Journal of Educational Measurement & Evaluation Studies
Vol.10, No. 29, Spring 2020

فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال دهم، شماره ۲۹، بهار ۱۳۹۹، صص ۱٤٤-٦۱

طراحی و کاربرد روش سنجش انطباقی کامپیوتری برای اجرای آزمون تولیمو در سازمان سنجش آموزش کشور[۱]

مریم مقدسین*

تاریخ دریافت مقاله: ۹۷/۱۲/۲۷
تاریخ پذیرش مقاله: ۹۸/۰۴/۰۵

چکیده

سنجش مهارت زبان انگلیسی در ارزیابی‌های خطیر به تعداد زیادی پرسش برای آزمون‌هایی به‌شیوه مداد-کاغذی (P&P) نیاز دارد؛ زیرا هر سال افراد بسیاری در این نوع آزمون‌های سرنوشت‌ساز شرکت می‌کنند. هدف از اجرای این پژوهش، طراحی و کاربرد سنجش انطباقی کامپیوتری (CAT) به‌عنوان گزینه‌ای برای سنجش مهارت زبان انگلیسی در سازمان سنجش آموزش کشور بود. CAT برخلاف آزمون‌های سنتی P&P که توالی گزینش پرسش‌ها در آن ثابت و یکنواخت است، از یک شیوه گزینش سؤال بهینه و انطباقی استفاده می‌کند. CAT، برآورد توانایی موقت را به‌طور بهینه مورد هدف قرار می‌دهد و به یک ملاک همگرایی مناسب برای برآورد توانایی می‌رسد و در نتیجه، به یک فرایند سنجش کوتاه‌تر، قابل اطمینان‌تر و کارآمدتر می‌رسد. مورد مطالعه مهارت زبان انگلیسی در مطالعه حاضر، آزمون تولیمو است. این پژوهش در دو مرحله اجرا شده است: در مرحله اول، نمونه‌ای از اجراهای مداد-کاغذی آزمون تولیمو (دوره ۱۱٤ تا ۱۲۳)، انتخاب و سپس، سؤال‌ها و توانایی آزمودنی‌ها مدرج‌سازی شد. در مرحله دوم، CATهای بهینه شبیه‌سازی شده به‌عنوان مبنایی برای ارزیابی صحت و کارایی CAT عملیاتی طراحی شدند. نتایج پژوهش نشان داد که برآورد پارامتر توانایی به روش بیشینه درست‌نمایی و ملاک توقف طول ثابت، بیشترین دقت در برآورد پارامتر توانایی آزمودنی‌ها را ایجاد می‌کنند. همچنین، آزمون CAT تولیمو که بر اساس خزانه سؤال بهینه شبیه‌سازی‌شدند، نسبت به CAT عملیاتی که بر اساس خزانه سؤال موجود طراحی شده است، به سؤال‌های کمتری نیاز دارد، درحالی‌که به نتایج دقیق‌تری نسبت به CAT عملیاتی در برآورد پارامتر توانایی منجر می‌شود. بنابراین، باوجود مناسب بودن سؤال‌های موجود در خزانه سؤال مدرج‌سازی شده آزمون تولیمو، طراحی سؤال‌هایی برای هدف اجرای آزمون به شیوه CAT، هم به‌صرفه‌تر است و هم دقیق‌تر پارامتر توانایی را برآورد می‌کند. همچنین، نتایج این مطالعه نشان داد که آزمون تولیمو هم به شیوه CAT شبیه‌سازی شده و CAT عملیاتی طراحی شده به شکلی کارآتر و دقیق‌تر نسبت به تولیمو مداد-کاغذی عمل می‌کند. یافته‌های این پژوهش، نشان می‌دهد که آزمون CAT تولیمو دارای پتانسیل بالایی در کارایی و دقت اندازه‌گیری توانایی زبان انگلیسی است.

واژگان کلیدی: سنجش انطباقی کامپیوتری (CAT)، آزمون تولیمو، برآورد پارامتر توانایی، خزانه سؤال، تعادل محتوایی و مواجهه سؤال

---

[۱]. این طرح با حمایت مالی سازمان سنجش آموزش کشور و از محل اعتبارات پژوهشی این سازمان تأمین اعتبار شده است.

