پس‌تجربی[2] آزمون انطباقی کامپیوتری را با استفاده از داده‌های آزمون ورودی دانشکده حقوق انجام دادند. این مطالعه اولین شبیه‌سازی انطباقی کامپیوتری با استفاده از آزمودنی‌های واقعی بود. در این مطالعه برآورد توانایی اصلی و اجرای انطباقی کامپیوتری مقایسه شد. با استفاده از پاسخ آزمودنی‌های واقعی به ۱۲۷ سؤال آزمون، دقت، کارآمدی بازیابی توانایی آزمودنی‌ها و تعداد سؤال‌های مورد نیاز برای تکمیل آزمون انطباقی کامپیوتری در سه سطح دقت (خطای استاندارد ۰/۳۱۶۲، ۰/۲۶۵ و ۰/۱۷۳) بررسی شد. یافته مطالعه نشان داد که ۱۲۷ سؤال برای اجرای انطباقی کامپیوتری آزمون برای تمام آزمودنی‌ها در سه سطح دقت کافی بوده است. کاربرد داده‌های واقعی در جریان آزمون انطباقی کامپیوتری به‌استثناء بالاترین سطح دقت منجر به بازیابی مناسب توانایی نشد.

کلندر (۲۰۱۱) کاربرد آزمون انطباقی کامپیوتری را با استفاده از مطالعه شبیه‌سازی و واقعی بررسی کرد. وی روش برآورد توانایی بیشینه درست‌نمایی[3] (ML) و پسین

---

1. Differential Aptitude Tests (DAT)
2. post-hoc simulations
3. Maximum likelihood (ML)

مورد انتظار[1] (EAP) و ملاک خاتمه آزمون طول ثابت آزمون[2] و خطای استاندارد ثابت[3] را در مطالعه خود در نظر گرفت. در پایان شبیه‌سازی بهترین استراتژی اجرای آزمون انطباقی کامپیوتری مشخص شد و با استفاده از آن آزمون انطباقی کامپیوتری واقعی اجرا شد. نتایج مطالعه نشان داد که همبستگی بین برآوردهای توانایی نسخه انطباقی کامپیوتری و مداد – کاغذی آزمون هنگامی که از روش برآورد پسین مورد انتظار استفاده می‌شود در مقایسه با روش بیشینه درست‌نمایی بیشتر است.

بولوت و کان (۲۰۱۲) نیز کاربرد آزمون انطباقی کامپیوتری را در آزمون‌های ورودی تحصیلات تکمیلی دانشگاه‌های ترکیه با استفاده از روش شبیه‌سازی پس‌تجربی بر اساس روش برآورد پسین مورد انتظار و ملاک‌های خاتمه خطای استاندارد ثابت بررسی کردند. نتایج مطالعه نشان داد که آزمون انطباقی کامپیوتری قادر به بازیابی دقیق در هر یک از خرده آزمون‌ها است. همبستگی بین برآوردهای توانایی آزمون کامل و انطباقی کامپیوتری در تمام خرده آزمون‌ها بالا بود.

همان‌طور که پیک (۲۰۰۵) با مرور مطالعات مقایسه‌پذیری نشان داد، نتایج حاصل از تغییر روش اجرای آزمون‌ها بر ویژگی‌های روان‌سنجی آزمون‌ها و توانایی برآورد شده آزمودنی‌ها متناقض است. تفاوت معنی‌دار بین روش اجرای آزمون‌ها بیشتر در مطالعات اولیه مشاهده می‌شود که دلیل این مسئله می‌تواند تازگی روش سنجش کامپیوتری برای آزمودنی‌ها و ناآشنایی با روش اجرای آزمون‌ها در مطالعات اولیه باشد. همچنین تفاوت مطالعات مقایسه‌پذیری ممکن است به دلیل دامنه وسیعی از تغییرپذیری در آزمون‌ها ازجمله حوزه‌های محتوایی، ویژگی‌های آزمودنی‌ها، طرح‌های جمع‌آوری داده‌ها و شکل آزمون نیز باشد. با توجه به پیشرفت مداوم فناوری کامپیوتر، یافته‌های مطالعات مقایسه‌پذیری قبلی را نمی‌توان به موقعیت‌های مشابه تعمیم داد (پوگیو و همکاران[4]، ۲۰۰۵). تصمیم‌گیری دقیق و شفاف‌سازی نتایج حاصل مستلزم انجام مطالعات دقیق و کنترل شده است، لذا هر مؤسسه آزمون‌سازی که قصد تغییر روش اجرای آزمون‌ها از مداد- کاغذی به کامپیوتری (خطی و انطباقی) را دارد،

---

1. Expected a Posteriori Estimation (EAP)
2. fixed length of test
3. fixed standard error
4. Poggio et al

باید مطالعه مقایسه‌پذیری خاص خود را انجام دهد (پومریچ[1]، ۲۰۰٤؛ تسایی[2] و شین، ۲۰۱۳).

تغییر روش اجرای آزمون از مداد – کاغذی به کامپیوتری موضوعی است که باید در آزمون‌سازی ایران به‌ویژه در مورد آزمون‌های خطیر به‌منظور بهینه‌سازی ساخت، اجرا و نمره‌دهی آزمون‌ها مورد توجه قرار گیرد. ایجاد سازوکارهایی به‌منظور سنجش کارآمد در سطح گسترده و با هزینه قابل قبول امری بسیار مهم است. کاربرد آزمون‌های کامپیوتری و انطباقی کامپیوتری راهکار مناسبی برای تحقق این اهداف است اما مسئله قابل تأمل این است که کاربرد کامپیوتر در سنجش را نم ی‌توان صرفاً بر اساس احساس نیاز و ضرورت حرکت به سوی بهتر شدن انجام داد، بلکه باید با پشتوانه مطالعات کافی و کارآمد، فرایند پرمخاطره جایگزینی آزمون‌های کامپیوتری را سازمان‌دهی و اجرا کرد. بر این اساس لازم است برای انجام مطالعات هدفمند درباره کاربرد سنجش کامپیوتری به‌طور کل و سنجش انطباقی کامپیوتری به‌طور خاص به‌منظور بررسی مشکلات فراروی فرایند تبدیل و بهینه‌سازی روش اجرای آزمون‌های خطیر در مراکز و مؤسسات آزمون‌سازی خصوصاً در سازمان سنجش آموزش کشور به‌عنوان متولی اصلی برگزاری آزمون‌های خطیر، برنامه‌ریزی صورت پذیرد. در راستای این هدف، پژوهش حاضر به بررسی مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد – کاغذی و تعیین الگوریتم بهینه آزمون انطباقی کامپیوتری بر اساس روش‌های مختلف برآورد توانایی (بیشینه درست‌نمایی و پسین مورد انتظار) و ملاک خاتمه آزمون (خطای استاندارد ثابت و طول ثابت آزمون) در خرده آزمون ریاضی آزمون سراسری به‌عنوان آزمون خطیر پرداخته است.

روش پژوهش

پژوهش حاضر با توجه به ماهیت مسئله، کاربرد روش‌های شبیه‌سازی و دست‌کاری متغیرهای مورد بررسی تحقیق تجربی است. گردآوری داده‌ها با استفاده از پاسخ‌های شرکت‌کنندگان آزمون سراسری گروه آزمایشی ریاضی و فنی سال ۱۳۹۲ به سؤال‌های آزمون ریاضی به‌عنوان آزمون خطیر انجام شد. این آزمون به دلیل اهمیت و ماهیت

1. Pommerich
2. Tsai

صفت مورد اندازه‌گیری آن انتخاب شده است. با این همه نتایج پژوهش با تحلیل‌های مشابه قابل تعمیم به سایر خرده آزمون‌های آزمون سراسری و همچنین سایر آزمون‌های خطیر است. پس از حذف الگوهای پاسخ کامل، نمونه تصادفی به حجم ۱۰۰۰ نفر انتخاب شد. مفروضه تک‌بعدی بودن خرده آزمون ریاضی بر اساس روش تحلیل عاملی غیرخطی[1] با استفاده از نرم‌افزار NOHARM4 (مک‌دونالد و فریزر[2]، ۲۰۰۳) بررسی شد. برازش مدل‌های سؤال – پاسخ، ارزیابی و مدرج‌سازی[3] ۵۵ سؤال آزمون ریاضی بر اساس مدل لجستیک سه‌پارامتری با نرم‌افزار Bilog-MG3 (زیموسکی و همکاران[4]، ۲۰۰۳) انجام شد. سپس شبیه‌سازی داده‌ها بر اساس پارامترهای برآوردشده سؤال‌ها و توانایی آزمودنی‌ها با استفاده از نرم‌افزار WinGen (هان[5]، ۲۰۰۷) صورت گرفت. از آنجایی که پاسخ‌های واقعی تبلور کامل آزمودنی‌ها از نظر تجارب آموزشی، فرایندهای شناختی، رفتار آزمون و شرایط روانی و جسمانی زمان آزمون است، کاربرد پارامترهای سؤال‌های واقعی و توانایی آزمودنی‌های واقعی تضمین می‌کند که داده‌های شبیه‌سازی شده تطابق بیشتری با داده‌های واقعی داشته باشند (وانگ، پان و هریس، ۱۹۹۹).

به‌منظور کاهش سوگیری بالقوه چندین مجموعه داده شبیه‌سازی شد. تعداد تکرارهای شبیه‌سازی بستگی به هدف مطالعه، میزان کاهش مورد نظر در واریانس نمونه‌گیری پارامتر و توان[6] مورد نیاز آزمون‌های آماری دارد. هارول و همکاران (۱۹۹۶) حداقل ۲۵ تکرار را توصیه کرده‌اند. در مطالعه حاضر برای کسب اطمینان از حداقل بودن سوگیری نمونه و به‌منظور دستیابی به توان آماری مطلوب ۴۰ تکرار صورت گرفت. پس از شبیه‌سازی داده‌ها مجدداً پارامتر توانایی آزمودنی‌های هر مجموعه برآورد شد. توزیع نمونه‌برداری توانایی برآورد شده آزمودنی‌ها به‌گونه‌ای تنظیم شد که دارای میانگین صفر و انحراف استاندارد یک باشد.

---

1. non-linear factor analysis
2. McDonald& Fraser
3. calibration
4. Zimowski et al
5. Han
6. Power

شبیه‌سازی پس‌تجربی آزمون انطباقی کامپیوتری برای هر یک از ٤٠ مجموعه با نرم نرم‌افزار Firestar (چویی، پودرابسکی و مکینی[1]، ۲۰۱۱) انجام شد. فرایند شبیه‌سازی پس‌تجربی آزمون انطباقی برای هر آزمودنی به‌صورت جداگانه صورت می‌گیرد، به این ترتیب که نرم‌افزار بر اساس قاعده شروع و ملاک گزینش تعیین شده، سؤالی را از بانک سؤال انتخاب و به آزمودنی ارائه می‌کند. سپس نرم‌افزار پاسخ آزمودنی را که بر اساس آزمون مداد- کاغذی شبیه‌سازی شده، بررسی و توانایی آزمودنی را برآورد می‌کند. در ادامه با توجه به پاسخ آزمودنی و ملاک گزینش، سؤال بعدی از بانک انتخاب می‌شود، پاسخ جدید دوباره بررسی 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.

در مطالعه حاضر دو استراتژی روش‌های برآورد توانایی بیشینه درست‌نمایی و پسین مورد انتظار و هم چنین قاعده خاتمه خطای استاندارد ثابت و طول ثابت بررسی شد. بر اساس ملاک خطای استاندارد ثابت، آزمون زمانی خاتمه می‌یابد که خطای استاندارد برآورد توانایی کمتر از مقدار تعیین شده باشد. در ملاک خطای استاندارد ثابت چهار سطح خطای استاندارد ۰/۵، ۰/٤، ۰/۳ و ۰/۲ در نظر گرفته شد. در ملاک طول ثابت سه سطح مختلف آزمون ۲۵، ۳۵ و ٤۵ سؤالی بررسی شد. به این ترتیب برای هر مجموعه داده تولید شده ١٤ شبیه‌سازی انطباقی کامپیوتری انجام شد (دو روش برآورد توانایی × ۷ ملاک خاتمه آزمون) و با توجه به اینکه ٤٠ مجموعه داده تولید شده است در کل ٥٦٠ شبیه‌سازی انطباقی کامپیوتری انجام شد. برای تمام آزمودنی‌ها نقطه شروع آزمون برآورد توانایی صفر در نظر گرفته شد و گزینش سؤال‌ها بر اساس بیشینه آگاهی[2] (MI) صورت گرفت. این قاعده مبتنی بر انتخاب سؤالی با بیشترین تابع آگاهی در سطح توانایی است. روش بیشینه آگاهی به‌طور تکرارشونده سؤال بعدی را به‌گونه‌ای انتخاب می‌کند که بیشینه آگاهی را در سطح توانایی برآورد شده داشته باشد. در جریان آزمون انطباقی کامپیوتری هر سؤال انتخاب شده بیشترین افزایش در آگاهی و بیشترین کاهش در خطای استاندارد را فراهم می‌کند. این قاعده کارآمدی آزمون انطباقی کامپیوتری را با کاهش سریع خطای

1. Choi, Podrabsky & McKinney
2. Maximum Information (MI)

استاندارد برآورد توانایی بیشینه می‌سازد. با توجه به اینکه اگر تعداد سؤال‌های اجرا شده در آزمون کمتر از ۱۰ سؤال باشد روایی آزمون به مخاطره می‌افتد در شبیه‌سازی‌های حاضر حداقل طول آزمون ۱۰ سؤال در نظر گرفته شد (آیهان[1]، ۲۰۱۵). در روش بیشینه درست‌نمایی برای به دست آوردن الگوی پاسخ مرکب (یک پاسخ درست/ یک پاسخ نادرست) اولین سؤال از میان پنج سؤال با دشواری متوسط انتخاب شد، بر اساس پاسخ آزمودنی، سؤال بعدی از میان پنج سؤال از دشوارترین یا ساده‌ترین سؤال‌های بانک انتخاب شد. در صورت ارائه نشدن الگوی مورد نظر، برآورد توانایی آزمودنی بی‌نهایت در نظر گرفته شده است. در مطالعه حاضر محدودیت محتوایی اعمال نشده است چون این پژوهش اولین مطالعه در نوع خود است، بنابراین باید پیش از اعمال محدودیت محتوایی، اطلاعات آماری پایه به دست آید.

مقایسه‌پذیری برآورد پارامتر توانایی خرده آزمون ریاضی بر اساس ملاک روان‌سنجی در سطح آزمون انجام شد. ملاک ویز و گیبونز[2] (۲۰۰۷) برای مقایسه عملکرد آزمون انطباقی کامپیوتری با آزمون مرسوم مداد- کاغذی مورد استفاده قرار گرفت که عبارتند از:

۱- همبستگی برآورد توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی

۲- میانگین تعداد سؤال‌های اجرای انطباقی کامپیوتری برای بازیابی برآورد توانایی با خطای استاندارد تعیین شده

۳- سوگیری[3] (میانگین تفاوت برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی)

٤- دقت[4] (میانگین قدر مطلق تفاوت برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی)

۵- ریشه میانگین مجذور تفاوت[1] (RMSD) برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی

---

1. Ayhan
2. Gibbons
3. bias
4. accuracy

یافته‌های پژوهش

نتایج تحلیل ابعاد آزمون ریاضی بر اساس روش تحلیل عاملی غیرخطی با استفاده از نرم‌افزار NOHARM4 (مک‌دونالد و فریزر، ۲۰۰۳) در جدول (۱) ارائه شده است. برنامه NOHARM4 ریشه میانگین مجذورات پس‌مانده‌ها[2] (RMSR) و شاخص خوبی برازندگی تاناکا[3] (۱۹۹۳) را به‌عنوان شاخص‌های برازش محاسبه می‌کند. مقادیر کوچک ریشه میانگین مجذورات پس‌مانده‌ها بیانگر برازش مدل با داده‌ها است. در خصوص شاخص تاناکا نیز مک‌دونالد (۱۹۹۷) عنوان می‌کند که مقدار ۰/۹۰ بیانگر برازش قابل قبول و ۰/۹۵ بیانگر برازش خوب مدل با داده‌ها است. در صورتی که این شاخص برابر یک باشد نشان‌دهنده برازش کامل است (مینایی و فلسفی‌نژاد، ۱۳۸۹).