* استادیار گروه روان‌شناسی بالینی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه خوارزمی، تهران، ایران (نویسنده مسئول: mmoghadasin@yahoo.com)

مقدمه

سه دهه است که سنجش انطباقی کامپیوتری[1](CAT) ایجاد شده است. درواقع ایده اصلی CAT اجرای سؤال‌هایی است که برای توانایی هر آزمودنی کاملاً متناسب و مناسب باشد (ریکیسی[2] ۱۹۸۹). در سنجش انطباقی کامپیوتری (CAT)، سؤال‌ها به صورت متغیر و چرخشی[3] بر اساس سطح توانایی آزمودنی انتخاب می‌شوند. برای انتخاب سؤال‌ها با مشخصات بهینه اندازه‌گیری در سطح توانایی برآورد شده آزمودنی، روش‌های انطباقی به‌کار می‌رود. CAT دارای مزیت‌های مشابهی است همانند آزمون‌های دیگری که مبتنی بر کامپیوتر هستند. این مزیت‌ها عبارت‌اند از؛ افزایش انعطاف و ارتباط با سیستم‌های اجرایی آزمون. افزون بر این، دو مزیت دیگر وجود دارد که تنها مختص CAT هستند، یکی از این مزیت‌ها این است که طول تست می‌تواند تا تقریباً ٤٠ درصد کم شود، بدون آنکه دقت اندازه‌گیری کاهش یابد، مزیت دیگر این است که، آزمودنی‌ها با دریافت پرسش‌هایی که برای آنها خیلی آسان یا خیلی سخت است، نا امید نمی‌شوند (واینر و همکاران[4] ۱۹۹۰؛ ون‌در لیندن و گلاس[5] ۲۰۰۰). سنجش انطباقی کامپیوتری در اصل ادغام تست‌های کامپیوتری و سنجش انطباقی است (چانگ[6] ۲۰۱۵)

در مجموع، CAT تئوری آزمون‌سازی و همچنین، تکنولوژی را به منظور تصمیم‌گیری بهتر برای اندازه‌گیری، با یکدیگر تلفیق می‌کند. CAT سه مؤلفه بسیار مهم و کلی در همه اجراهای خود دارد: اول، نقطه شروع آزمون؛ دوم، انتخاب سؤال‌های متوالی و برآورد توانایی آزمودنی و سوم قاعده اتمام آزمون[7](چانگ، ۲۰۰٤؛ ۲۰۱٤؛ فانکوک‌کرواد[8] ۲۰۱۲). سنجش انطباقی کامپیوتری نوعی سنجش مناسب[9]است، زیرا سؤال‌ها بر اساس توانایی آزمودنی‌ها انتخاب می‌شوند. در این نوع

---

1. Computerized Adaptive Testing
2. Reckase
3. on the fly
4. Wainer et al
5. Van der Linden & Glas
6. Chang
7. Termination criterion
8. Phankokkruad
9. tailored

سنجش، آزمون با سؤالی آغاز می‌شود که درجه دشواری متوسط دارد، اگر آزمودنی به آن سؤال، پاسخ درست داد، سؤال بعدی که برای اجرا انتخاب می‌شود، دشوارتر خواهد بود، اگر پاسخ غلط دهد، سؤال آسان‌تری برای اجرا انتخاب می‌شود. سنجش تا زمانی ادامه می‌یابد که ملاک توقف آزمون برقرار شود (کانجاناوسی[1] ۲۰۱۲).

برای ساخت و طراحی یک CAT بهینه، علاوه بر سه مؤلفه ذکر شده، مؤلفه‌های دیگری نیز اهمیت دارد. مدل نظریه سؤال پاسخ[2] مناسبی که سؤال‌ها بر اساس آن مدرج می‌شوند (دی‌آیالا[3] ۲۰۰۹)، خزانه سؤال مدرج شده[4] (ریکیسی، ۱۹۸۹)، کنترل مواجهه سؤال[5] و تعادل محتوایی[6] آزمون. این دو مؤلفه اخیر، در انتخاب سؤال محدودیت‌هایی وارد می‌کنند، به‌طوری‌که سؤال‌هایی انتخاب می‌شوند که نه‌تنها ویژگی‌های آماری‌شان بلکه ویژگی‌های محتوایی و امنیت آنها نیز مدنظر باشد (برگستروم و لانز[7] ۱۹۹۹). لازم به ذکر است که سیستم‌های CAT از نظر تئوریکی، مبتنی بر خصوصیات تئوری سؤال پاسخ (IRT) است. در IRT پارامترهای توانایی و سؤال جدا از یکدیگرند. فرض بر این است که پارامترهای سؤال، برای مقادیر متفاوت پارامترهای توانایی نامتغیر هستند؛ بنابراین، سؤال‌ها می‌توانند مدرج شوند و پارامترهای سؤال نیز می‌توانند در خزانه سؤال مدرج شوند. از داخل همین خزانه‌ها، سؤال‌هایی انتخاب می‌شوند که حداکثر آگاهی در پارامتر توانایی برآورد شده را فراهم می‌کنند (ولدکمپ، ورشور و ایگن، ۲۰۰۷). همچنین، اخیراً به‌دلیل آنکه روش‌های CAT برای سنجش‌های سرنوشت‌ساز فراوان به کار می‌روند، تعادل محتوایی (چنگ و چانگ[8] ۲۰۰۹) و قواعد امنیتی، مانند کنترل مواجهه سؤال[9] (سیمپسون و هتر[10] ۱۹۸۵) از اهمیت به‌سزایی برخوردار می‌شوند. از دیگر مؤلفه‌های مهم CAT که پژوهش‌های محدودی به آن اختصاص یافته، خزانه سؤال است. ویژگی‌های جذاب