در مطالعه حاضر به‌منظور بررسی دقیق، تحلیل بر اساس راه حل تک‌بعدی و دوبعدی انجام شد. همان‌گونه که در جدول (۱) مشاهده می‌شود در راه حل تک‌بعدی مقدار ریشه میانگین مجذورات پس‌مانده‌ها برابر ۰/۰۰۵ و شاخص تاناکا برابر ۰/۹۷ است که بیانگر تک‌بعدی بودن آزمون است. مقدار ریشه میانگین مجذورات پس‌مانده‌ها راه حل دوبعدی ۰/۰۰٤ است که در مقایسه با ریشه میانگین مجذورات پس‌مانده‌ها راه حل تک‌بعدی کاهش ناچیزی را نشان می‌دهد. شاخص تاناکا تحلیل دوبعدی نیز در مقایسه با تحلیل تک‌بعدی افزایش بسیار کمی را نشان می‌دهد. بر این اساس می‌توان نتیجه گرفت که یک خصیصه مکنون به‌وسیله سؤال‌های آزمون ریاضی اندازه‌گیری شده است و تحلیل آزمون بر اساس مدل‌های تک‌بعدی نظریه سؤال- پاسخ امکان‌پذیر است.

جدول (۱) تحلیل عاملی غیرخطی آزمون ریاضی

تعداد ابعاد	ریشه میانگین مجذور باقیمانده‌ها	شاخص برازش تاناکا
تک‌بعدی	۰/۰۰۵	۰/۹۷
دوبعدی	۰/۰۰٤	۰/۹۸

1. Root mean squared difference (RMSD)
2. Root Mean Square Residuals (RMSR)
3. Tanaka Goodness of Fit

برازش مدل‌های لجستیک یک، دو و سه پارامتری سؤال – پاسخ بر اساس شاخص لگاریتم درست‌نمایی[1] بررسی شد. تفاوت لگاریتم درست‌نمایی (2Log Likelihood -) مدل یک و دو پارامتری برابر ۷۸۳/۳۷۵۲ بود که با درجه آزادی df=٥٤ در سطح ۰/۰۱ معنادار بود. تفاوت لگاریتم درست‌نمایی (2Log Likelihood -) مدل دو و سه پارامتری نیز برابر ۹۷۳/٦٤۰۱ که با درجه آزادی df=٥٥ در سطح ۰/۰۱ معنی‌دار بود. بنابراین نتیجه گرفته شد که مدل سه‌پارامتری برازش بهتری با مجموعه داده‌ها دارد و مدرج‌سازی سؤال‌های آزمون ریاضی گروه آزمایشی ریاضی و فنی سال ۱۳۹۲ بر اساس مدل لجستیک سه‌پارامتری انجام شد. در جریان مدرج‌سازی سؤال شماره ۲۹ به دلیل داشتن ضریب همبستگی دورشته‌ای[2] منفی از فرایند تحلیل حذف شد. میانگین و انحراف استاندارد پارامتر تشخیص[3] (a)، دشواری[4] (b) و حدس[5] (c) سؤال‌ها در جدول (۲) ارائه شده است.

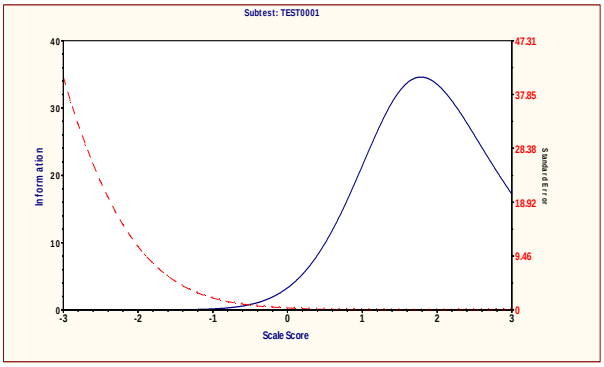
جدول (۲) آماره‌های برآورد پارامترهای سؤال‌های آزمون ریاضی

پارامتر	میانگین	انحراف استاندارد	حداقل	حداکثر
تشخیص(a)	۱/۸۳۸	۰/٤۷٤	۰/٥۹	۲/۸۱
دشواری (b)	۲/۲۹	۱/۲۱	۰/۷	٤>
حدس (c)	۰/۰٤	۰/۰۳	۰/۰۱	۰/۱۲

همان‌گونه که مشاهده می‌شود سؤال‌های خرده آزمون ریاضی بسیار دشوار هستند. بیشتر سؤال‌های آزمون پارامتر تشخیص مطلوبی دارند. تابع آگاهی خرده آزمون ریاضی در شکل (۱) ارائه شده است. بیشینه آگاهی آزمون در سطح توانایی ۱/۷٥= θ است و برابر با ۳٤/٥٥ است. آزمون ریاضی در دامنه توانایی ۱/٦ تا ۲/۲ توانایی آزمودنی‌ها را به‌خوبی برآورد می‌کند چون میزان آگاهی آزمون در این دامنه بالا است. توانایی آزمودنی‌های گروه نمونه بر اساس ٥٤ سؤال آزمون با استفاده از روش بیشینه

1. log-likelihood
2. biserial Correlation
3. discrimination Parameter
4. difficulty
5. guessing

درست‌نمایی و پسین مورد انتظار برآورد شد. میانگین و انحراف استاندارد برآورد پارامتر توانایی آزمودنی‌ها در روش پسین مورد انتظار به ترتیب برابر ۰/۰۱- و ۰/۸۹ و در روش بیشینه درست‌نمایی برابر با ۰/۰۱- و ۱/۱ بود.



شکل (۱) نمودار تابع آگاهی آزمون ریاضی

شبیه‌سازی ٤٠ مجموعه داده بر اساس پارامترهای برآوردشده سؤال‌ها و توانایی آزمودنی‌ها با استفاده از نرم‌افزار WinGen (هان، ۲۰۰۷) انجام شد. در ادامه شبیه‌سازی پس‌تجربی آزمون انطباقی کامپیوتری برای هر یک از مجموعه داده‌ها بر اساس روش برآورد بیشینه درست‌نمایی و پسین مورد انتظار و ملاک خاتمه خطای استاندارد ثابت و طول ثابت صورت گرفت. مقادیر همبستگی، سوگیری، دقت و ریشه میانگین مجذور تفاوت برآورد توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی خرده آزمون ریاضی محاسبه شد. میانگین همبستگی برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی بر اساس هر یک از روش‌های برآورد توانایی و ملاک خاتمه در جدول (۳) ارائه شده است. در انجام محاسبات آزمودنی‌هایی که توانایی آنها در روش بیشینه درست‌نمایی به دلیل نداشتن الگوی پاسخ مرکب (پاسخ درست/ پاسخ نادرست) بی‌نهایت برآورد شده بود از تحلیل کنار گذاشته شدند.

جدول (۳) میانگین همبستگی برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی آزمون ریاضی

روش برآورد توانایی	ملاک خاتمه آزمون						
	خطای استاندارد ثابت				طول ثابت آزمون		
	۰/۲	۰/۳	۰/٤	۰/۵	۲۵	۳۵	٤۵
EAP	۰/۹۸	۰/۹۷	۰/۹٦	۰/۹٦	۰/۹۸	۰/۹۸	۰/۹۸
ML	۰/۹۳	۰/۹۲	۰/۹۲	۰/۹۱	۰/۸۸	۰/۹۲	۰/۹۲

همان‌گونه که مشاهده می‌شود برآورد توانایی حاصل از اجرای انطباقی کامپیوتری و مداد-کاغذی آزمون ریاضی همبستگی بالایی دارند. مقادیر همبستگی‌ها در تمامی شرایط ۰/۸۸ یا بیشتر و از نظر آماری در سطح ۰/۰۱ معنی‌دار بودند. در هر دو روش برآورد توانایی بیشینه درست‌نمایی و پسین مورد انتظار با کاهش مقدار خطای استاندارد (افزایش دقت) همبستگی افزایش یافته است. در ملاک خاتمه طول ثابت آزمون در روش برآورد بیشینه درست‌نمایی با افزایش طول آزمون از ۲۵ سؤال به ۳۵ سؤال، همبستگی افزایش یافته است اما با افزایش طول آزمون به ٤۵ سؤال، میزان همبستگی تغییر نکرده است. در روش پسین مورد انتظار نیز مقادیر همبستگی در ملاک خاتمه طول ثابت نامتغیر است. با نگاهی به جدول (۳) مشخص می‌شود که در تمامی موارد همبستگی بین برآورد توانایی مبتنی بر روش پسین مورد انتظار بیشتر از روش بیشینه درست‌نمایی است.

در جدول‌های (٤ و ۵) میانگین مقادیر سوگیری، دقت و ریشه میانگین مجذور تفاوت برآوردهای توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی آزمون ریاضی برای روش‌های برآورد توانایی و ملاک‌های خاتمه مختلف ارائه شده است. همان‌گونه که مشاهده می‌شود سوگیری منفی در برآورد توانایی در خرده آزمون ریاضی وجود داشته است. مقادیر سوگیری در روش برآورد پسین مورد انتظار کمتر از روش بیشینه درست‌نمایی است. در روش برآورد پسین مورد انتظار با کاهش خطای استاندارد مقدار سوگیری کاهش یافته است. در ملاک خاتمه طول ثابت آزمون در هر دو روش برآورد توانایی با افزایش طول آزمون سوگیری کاهش یافته است. میانگین قدر مطلق تفاوت برآوردهای توانایی آزمون انطباقی کامپیوتری و آزمون مداد- کاغذی در روش برآورد پسین مورد انتظار کمتر از روش بیشینه درست‌نمایی است که بیانگر بیشتر

بودن دقت روش پسین مورد انتظار در برآورد پارامتر توانایی آزمودنی‌ها نسبت به بیشینه درست‌نمایی است. در روش برآورد پسین مورد انتظار با کاهش خطای استاندارد، دقت برآورد پارامتر توانایی افزایش یافته است. در حالی که در روش بیشینه درست‌نمایی الگوی منظمی مشاهده نشده است. در ملاک خاتمه طول ثابت و روش بیشینه درست‌نمایی با افزایش طول آزمون دقت برآورد پارامتر توانایی افزایش یافته است.

در بررسی مقادیر ریشه میانگین مجذور تفاوت برآوردهای توانایی آزمون انطباقی کامپیوتری و مداد- کاغذی نیز مشخص می‌شود که این مقادیر در روش پسین مورد انتظار بر اساس هر دو ملاک خاتمه آزمون از روش بیشینه درست‌نمایی کمتر بوده است که بیانگر مشابهت برآورد توانایی اجرای انطباقی کامپیوتری و مداد – کاغذی آزمون ریاضی در روش برآورد پسین مورد انتظار است. مقادیر ریشه میانگین مجذور تفاوت در رویکرد پسین مورد انتظار در ملاک خاتمه خطای استاندارد ثابت با کاهش خطای استاندارد کاهش یافته است در حالی که در روش بیشینه درست‌نمایی الگوی نامنظمی دارد. مقادیر ریشه میانگین مجذور تفاوت در ملاک خاتمه طول ثابت در هر دو روش برآورد با افزایش طول آزمون کاهش یافته است.

جدول (٤) مقادیر سوگیری، دقت و ریشه میانگین مجذور تفاوت برآوردهای توانایی آزمون ریاضی- ملاک خاتمه خطای استاندارد ثابت

روش برآورد توانایی	ملاک خاتمه خطای استاندارد	سوگیری	دقت	RMSD
EAP	۰/۲	-۰/۱٤۷	۰/۲۳٤	۰/۲۷٦
	۰/۳	-۰/۱٤۸	۰/۲۳۹	۰/۲۸٦
	۰/٤	-۰/۱۵۱	۰/۲۵۱	۰/۳۰٤
	۰/۵	-۰/۱۵۲	۰/۲٦۳	۰/۳۱۳
ML	۰/۲	-۰/۳۱۹	۰/۳۵٤	۰/۵۵۷
	۰/۳	-۰/۳۰۲	۰/٤۰	۰/۵٦
	۰/٤	-۰/۲۸٤	۰/۳۸٦	۰/۵٤۹
	۰/۵	-۰/۲٦۷	۰/۳۸۵	۰/۵٦۱

جدول (۵) مقادیر سوگیری، دقت و ریشه میانگین مجذور تفاوت برآوردهای توانایی آزمون ریاضی - ملاک خاتمه طول ثابت

روش برآورد توانایی	ملاک خاتمه طول آزمون	سوگیری	دقت	RMSD
EAP	۲۵	-۰/۱۶	۰/۲۲۹	۰/۲۷۲
	۳۵	-۰/۱۵۳	۰/۲۲۳	۰/۲۵۸
	۴۵	-۰/۱۵۱	۰/۲۲۵	۰/۲۴۷
ML	۲۵	-۰/۳۳۵	۰/۴۶۵	۰/۷۲۳
	۳۵	-۰/۳۱۸	۰/۴۲	۰/۶۰۲
	۴۵	-۰/۳۰۷	۰/۴۰۶	۰/۵۷۳

با توجه به مقادیر همبستگی می‌توان نتیجه گرفت که برآورد توانایی اجرای انطباقی کامپیوتری آزمون ریاضی در راستای برآورد توانایی آزمون کامل بوده است. مقادیر سوگیری، دقت و ریشه میانگین مجذور تفاوت برآوردهای توانایی آزمون انطباقی کامپیوتری و مداد- کاغذی نیز بیانگر آن است که برآوردهای توانایی اجرای انطباقی کامپیوتری آزمون در روش پسین مورد انتظار برای هر دو ملاک خاتمه خطای استاندارد ثابت و طول ثابت قادر به بازیابی دقیق توانایی مداد - کاغذی بوده است در حالی که برآوردهای توانایی اجرای انطباقی کامپیوتری آزمون در روش بیشینه درست‌نمایی قادر به بازیابی برآورد توانایی آزمون کامل نبوده است. به‌منظور اظهار نظر قطعی در این خصوص یافته‌های مربوط به قاعده خاتمه آزمون در ارتباط با روش‌های برآورد توانایی بررسی شد. در ملاک خاتمه خطای استاندارد، کاهش تعداد سؤال‌ها و در ملاک طول ثابت آزمون مقادیر خطای استاندارد مورد مقایسه قرار گرفت. در جدول (۶) میانگین تعداد سؤال‌های اجرا شده در شبیه‌سازی‌های پس‌تجربی برای هریک از روش‌های برآورد توانایی در سطوح مختلف خطای استاندارد ارائه شده است.