---

1. Kanjanawasee
2. Item response theory model
3. De Ayala
4. Calibrated Item pool
5. item exposure
6. content balancing
7. Bergstrom & Lunz
8. Cheng & Chang
9. Exposure Control
10. Sympson & Hetter

CAT، در صورتی تحقق می‌یابند که سؤال‌های موجود در خزانه‌ای که برای اجرا به کار می‌روند، مناسب باشند (ریکیسی، ۲۰۱۰). مسلم است که افزایش کیفیت خزانه سؤال، نحوه عملکرد الگوریتم‌های سنجش انطباقی را بهبود می‌بخشد. بهترین و حتی جذاب‌ترین برنامه‌های سنجش انطباقی، اگر بر اساس خزانه سؤال محدود و سؤال‌هایی که کیفیت ضعیفی دارند بنا شوند، مطلوب نخواهند بود (فلاگر[1] به نقل از واینر و همکاران، ۲۰۰۰). بنابراین، سنجش انطباقی کامپیوتری، به خزانه سؤالی نیاز دارد که خوب طراحی شده و شامل تعداد مناسبی از سؤال‌ها باشد، به‌طوری که آزمون‌های مجزایی برای مطابقت با سطوح توانایی آزمودنی‌ها فراهم سازد. یک خزانه سؤال بهینه باید شامل سؤال‌هایی با تعادل محتوایی مناسب که به استفاده بهینه از سؤال منجر شود، باشد؛ خزانه‌ای که هزینه طراحی سؤال را کاهش دهد (گو[2] و ریکیسی، ۲۰۰۷؛ مقدسین، فلسفی‌نژاد، دلاور، جمالی و فرخی، ١٣٩٤). هنگامی خزانه سؤال بهینه خواهیم داشت که هر زمان الگوریتم انتخاب سؤال CAT، سؤالی را برای اجرا جستجو می‌کند، دقیقاً همان سؤالی که مطلوب و مورد نظر ما است، در خزانه سؤال موجود باشد (ریکیسی، ۲۰۱۰). خزانه‌های سؤال در سنجش انطباقی باید به پیش‌فرض‌های مدل روان‌سنجی که زیربنای مدرج‌سازی، اجرا و نمره‌گذاری است، توجه کنند. بنابراین، تلاشی که برای نوشتن خزانه سؤال‌های سنجش انطباقی لازم است، بسیار بیشتر از آزمون‌های مداد-کاغذی است (میلمن و آرتر[3] ١٩٨٤). خزانه سؤال بهینه باید بر اساس مؤلفه‌های دیگر CAT، یعنی طول آزمون، توزیع مورد انتظار توانایی در جامعه آزمودنی‌ها، برآورد توانایی، شیوه‌های انتخاب سؤال و نسبت‌های مواجهه و نرخ همپوشی سؤال (آزمون) نیز تعیین شود. توجه به تمام مؤلفه‌های CAT که توسط ریکیسی (۱۹۸۹) تعیین شده، به‌طور هم‌زمان الزامی است (برگستروم و لانز، ۱۹۹۹). به دلیل اینکه هدف سنجش انطباقی کامپیوتری بالا بردن آگاهی کل تست‌های سرهم‌بندی شده از طریق انتخاب سؤال با آگاهی بالا از بانک سؤال مدرج‌شده است (که به‌طور ایدئالی با یک مدل مناسب مدرج‌سازی شده است) (ون‌درلیندن و پاشلی، ۲۰۱۰)، بنابراین ممکن است در اجراهای اولیه سؤال‌های بسیار