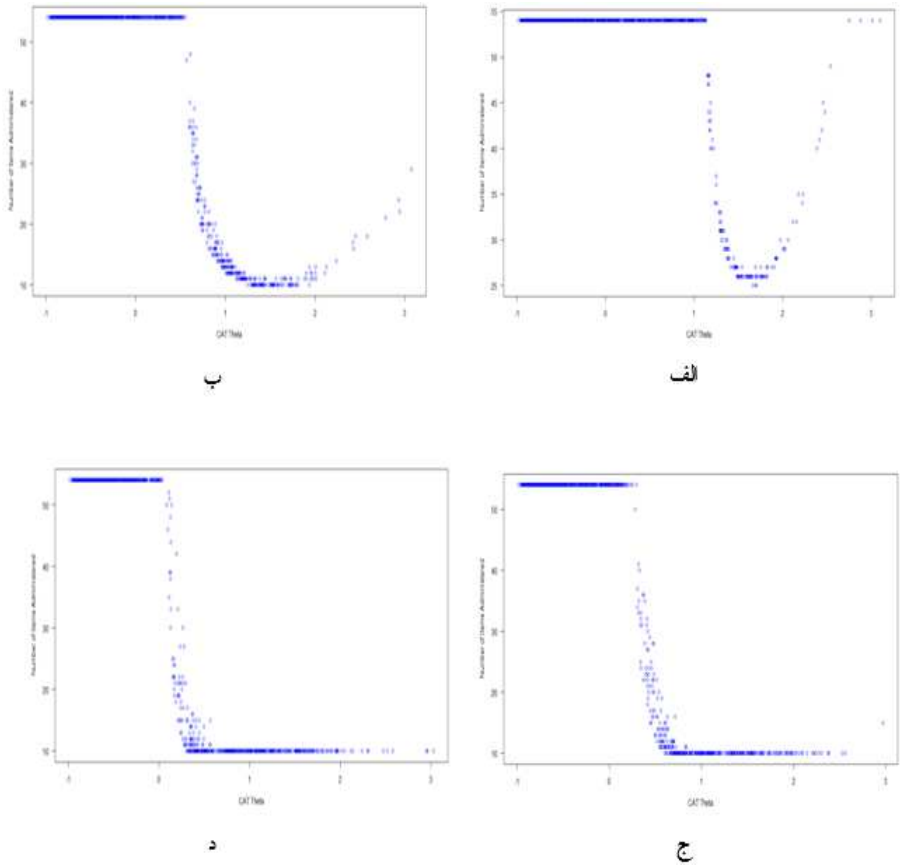
جدول (٦) میانگین تعداد سؤال‌های اجرا شده در شبیه‌سازی پس‌تجربی

روش برآورد توانایی	خطای استاندارد	تعداد سؤال‌ها	
		میانگین	درصد کاهش
EAP	۰/۲	۵۰/۷۸	٪٦
	۰/۳	٤۲/٤۵	٪۲۱
	۰/٤	۳۸/۱۹	٪۲۹
	۰/۵	۳۵/٤٦	٪۳٤
ML	۰/۲	۵۰/٤٦	٪۷
	۰/۳	٤۰/٤٦	٪۲۵
	۰/٤	۳٤/۰۹	٪۳۷
	۰/۵	۲۹/۹۱	٪٤۵

همان‌گونه که مشاهده می‌شود شبیه‌سازی پس‌تجربی روش پسین مورد انتظار در مقایسه با روش بیشینه درست‌نمایی مستلزم اجرای تعداد سؤال بیشتری بوده است. تعداد سؤال‌های مورد نیاز برای مقادیر خطای استاندارد پایین بیشتر از خطای استاندارد بالا است. دستیابی به ملاک خطای استاندارد ۰/۲ در روش بیشینه درست‌نمایی مستلزم اجرای به‌طور متوسط ۵۰ سؤال و در روش پسین مورد انتظار ۵۱ سؤال است. اجرای انطباقی کامپیوتری خرده آزمون ریاضی هنگامی که ملاک خاتمه خطای استاندارد ۰/۲ در نظر گرفته شده است، صرفاً کاهش ۷ درصدی در روش بیشینه درست‌نمایی و ٦ درصدی در روش پسین مورد انتظار را در تعداد سؤال‌های اجرا شده در مقایسه با آزمون مداد-کاغذی به دنبال داشته است که با توجه به هزینه‌ها و زیرساخت‌های لازم اجرای انطباقی کامپیوتری آزمون‌ها این میزان کاهش مطلوب نیست. ملاک خطای استاندارد ۰/۳ منجر به کاهش ۲۵ درصدی تعداد سؤال‌های اجرا شده در روش بیشینه درست‌نمایی و ۲۱ درصدی در روش پسین مورد انتظار شده است. ملاک خطای استاندارد ۰/٤ و ۰/۵ نیز در روش بیشینه درست‌نمایی طول آزمون را به ترتیب ۳۷ و ٤۵ درصد و در روش پسین مورد انتظار طول آزمون را ۲۹ و ۳٤ درصد کاهش داده است. با توجه به مقادیر کاهش در تعداد سؤال‌های اجرا شده، پایایی و همبستگی برآورد پارامتر توانایی اجرای انطباقی کامپیوتری و مداد-کاغذی ملاک خاتمه ۰/۳ و ۰/٤ مناسب به نظر می‌رسند. با لحاظ کردن مقادیر سوگیری و دقت برآورد پارامتر توانایی مشخص می‌شود که ملاک خطای استاندارد

۰/۳ ملاک بهینه برای دستیابی به اهداف پایایی مناسب، طول منطقی آزمون و بازیابی برآورد توانایی در اجرای انطباقی کامپیوتری خرده آزمون ریاضی است.

از آنجایی که در آزمون‌های ورودی دانشگاه‌ها هدف انتخاب توانمندترین افراد است، سؤال‌های آزمون به‌گونه‌ای طراحی می‌شوند که سطوح بالا پیوستار توانایی را به دقت اندازه‌گیری کنند. کاربرد سؤال‌های آزمون ریاضی به‌عنوان بانک سؤال شبیه‌سازی انطباقی کامپیوتری در مطالعه حاضر به دلیل محدود و دشوار بودن سؤال‌ها صرفاً منجر به سنجش دقیق سطوح بالا توانایی شده است. بر این اساس ملاک‌های خاتمه خطای استاندارد ثابت برای آزمودنی‌ها با سطوح توانایی پایین حتی با وجود اجرای تمامی سؤال‌های بانک محقق نشده است. در اجرای واقعی آزمون‌های انطباقی کامپیوتری بانک‌های سؤال بزرگ استفاده می‌شود و به این ترتیب برای تمامی سطوح توانایی سؤال مناسب در بانک موجود است. در شکل (۲) تعداد سؤال‌های اجرا شده در سطوح مختلف توانایی برای ملاک خاتمه خطای استاندارد ثابت ۰/۲، ۰/۳، ۰/٤ و ۰/۵ در روش پسین مورد انتظار ارائه شده است. این نمودارها مربوط به یکی از شبیه‌سازی‌های انجام شده است. همان‌گونه که مشاهده می‌شود در تمام موارد در سطوح پایین پیوستار توانایی تمامی سؤال‌های بانک (٥٤ سؤال) برای آزمودنی‌ها اجرا شده است. درنتیجه در شبیه‌سازی پس‌تجربی خرده آزمون ریاضی کاهش تعداد سؤال‌ها و ارائه سؤال متناسب با سطح توانایی که از اهداف اجرای انطباقی کامپیوتری آزمون‌ها محسوب می‌شوند، صرفاً در مورد آزمودنی‌های با سطوح توانایی متوسط و بالا محقق شده است.



شکل (۲) تعداد سؤال‌های اجرا شده در روش برآورد پسین مورد انتظار: الف) ملاک خاتمه خطای استاندارد ثابت ۰/۲ ب) ملاک خاتمه خطای استاندارد ثابت ۰/۳ ج) ملاک خاتمه خطای استاندارد ثابت ۰/٤ د) ملاک خاتمه خطای استاندارد ثابت ۰/۵

پس از بررسی میزان کاهش تعداد سؤال‌ها در ملاک خاتمه خطای استاندارد ثابت، به شیوه مشابه میزان خطای استاندارد برآورد توانایی در ملاک طول ثابت آزمون، بررسی شد. میانه مقادیر خطای استاندارد برآورد توانایی (به دلیل چولگی توزیع خطای استاندارد برآورد توانایی مقادیر میانه مورد بررسی قرار گرفت) در جدول (۷) ارائه شده است. در هر دو روش برآورد توانایی با افزایش طول آزمون مقدار خطای استاندارد کاهش و دقت برآورد افزایش یافته است. مقدار خطای استاندارد برآورد

روش پسین مورد انتظار نسبت به روش بیشینه درست‌نمایی بیشتر است. در روش برآورد بیشینه درست‌نمایی در آزمون ۳۵ و ٤٥ سؤالی میزان خطای استاندارد به سطح ۰/٤ می‌رسد، در حالی که در روش پسین مورد انتظار حتی در آزمون ٤٥ سؤالی نیز میانه مقادیر خطای استاندارد برآورد زیاد است. این نتایج به دلیل محدود بودن تعداد سؤال‌های بانک از یک سو و دشوار بودن سؤال‌ها از سوی دیگر است که امکان ارائه سؤال متناسب در سطوح پایین پیوستار توانایی را محقق نمی‌سازد. درنتیجه آزمون انطباقی کامپیوتری برآورد دقیقی از توانایی آزمودنی‌های سطوح پایین توانایی ارائه نمی‌کند. بنابراین به‌منظور کسب نتایج دقیق، تحلیل مقادیر خطای استاندارد برآورد توانایی مجدداً با حذف آزمودنی‌های سطوح پایین توانایی انجام شد (جدول ۸). بر این اساس آزمودنی‌ها با توانایی کمتر از صفر از تحلیل کنار گذاشته شدند.

جدول (۷) میانه خطای استاندارد برآورد توانایی

روش برآورد توانایی	طول ثابت آزمون		
	۲۵	۳۵	٤٥
EAP	۰/٦۳	۰/٥٥	۰/٥۱
ML	۰/٤۳	۰/۳۸	۰/۳٦

جدول (۸) میانه خطای استاندارد برآورد توانایی (θ>۰)

روش برآورد توانایی	طول ثابت آزمون		
	۲۵	۳۵	٤٥
EAP	۰/۳	۰/۲۹	۰/۲۸
ML	۰/۲۸	۰/۲٦	۰/۲٦

همان‌گونه که مشاهده می‌شود با حذف آزمودنی‌هایی که برآورد توانایی آنها با دقت صورت نگرفته است، میانه خطای استاندارد برآورد توانایی کاهش یافته است. میانه خطا در هر دو روش برآورد توانایی در سطح ۰/۳ و کمتر است که نشان‌دهنده برآورد دقیق توانایی آزمودنی‌ها است. بررسی مقادیر خطای استاندارد برای آزمون‌ها با طول مختلف نشان می‌دهد که روش بیشینه درست‌نمایی خطای استاندارد کمتری تولید می کند و بنابراین همان‌گونه که انتظار می‌رود مستلزم تعداد سؤال‌های کمتری است. اگرچه تعداد سؤال‌های اجرا شده در روش برآورد پسین مورد انتظار نسبت به روش

بیشینه درست‌نمایی بیشتر است، اما رویکرد پسین مورد انتظار به دلیل مقادیر همبستگی بالا برآورد توانایی اجرای انطباقی کامپیوتری و مداد-کاغذی، بالا بودن دقت برآورد و پایین بودن سوگیری در برآورد توانایی، روش برآورد مطلوب در اجرای انطباقی کامپیوتری آزمون ریاضی آزمون سراسری محسوب می‌شود.

بحث و نتیجه‌گیری

نظر به تمایل فزاینده کاربرد کامپیوترها در سنجش در قالب آزمون‌های کامپیوتری و انطباقی کامپیوتری برای بهینه‌سازی سنجش، بررسی چالش‌های عملی و نظری فرایند تغییر روش اجرای آزمون‌ها اهمیت به‌سزایی یافته است. مطالعه حاضر به‌منظور بررسی مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد – کاغذی خرده آزمون ریاضی آزمون سراسری و تعیین الگوریتم بهینه آزمون انطباقی کامپیوتری بر اساس روش‌های مختلف برآورد توانایی و ملاک خاتمه آزمون با استفاده از روش‌های شبیه‌سازی انجام شد. شبیه‌سازی‌های پس‌تجربی خرده آزمون ریاضی نتایجی در تأیید مقایسه‌پذیری برآورد توانایی اجرای انطباقی کامپیوتری و مداد – کاغذی ارائه کرد. همبستگی برآورد توانایی اجرای مداد – کاغذی و انطباقی کامپیوتری آزمون ریاضی در هر دو روش بیشینه درست‌نمایی و پسین مورد انتظار و تمام سطوح ملاک خاتمه خطای استاندارد ثابت و طول ثابت بالا و معنی‌دار بود که بیانگر همسویی برآورد توانایی اجرای انطباقی کامپیوتری با آزمون کامل است. این یافته با نتایج مطالعات کلندر (۲۰۱۱)، بولت و کان (۲۰۱۲) و آیهان (۲۰۱۵) هم‌خوانی دارد.

روش برآورد توانایی از مؤلفه‌های مهم الگوریتم اجرایی آزمون انطباقی کامپیوتری محسوب می‌شود؛ چون نه‌تنها پیامد نهایی آزمون انطباقی کامپیوتری را تحت تأثیر قرار می‌دهد بلکه فرایند گزینش سؤال‌ها و خاتمه آزمون را نیز متأثر می‌سازد. مقایسه روش‌های برآورد توانایی بیشینه درست‌نمایی و پسین مورد انتظار بر اساس مقادیر همبستگی، سوگیری، دقت و ریشه میانگین مجذور تفاوت برآورد توانایی اجرای انطباقی کامپیوتری و مداد - کاغذی خرده آزمون ریاضی صورت گرفت. مقادیر همبستگی در روش پسین مورد انتظار در تمامی سطوح خطای استاندارد و طول آزمون بیشتر از روش بیشینه درست‌نمایی بود. مقادیر سوگیری، دقت و ریشه میانگین