---

[1]. Flaugher
[2]. Gu
[3]. Millman & Arter

مناسب زودتر انتخاب شوند و مواجهه سؤال را در اجراهای بعدی به بیشینه برسانند. این قضیه می‌تواند به مواجهه بیش از حد سؤال‌ها و پایین آمدن دقت اجرای CAT در برآورد توانایی آزمودنی منجر شود و به دنبال آن موجب لو رفتن سؤال‌های مناسب شود که به‌نوبه خود امنیت آزمون را پایین می‌آورد. از این‌رو، در آزمون‌های خطیر و سرنوشت‌ساز که امنیت آزمون از اهمیت بسزایی برخوردار است باید از روش کنترل مواجهه سؤال مناسب استفاده شود (سیمسون و هتر، ۱۹۸۵). در حال حاضر، CAT به‌طور گسترده‌ای برای اجرای آزمون‌های سرنوشت‌سازی[1] همچون آزمون سنجش توانایی‌ها و معلومات فارغ‌التحصیلان[2](GRE) و آزمون استعداد تحصیلی ویژه رشته مدیریت[3](GMAT) به‌کار می‌رود (چانگ، ۲۰۰٤). به دلیل اینکه این نوع آزمون‌ها ملاکی برای انتخاب و تصمیم‌گیری در مورد آینده آزمودنی‌ها هستند، باید در اجراهای مداد-کاغذی سنتی این نوع آزمون‌ها اعتبار و روایی اجراها در طول زمان رعایت شود. استفاده CAT این نوع آزمون‌ها، مسائلی دارند که گاهی روایی و اعتبار آزمون‌ها را به خطر می‌اندازند، این مسائل همه معطوف به امنیت آزمون است (چانگ، تویی[4]، ۱۹۹۸؛ دیوی و نرینگ[5] ۲۰۰۲؛ بارادا، اولا، پانسودا و اباد[6] ۲۰۰۹؛ فرنچ و تامسون[7] ۲۰۰۳). این واضح است که کاربرد CAT این نوع آزمون‌ها به خزانه سؤال بسیار بزرگی نیاز دارد، امّا تنها مجموعه‌ای از سؤال‌ها ویژگی‌های بسیار مناسبی دارند و به‌طور فراوان استفاده می‌شوند. در چنین، موقعیتی احتمال به‌خاطر سپردن پاسخ‌ها بالا می‌رود، اگر پاسخ‌ها بین آزمودنی‌ها به اشتراک گذاشته شود، اعتبار آزمون را به خطر می‌اندازد (جورجیادو، تریانتافیلو و اکونومید[8] ۲۰۰۷؛ لی و داود[9] ۲۰۱۲). بنابراین، برای جلوگیری از این مشکل، گسترش و ایجاد یک خزانه سؤال بهینه، فرایندی طولانی و پرزحمت است. برای طراحان آزمون‌های CAT نامطلوب است که

---

1. High Stake
2. Graduate Record Examination
3. Graduate Management Admission Test
4. Twu
5. Davey & Nering
6. Barrada, Olea, Ponsoda & Abad
7. French & Thompson
8. Georgiadou, Triantafillou & Economide
9. Lee & Dodd

تنها از درصدی از سؤال‌های خزانه استفاده کنند، آنها تمایل دارند از تمام سؤال‌های خزانه به‌طور کامل استفاده کنند (ریولتا و پانسودا،[1] ۱۹۹۸؛ اُزترک و داگن،[2] ۲۰۱۵). بنابراین، برنامه‌ریزی دقیق و بهینه برای طراحی خزانه سؤال برای این نوع آزمون‌ها به امنیت آزمون نیز کمک می‌کند. در طراحی یک آزمون CAT مناسب، تمام مؤلفه‌های CAT بایکدیگر ارتباط تنگاتنگ دارند. زیرا مهم‌ترین مؤلفه CAT و همچنین، کنترل مواجهه سؤال، خزانه سؤال است. البته ویژگی‌های سؤال این خزانه نیز از اهمیت بسزایی برخوردار است و عاملی است که بر مواجهه سؤال تاثیر می‌گذارد (ریولتا و پانسودا، ۱۹۹۸)؛ زیرا از دیدگاه روان‌سنجی، سؤال‌ها بر اساس میزان آگاهی متناظر با پارامتر و سطح توانایی انتخاب می‌شوند که البته هدف آزمون نیز عامل تعیین کننده‌ای در میزان آگاهی است. در آزمون‌های نرم مرجع، خزانه سؤال باید شامل سؤال‌هایی در دامنه خیلی آسان تا خیلی دشوار باشد، و توزیع پارامتر b باید یکنواخت باشد. در صورتی‌که، در آزمون‌های ملاک–مرجع، که هدف‌شان تشخیص بین آزمودنی‌های بالا و پایین نقطه‌ی برش[3]است، اغلب سؤال‌ها در خزانه سؤال، باید در سطح دشواری لازم برای تعیین این هدف باشند تا بیشترین میزان آگاهی را فراهم کنند (بوید،[4] ۲۰۰۳). پس در ساخت این نوع آزمون‌ها به شیوه CAT باید به این امر مهم توجه لازم را مبذول داشت. البته هنگامی که مواجهه سؤال نیز در این نوع آزمون‌ها کنترل می‌شود، احتمال انتخاب یک سؤال را نسبت به زمانی که کنترل نمی‌شود، متفاوت می‌کند (هان،[5] ۲۰۱۲). یعنی در حالی که بر کنترل مواجهه سؤال تمرکز داریم نباید دقت اندازه‌گیری پارامتر توانایی را نادیده بگیریم. برای این منظور، باید شیوه‌ای برای کنترل مواجهه سؤال انتخاب شود که بین امنیت آزمون و دقت اندازه‌گیری پارامتر توانایی تعادل ایجاد کند (بوید، ۲۰۰۳؛ بوید، داود و فیتزپاتریک،[6] ۲۰۱۳). سه شیوه معروف در ادبیات پژوهشی مربوط به CAT برای کنترل مواجهه سؤال وجود دارد، روش سیمپسون–هتر[7](SH) (سیمپسون و هتر، ۱۹۸۵؛ هان، ۲۰۱۱؛ دیویس[1]و داود،

---

1. Revuelta & Ponsoda
2. Ozturk & Dogan
3. cut off point
4. Boyd
5. Han
6. Fitzpatrick
7. The Sympson-Hetter (SH) Method

۲۰۰۵). روش فید-اوی[۲](FAM) که توسط هان (۲۰۱۱) توسعه داده شد و روش رندوماسکو[۳](RA) این روش برای جلوگیری از بیش مواجهه شدن سؤالی که بیشینه آگاهی را در انتخاب سؤال ایجاد می‌کند، به‌وجود آمده است. این روش را کینگسبوری و زارا[۴](۱۹۸۹) پیشنهاد کرده‌اند. پژوهشی که در سال ۲۰۱۵ برای مقایسه روش‌های مناسب برای کنترل مواجهه در آزمون‌های مقیاس وسیع توسط اُزترک و داگن (۲۰۱۵) انجام گرفت، نشان داد زمانی که روش بیشینه آگاهی به‌عنوان روش انتخاب سؤال به‌کار می‌رود، روش کنترل مواجهه سیمپسون-هتر بالاترین راستی‌آزمایی[۵] برای برآورد توانایی را دارد، البته کمتر از روشی که کنترل مواجهه سؤال اعمال نمی‌شود؛ زیرا در حالتی‌که انتخاب سؤال بر اساس روش بیشینه آگاهی است، کنترل نکردن مواجهه سؤال بیشترین میزان دقت را ایجاد می‌کند ولی امنیت آزمون پایین می‌آید. همچنین، ریشه میانگین مجذور خطا (RMSE)[۶] در روش رندوماسکو (RA) در آزمون‌هایی با تعداد اندک سؤال، کمترین مقدار و میزان اریب[۷] نیز در این روش نسبت به سایر روش‌ها کمتر است.

لازم به ذکر است که از سنجش انطباقی کامپیوتری در مراکز و سازمان‌های سنجشی و آموزشی بسیاری از کشورهای پیشرفته استفاده می‌شود. به‌طوری که استفاده از نسخه‌های CAT این آزمون‌های سرنوشت‌ساز به دلیل دقت بالا و سهولت در اجرا، جایگزین آزمون‌های مداد-کاغذی شدند. برای مثال، نسخه CAT آزمون GRE و آزمون استعداد شغلی نیروهای مسلح[۸](ASVAB)، هم اکنون در دسترس بسیاری از کشورهای پیشرفته قرار دارد. مؤسسه ملّی اندازه‌گیری آموزشی (CITO) در هلند، چندین CAT تا به حال اجرا کرده است؛ مانند، MATCAT، (CITO، ۱۹۹۹)، TURCAT، (CITO، ۲۰۰۸)، DSLCAT، (CITO، ۲۰۰۲) و kindergartenCAT. MATCAT که برای تشخیص نقص‌های ریاضی در

---

1. Davi
2. The Fade-Away Method (FAM)
3. The Randomesque Method (RA)
4. Kingsbury & Zara
5. Fidelity
6. Root Mean Square Error (RMSE)
7. Bias
8. Armed Services Vocational Aptitude Battery