مجذور تفاوت برآورد توانایی اجرای انطباقی کامپیوتری و مداد- کاغذی خرده آزمون ریاضی در روش پسین مورد انتظار بر اساس هر دو ملاک خاتمه آزمون کمتر از روش بیشینه درست‌نمایی بود که بیانگر مشابهت بیشتر برآورد توانایی اجرای انطباقی کامپیوتری و مداد – کاغذی آزمون ریاضی است. با توجه به یافته‌های حاصل می‌توان نتیجه گرفت که روش پسین مورد انتظار منجر به بازیابی دقیق‌تر برآورد توانایی در چهارچوب اجرای انطباقی کامپیوتری شده است در حالی که روش بیشینه درست‌نمایی قادر به بازیابی دقیق توانایی نبوده است. به این ترتیب رویکرد پسین مورد انتظار روش مطلوب برآورد توانایی در اجرای انطباقی کامپیوتری خرده آزمون ریاضی آزمون سراسری محسوب می‌شود. این یافته با نتایج مطالعه کلندر(۲۰۱۱) و آیهان (۲۰۱۵) هم‌خوانی ندارد آنها در مطالعه خود نشان دادند که هر دو روش برآورد توانایی پسین مورد انتظار و بیشینه درست‌نمایی در چهارچوب آزمون انطباقی کامپیوتری مناسب است. البته در هر دو مطالعه، مقایسه روش‌ها صرفاً بر اساس مقادیر همبستگی صورت گرفته است در حالی که مقایسه دقیق و جامع روش‌های برآورد توانایی مستلزم به‌کارگیری ملاک‌های چندگانه است. برتری روش پسین مورد انتظار در اجرای انطباقی خرده آزمون ریاضی با یافته‌های مطالعه باک و میسلوی (۱۹۸۲) و واینر و همکاران[1] (۲۰۰۰) همسو است. به نظر می‌رسد مطلوبیت روش پسین مورد انتظار به دلیل فرض توزیع پیشین توانایی آزمودنی‌ها است که امکان برآورد دقیق‌تر توانایی و درنتیجه بازیابی بهتر توانایی را به دلیل فراهم ساختن اطلاعات بیشتر در اجرای انطباقی کامپیوتری آزمون در مقایسه با روش بیشینه درست‌نمایی ممکن می‌سازد (کلندر، ۲۰۱۱). روش برآورد توانایی پسین مورد انتظار برخلاف روش بیشینه درست‌نمایی مبتنی بر فرایند تکرارشونده نیست و از برآوردگر فرم بسته استفاده می‌کند. به‌علاوه روش پسین مورد انتظار امکان ارائه برآورد توانایی آزمودنی‌ها با الگوی پاسخ کامل را فراهم می‌سازد در حالی که روش درست‌نمایی قادر به برآورد توانایی آزمودنی‌ها در الگوی پاسخ کامل نیست. این ویژگی‌ها برتری روش پسین مورد انتظار را در چهارچوب آزمون‌های انطباقی کامپیوتری توجیه می‌کند. در آزمون‌های خطیر مانند آزمون سراسری احتمال محقق نشدن الگوی مرکب (یک پاسخ

[1]. Wainer et al

درست و یک پاسخ نادرست) وجود دارد درنتیجه روش برآورد پسین مورد انتظار مناسب‌تر است.

بررسی ادبیات پژوهشی در خصوص ملاک خاتمه سنجش انطباقی کامپیوتری مؤید مطلوبیت ملاک خطای استاندارد ثابت است (بابکوک و ویز، ۲۰۰۰؛ استوکینگ[1]، ۱۹۸۷)، چون این ملاک تضمین‌کننده برآورد پایای توانایی آزمودنی‌ها است. هنگامی که ملاک خاتمه خطای استاندارد ثابت استفاده می‌شود، تعداد سؤال‌های مورد نیاز برای برآورد توانایی اهمیت می‌یابد چون برآورد پایای توانایی بدون هیچ‌گونه کاهشی در تعداد سؤال‌ها مطلوب و منطقی نیست. از سوی دیگر ملاک خاتمه طول ثابت آزمون مانع از اجرای تعداد زیادی سؤال برای آزمودنی‌ها می‌شود اما در این حالت ممکن است پس از ارائه تمامی سؤال‌ها سطح توانایی آزمودنی با پایایی مورد نظر برآورد نشود. نتایج مطالعه حاضر بیانگر کاهش تعداد سؤال‌های اجرا شده در اجرای انطباقی کامپیوتری آزمون، در تمام سطوح خطای استاندارد در هر دو روش برآورد توانایی بیشینه درست‌نمایی و پسین مورد انتظار بود، البته در سطح خطای استاندارد ۰/۲ در هر دو روش برآورد توانایی مقدار کاهش خیلی کم بود که با توجه به هزینه‌ها و زیرساخت‌های لازم برای اجرای انطباقی کامپیوتری آزمون‌ها این میزان کاهش مطلوب نیست و درنتیجه کاربرد این ملاک در اجرای انطباقی کامپیوتری آزمون‌های خطیر گزینه منطقی محسوب نمی‌شود. ملاک خطای استاندارد ۰/۳ توافق مطلوبی میان کاهش تعداد سؤال‌های اجرا شده، بازیابی دقیق توانایی آزمودنی‌ها، سوگیری و دقت برآورد پارامتر توانایی ایجاد کرده است. بنابراین به نظر می‌رسد ملاک خاتمه بهینه در اجرای انطباقی کامپیوتری آزمون ریاضی باشد. بابکوک و ویز (۲۰۰۹) نیز تأکید داشتند که در آزمون‌های انطباقی کامپیوتری به‌منظور سنجش دقیق توانایی از نظر سوگیری و ریشه میانگین مجذور خطا، ملاک خطای استاندارد برابر یا کوچک‌تر از ۰/۳ مورد استفاده قرار گیرد. نتایج مطالعات کلندر (۲۰۱۱)، بولت و کان (۲۰۱۲) و آیهان (۲۰۱۵) نیز مبین برتری ملاک خطای استاندارد ۰/۳ است. در ملاک خاتمه طول ثابت نیز نتایج مطالعه بیانگر مطلوبیت آزمون ۳۵ سؤالی در بازیابی برآورد توانایی در اجرای انطباقی کامپیوتری خرده آزمون ریاضی بود. مقایسه دو ملاک خاتمه خطای

---

1. Stocking

استاندارد ثابت و طول ثابت آزمون نشان داد که ملاک خطای استاندارد ثابت اگرچه مستلزم اجرای تعداد سؤال‌های بیشتری است اما به دلیل تولید برآوردهای توانایی پایاتر برای اهداف اجرای انطباقی کامپیوتری آزمون مناسب‌تر است. به‌طور کلی نتایج مطالعه نشان داد که روش برآورد پسین مورد انتظار و ملاک خاتمه خطای استاندارد ثابت ۰/۳ الگوریتم بهینه دستیابی به اهداف پایایی مناسب، طول منطقی آزمون و بازیابی برآورد توانایی در اجرای انطباقی کامپیوتری خرده آزمون ریاضی است.

یکی از مزایای اصلی اجرای انطباقی کامپیوتری آزمون‌ها مبنی بر پایایی بیشتر با تعداد سؤال‌های کمتر در مطالعه حاضر تأیید شد. نتایج مطالعه نشان داد که اجرای انطباقی کامپیوتری خرده آزمون ریاضی قادر به بازیابی توانایی آزمودنی‌ها با تعداد سؤال‌های کمتر نسبت به شکل کامل آزمون است. اگرچه آزمودنی‌ها با توانایی پایین هنوز می بایست به تمام سؤال‌های آزمون پاسخ می‌دادند، آزمودنی‌های توانمند با تعداد سؤال‌های کمتری و با دقت بیشتری اندازه‌گیری شدند. عدم تحقق امکان سنجش دقیق آزمودنی‌های سطوح پایین توانایی در مطالعه حاضر به دلیل محدودیت تعداد سؤال‌های بانک بود که منجر شد سؤال‌های بانک فقط در دامنه خاصی از پیوستار توانایی (سطوح بالا) آگاهی‌دهنده باشند. در سطوح پایین توانایی به جای سؤال‌های بهینه، سؤال‌های کمتر از حد مطلوب اجرا شد و درنتیجه دستیابی به سطح دقت از پیش تعیین شده میسر نشده است. بی‌تردید تهیه بانک سؤال بهینه از مهم‌ترین فعالیت‌ها در اجرای انطباقی کامپیوتری آزمون‌های خطیر محسوب می‌شود. به‌منظور فراهم ساختن امکان سنجش دقیق و کارآمد آزمودنی‌ها در سراسر پیوستار توانایی، بانک سؤال باید دارای تعداد سؤال مکفی با کیفیت بالا باشد و سؤال‌ها نیز به گونه مناسبی در سراسر پیوستار توانایی توزیع شده باشد تا الگوریتم آزمون انطباقی کامپیوتری قادر به انتخاب مناسب‌ترین سؤال برای آزمودنی‌های سطوح مختلف توانایی باشد. به‌علاوه تأمین امنیت بانک سؤال و کنترل میزان افشای سؤال‌ها در جریان آزمون انطباقی کامپیوتری نیز مستلزم وجود بانک سؤال بزرگ است.