دانشجویان ایجاد شده است (ورشور و استریتمن[1] ۲۰۰۰). همچنین، TURCAT، برای سنجش مهارت زبان ترکی به‌عنوان زبان دوم است. DLSCAT نیز زبان هلندی را به عنوان زبان دوم می‌سنجد و kindergartenCAT شامل آزمون‌هایی برای اندازه‌گیری ترتیب، زبان، توانایی جهت‌یابی زمانی و مکانی کودکان است (ایگن[2] ۲۰۰۴). این CATها تقریباً همانند همه سیستم‌های CAT با کاربرد سؤال‌ها در خزانه سؤال سروکار دارد که به‌طور متنوعی در سرتاسر پیوستار توانایی توزیع می‌شوند (ولدکمپ، ورشور و ایگن[3] ۲۰۰۷). مزیت‌های سنجش انطباقی روزبه‌روز کاربرد آنها را در اغلب کشورهای پیشرفته بیشتر کرده امّا آنها را با دشواری‌ها و صرف مخارج بالایی نیز روبرو کرده است. در ایران، یکی از آزمون‌هایی که برای سنجش توانایی و مهارت زبان انگلیسی در سازمان سنجش آموزش کشور چندین سال است که برگزار می‌شود، آزمون تولیمو است. تولیمو یا TOLIMO، مخفف The Test of Language by the Iranian Measurement Organization است که در سازمان سنجش آموزش کشور طراحی و اجرا می‌شود و درواقع آزمونی استاندارد برای تعیین سطح دانش زبان انگلیسی است. تولیمو آزمون زبان انگلیسی پیشرفته برای دوره‌های فوق لیسانس و دکتری است. این آزمون شامل ۱۴۰ سؤال چهارگزینه‌ای، شامل؛ ساختار و نوشتار زبانی (۴۰ سؤال)، خواندن و درک مطلب (۵۰ سؤال)، درک مطلب شفاهی (۵۰ سؤال) و یک سؤال در بخش نوشتاری در حیطه‌های زبانی نوشتاری است (سایت سازمان سنجش آموزش کشور، ۱۳۹۸). این آزمون نیز همانند سایر آزمون‌ها (GRE، GMAT، ASVAB و...) یک آزمون سرنوشت‌ساز مهارت‌سنج یا ملاک‌مرجع است و قابلیت اجرا به‌صورت CAT را دارد، امّا تا به‌حال به شکل مداد-کاغذی اجرا می‌شده است. در شرایطی که این‌گونه آزمون‌ها به شیوه CAT طراحی می‌شوند، اغلب آزمون با سؤالی آغاز می‌شود که پارامتر b سؤال به صفر نزدیک می‌شود، روش بیشینه آگاهی[4] به‌عنوان روشی برای انتخاب سؤال استفاده می‌شود و همچنین، از روش برآورد بیشینه درست‌نمایی[5] برای

---

1. Verschoor & Straetmans
2. Eggen
3. Veldkamp, Vershoor & Eggen
4. maximum information
5. maximum likelihood estimation

برآورد پارامتر توانایی استفاده می‌شود. تعادل محتوایی در چنین آزمون‌های ملاک‌مرجعی بسیار مهم است؛ زیرا هدف اصلی آزمون سنجش همه واقعیت‌های مورد اندازه‌گیری است (اُزترک و داگن، ۲۰۱۵). در سال ۲۰۱۲، در یکی از مراکز معتبر آزمون‌سازی چین نیز مطالعاتی شبیه‌سازی شده برای ارزیابی یادگیری زبان چینی به‌عنوان یک زبان خارجی[1] (CFL) به‌شیوه CAT اجرا شد. این آزمون شامل دو قسمت بود، درک مطلب شنیداری و دیداری (خواندن). هر سؤال هنگامی‌که ارائه می‌شد زمان محدودی برای پاسخ داشت، به‌طوری‌که، هر وقت زمان هر پاسخ به پایان می‌رسید، سیستم به‌صورت خودکار به سؤال بعدی می‌رفت. زمان‌بندی هر سؤال این‌گونه تعبیه شده بود که برای هر سؤال فرصت دوبار خواندن وجود داشته باشد و بین هر بار خواندن ۵ ثانیه فاصله گذاشته می‌شد. سؤال‌ها بر اساس مدل سه‌پارامتری مدرج شده بودند (وانگ، کوا، چائو و تسای[2] ۲۰۱۵). در این پژوهش، هیچ نوع مطالعه دقیق برای ساخت یک برنامه CAT حرفه‌ای و بی‌نقص انجام نگرفته بود، در خصوص بررسی بهترین شیوه برآورد توانایی آزمودنی‌ها، قاعده شروع و اتمام، روش طراحی خزانه سؤال، نحوه کنترل تعادل محتوایی و اینکه آیا مواجهه سؤال کنترل شده است یا خیر، بررسی نشده بود. همچنین، مطالعه‌ای دیگر توسط کمیته ارزیابی دانشجویان پرستاری در آمریکا برای طراحی امتحانات کسب مجوز پرستاری[3] به‌شیوه CAT انجام گرفت (کمیته ملی برد تخصص پرستاری[4] ۲۰۱۱). پژوهش‌هایی که توسط مرکز آزمون‌سازی آنها انجام گرفت بیشترین تمرکز را بر الگوریتم‌های انتخاب سؤال، نحوه ساخت خزانه سؤال قرار داد، علی‌رغم اینکه، نتایج بهبودیافته‌ای نسبت به اجراهای مداد-کاغذی دریافت کردند، ولی برای شیوه کنترل مواجهه سؤال هیچ نوع برنامه‌ریزی نشد. پژوهشی دیگر که به منظور برآورد متوالی پارامتر توانایی در طول‌های متفاوت آزمون سنجش انطباقی کامپیوتری انجام گرفت، مورد مطالعه بر داده‌های ارزیابی ملی پیشرفت تحصیلی[5] (NAEP) مبتنی بود. در این پژوهش، مبنای مقایسه نتایج، یک مطالعه شبیه‌سازی شده بود، امّا به دلیل اینکه خزانه سؤال

---

1. Chinese as a foreign language (CFL)
2. Wang, Kuo, Chao & Tsai
3. nurse licensure and certification examinations
4. National Council of State Boards of Nursing
5. National Assessment of Educational Progress (NAEP)

شبیه‌سازی شده از قوانین یکسانی نسبت به بانک سؤال واقعی پیروی نمی‌کرد، نتایج قابل مقایسه‌ای به دست نیامد. زیرا در این پژوهش برای طراحی خزانه‌های سؤال از روش یا الگویی معین استفاده نشد (چانگ و ینگ[1] ۲۰۰٤). با توجه به بررسی پیشینه پژوهشی در خصوص طراحی CAT برای آزمون‌های خطیر، و همچنین بررسی‌های انجام گرفته توسط نویسنده این پژوهش تا سال ۲۰۱۹، پژوهشی پیدا نشد که هم‌زمان، تمام مؤلفه‌های CAT را درنظر گرفته باشد. البته لازم به ذکر است که اجرا یا گزارش کامل تمام مؤلفه‌های CAT در یک پژوهش کار پیچیده‌ای است، ولی در پژوهش حاضر تلاش شد به تمام مؤلفه‌ها حتی در حد کنترل واریانس مؤلفه‌ای که مورد مطالعه قرار نگرفته نیز توجه لازم صورت گیرد. طبق بررسی‌ها و همچنین، پژوهش‌های انجام گرفته توسط نویسنده مقاله فعلی در بین سال‌های ۱۳۹۵ تا ۱۳۹۷، الگوریتمی برای طراحی برنامه CAT برای آزمون‌های خطیر همچون تولیمو برای پژوهش فعلی نوشته شد که در آن به مؤلفه خزانه سؤال بهینه، نقطه شروع آزمون، شیوه انتخاب سؤال، شیوه برآورد توانایی، روش کنترل مواجهه سؤال و تعادل محتوایی توجه شد. قسمتی از نتایج این پژوهش در مرجع (مقدسین، ۱۳۹۵) گزارش شده است.[2] در پژوهش حاضر که مورد مطالعه آزمون تولیمو است، علاوه بر کاربرد الگوریتم ذکر شده، شیوه برآورد توانایی و قاعده توقف آزمون به‌عنوان دو مؤلفه مهم در کاربرد CAT وارد الگوریتم شدند و مورد مطالعه قرار گرفتند. همچنین، روش طراحی خزانه سؤال و کنترل مواجهه سؤال نیز مؤلفه‌هایی هستند که واریانس‌شان مورد مطالعه قرار گرفت. از آنجا که در آزمون تولیمو که یک آزمون ملاک‌مرجع و مهارت‌سنج است، تعادل محتوایی مؤلفه بسیار اهمیت دارد، در مطالعات شبیه‌سازی شده و عملیاتی CAT، این مؤلفه کنترل شد. در پایان به دلیل بالا بردن انگیزش آزمودنی‌ها و همچنین پیدا کردن نقطه شروع برآورد، سؤال آغازین برای همه آزمودنی‌ها یکسان و نزدیک به پارامتر b مساوی با صفر انتخاب شد، این ملاک در بسیاری از پژوهش‌ها مورد تأیید قرار گرفته است (ریکیسی، ۲۰۱۰؛ آزترک و داگن، ۲۰۱۵؛ کانجاناوسی، ۲۰۱۲).

---

[1]. Ying

[2]. برای کسب اطلاعات بیشتر در خصوص نحوه طراحی خزانه سؤال بهینه به آن مراجعه شود.

هدف اصلی از اجرای پژوهش حاضر، مطالعه برای طراحی و کاربرد سنجش انطباقی کامپیوتری برای آزمون تولیمو در سازمان سنجش است. هدف اول فرعی پژوهش حاضر، طراحی و اجرای یک آزمون CAT عملیاتی بر اساس نمونه‌ای از خزانه سؤال موجود آزمون تولیمو و اجرای آزمون CAT عملیاتی بر اساس برآورد پارامترهای توانایی آزمودنی‌های واقعی آزمون تولیمو است. هدف فرعی دوم این پژوهش، طراحی و اجرای آزمون CAT شبیه‌سازی شده با ویژگی‌های بهینه، به‌عنوان مبنایی برای مقایسه با CAT عملیاتی است، و هدف سوم فرعی این پژوهش، مقایسه برآورد توانایی و کارایی بین اجرای آزمون تولیمو به شیوه مداد-کاغذی با اجرای CAT عملیاتی و CAT شبیه‌سازی شده است.