ذکر این نکته ضروری است که در مطالعه حاضر قابلیت اجرای انطباقی خرده آزمون ریاضی به‌عنوان آزمون خطیر از منظر روان‌سنجی، بررسی شده است. تبدیل روش اجرای آزمون‌های خطیر ازجمله آزمون‌های ورودی از شکل مداد- کاغذی به انطباقی کامپیوتری مستلزم مطالعات جامع و گسترده در سایر حوزه‌ها ازجمله منابع مالی مورد

نیاز، تجهیزات و زیرساخت‌های لازم، میزان آمادگی جامعه برای پذیرش آزمون‌های کامپیوتری، تهیه بانک سؤال، ارزیابی عادلانه بودن آزمون و امنیت آزمون است که لازم است پیش از هرگونه تصمیم‌گیری در خصوص تغییر روش اجرای آزمون‌ها برای انجام مطالعات مرتبط، برنامه‌ریزی شود.

در مطالعه حاضر مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد کاغذی بر اساس روش شبیه‌سازی تجربی و با استفاده از پارامترهای برآورد شده سؤال‌ها و توانایی آزمودنی‌ها واقعی انجام شده است. اگرچه کاربرد پارامترهای واقعی سؤال‌ها و توانایی آزمودنی‌های تضمین می‌کند که داده‌های شبیه‌سازی شده تطابق بیشتری با داده‌های واقعی داشته باشند اما در اجرای آزمون انطباقی کامپیوتری واقعی عوامل مختلفی ازجمله آشنایی با کامپیوتر، اضطراب کار با کامپیوتر و حتی ویژگی‌های شخصیتی آزمودنی‌ها پاسخ‌گویی به سؤال‌ها را تحت تأثیر قرار می‌دهد. در مطالعه حاضر با توجه به هدف پژوهش این متغیرها مورد توجه قرار نگرفته است لذا در تعبیر، تفسیر و تعمیم نتایج، این نکات باید در نظر گرفته شود.

یکی از چالش‌های عملی کاربرد سنجش انطباقی کامپیوتری در برآورد توانایی مربوط به احتمال افشای سؤال‌ها در اجرای آزمون انطباقی کامپیوتری برای افراد مختلف است. این مسئله تا حد زیادی تحت تأثیر حجم بانک سؤال، تعداد آزمودنی‌ها و تعداد دفعات اجرای برنامه آزمون انطباقی کامپیوتری است. در پژوهش حاضر این عامل مدنظر قرار نگرفته است لذا بر اساس یافته‌های این مطالعه نمی‌توان در مورد مخاطراتی که ممکن است از این ناحیه کلیت سنجش انطباقی کامپیوتری را تحت‌الشعاع خود قرار دهد، اظهار نظر کرد. پیشنهاد می‌شود در مطالعات بعدی بررسی مقایسه‌پذیری برآورد توانایی آزمون انطباقی کامپیوتری و آزمون مداد – کاغذی با استفاده از بانک سؤال بهینه و با اعمال محدودیت محتوایی و کنترل میزان افشای سؤال‌های بانک انجام شود.

منابع

بابایی، محمود (۱۳۸۹). *مقدمه‌ای بر یادگیری الکترونیکی*. تهران: انتشارات پژوهشگاه علوم و فن‌آوری اطلاعات ایران، نشر چاپار.

مینایی، اصغر و فلسفی‌نژاد، محمدرضا (۱۳۸۹). روش‌های سنجش تک‌بعدی بودن سؤال‌ها در مدل‌های دو ارزشی IRT. *فصلنامه اندازه‌گیری تربیتی*، ۱ (۳)، ۷۱ – ۱۰۰.

Alkhadher, O.; Clarke, D. D. & Anderson, N. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the Differential Aptitude Tests. *Journal of occupational and organizational psychology*, *71* (3), 205-217.

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.

Ayhan, A. S. (2015). *Comparability of Scores from CAT and Paper Pencil Implementations of Students Selection Examination to Higher Education*. Master's thesis, Ihsan Doğramacı Bilkent University.

Babcock, B. & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In *Proceedings of the 2009 GMAC conference on computerized adaptive testing* (Vol. 14).

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Computer-based testing and the Internet*. 201-217.

Bergstrom, B. A. (1992). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. ERIC Clearinghouse.

Bock, R. D. & Mislevy. R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement, 6*, 431-444.

Bringsjord, EL. (2001). *Computer-Adaptive Versus Paper-and-Pencil Testing Environments: An Experimental Analysis of*

Examinee Experience. Doctoral dissertation, University at Albany.

Bulut, O. & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*. 49. 61-80.

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.

Davey, T. & Pitoniak, M. J. (2006). Designing computerized adaptive tests. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In *Proceedings of the 2007 GMAC conference on computerized adaptive testing*

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.

Henly, S. J.; Klebe, K. J.; McBride, J. R. & Cudeck, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement*. 13 (4). 363-371.

Joubert, T., & Kriek, H. J. (2009). Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting. *SA Journal of Industrial Psychology, 35* (1), 78-88.

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

McDonald, R. P. & Fraser, C. (2003). *NOHARM: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Niagara College. Welland, Ontario. Retrieved from <http://noharm.niagararesearch.ca/nh4man/nhman.html>

Paek, P. (2005). Recent trends in comparability studies. *Pearson Educational Measurement*. Retrieved February, 1, 2010.

poggio, J.; Glasnapp, D. R.; Yang, X. & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning and Assessment, 3* (6).

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment*, 2 (6), 1-44.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, 52 (2), 127-141.

Ponsoda, V. (2000). Overview of the computerized adaptive testing special section. *Psicológica*, *21*, 115-120.

Schaeffer, G. A.; Bridgeman, B.; Golub-Smith, M. L.; Lewis, C.; Potenza, M. T. & Steffen, M. (1998). Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE" General Test. *ETS Research Report Series*. 2. i-25.

Schaeffer, G. A.; Reese, C. M.; Steffen, M.; McKinley, R. L. & Mills, C. N. (1993). Field Test of a Computer-Based GRE General Test. *ETS Research Report Series*. 1. i-47.

Schaeffer, G. A.; Steffen, M.; Golub-Smith, M. L.; Mills, C. N. & Durso, R. (1995). The introduction and comparability of the computer adaptive GRE general test. *ETS Research Report Series*, 1, i-48.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review*, 36, 263-277.

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Retrieved from <http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literature_review_of_comparability_report.pdf

Tsai, T. H., & Shin, C. D. (2013). A Score Comparability Study for the NBDHE Paper–Pencil versus Computer Versions. *Evaluation & the health professions*, *36* (2), 228-239.

Wainer, H.; Dorans, N. J.; Flaugher, R.; Green, B. F. & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurement and Research Service Bulletin*, *13*, 1-7.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38 (1), 19-49.

Wang, X. B.; Pan, W. & Harris, V. (1999). *Computerized Adaptive Testing Simulations Using Real Test Taker Responses*. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*, URL. http://www.psych.umn.edu/psylabs/catcentral/pdf files/cat07weiss&gibbons.pdf

Zimowski, M. F.; Muraki, E.; Mislevy, R. J. & Bock, R. D. (2003). *BILOG-MG3: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.

استناد به این مقاله:

شریفی یگانه، نگار؛ فلسفی‌نژاد، محمدرضا؛ دلاور، علی؛ فرخی، نورعلی و جمالی، احسان (۱۳۹۵). تعیین مقایسه‌پذیری برآورد پارامتر توانایی در سنجش انطباقی کامپیوتری و مداد-کاغذی. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ٦ (١٤)، ۲۰۳ – ۲۳٤.