روش پژوهش

از آنجایی‌که هدف از اجرای این پژوهش، طراحی و کاربرد سنجش انطباق کامپیوتری برای آزمون تولیمو بوده است؛ در این پژوهش، نخست ویژگی‌هایی یک آزمون CAT بهینه استخراج شد، سپس برای بررسی مزایا و کاستی‌های آن یک مطالعه شبیه‌سازی شده روی آزمون تولیمو انجام گرفت، به‌طوری که عملکرد CAT عملیاتی تولیمو مورد مقایسه با آن قرار گرفت. در پایان، برآورد توانایی آزمودنی‌ها در سه اجرای مداد-کاغذی، CAT عملیاتی و CAT شبیه‌سازی شده مورد مقایسه قرار گرفت. برنامه‌های شبیه‌سازی شده از طریق نسخه اصلی برنامه MATLAB (MathWorks, 2016)، به منظور شبیه‌سازی برنامه CAT طراحی شد. برنامه CAT عملیاتی نیز از طریق زبان PHP نوشته و از پایگاه داده MYSQL برای ذخیره‌سازی سؤال‌ها در خزانه استفاده شد. همچنین، برای ایجاد وزن‌های محتوایی به روش WDM از بسته نرم‌افزار "GAMS" (بروک، کندریچ و مروس[1]، ۱۹۸۸) استفاده شد. به منظور مطالعه آماره‌های سؤال‌ها و مدرج‌سازی پارامترهای آنها از نرم‌افزار SPSS-23 و BILOG-MG و برای بررسی تعداد عوامل زیربنایی آزمون از نرم‌افزار TESTFACT و NOHARM بهره گرفته شد. همچنین، برای بررسی استقلال موضعی سؤال‌های آزمون از نرم‌افزار R (پکیج، haven) استفاده شد.

---

[1]. Brooke, Kendrick & Meeraus

همچنین، مراحل زیر برای شبیه‌سازی CAT بهینه و عملیاتی انجام گرفت:

۱- مراحل شبیه‌سازی آزمون تولیمو به شیوه سنجش انطباقی کامپیوتری

گام اول، مدل‌یابی آزمون تولیمو به شیوه CAT: از آنجا که هدف از اجرای این پژوهش، طراحی و کاربرد آزمون تولیمو به شکل سنجش انطباقی در سازمان سنجش آموزش کشور بوده است، در این شبیه‌سازی تلاش شد تمام ویژگی‌های روان‌سنجی سؤال‌های واقعی آزمون تولیمو و همچنین، پارامتر توانایی آزمودنی‌های واقعی آزمون تولیمو که پس از مطالعات ویژگی‌های آن آزمون معلوم شد، به دقت وارد برنامه شبیه‌سازی شده و عملیاتی CAT شود. ۱- نقطه شروع آزمون: در هر سه بخش آزمون، سؤال مشترکی برای همه آزمودنی‌ها که درجه دشواری آن نزدیک به صفر بود (این مؤلفه ثابت نگه داشته شد) ارائه شد. زیرا بر اساس نظر (پارشال، اسپری، کالن و دیوی[1] ۲۰۰۲)، بهترین رویکرد برای انتخاب سؤال آغازین، سؤالی با دشواری متوسط است. به عبارتی اگر هیچ اطلاعی در مورد سطح توانایی آزمودنی نداشته باشیم، بهترین حدس ما این است که او همانند اکثریت آزمودنی‌های دیگر عمل می‌کند؛ ۲- در خصوص مؤلفه محتوای آزمون، آزمون‌ها بر اساس محتواهای از پیش تعیین شده توسط متخصصان موضوعی زبان انگلیسی (٤ متخصص) مشخص شد. همچنین لازم به ذکر است که، عامل تعادل محتوایی در تمام اجراها ثابت و کنترل شده، نگه داشته شد؛ به‌طوری که، خزانه سؤال بر اساس سه حوزه محتوایی (ساختار و نوشتار زبانی، خواندن و درک مطلب، درک مطلب شفاهی) تقسیم‌بندی شد. در این پژوهش برای وارد کردن تعادل محتوایی از روش برنامه‌نویسی اعداد صحیح (WDM) برای وزن دادن به محتواهای تعیین شده توسط طراحان استفاده شد؛ ۳ - همچنین، روش انتخاب سؤال‌ها روش بیشینه آگاهی (MI) به همراه جدول آگاهی در نظر گرفته شد؛ ٤- به منظور برآورد توانایی آزمودنی در طول اجرای آزمون، از روش پسین مورد انتظار (EAP)[2] و روش برآورد بیشینه درست‌نمایی استفاده شد. البته زمانی‌که از روش بیشینه درست‌نمایی برای برآورد پارامتر توانایی استفاده شد، پیش از اینکه آزمودنی‌ها در الگوی پاسخ خود حداقل دو پاسخ صحیح و غلط ایجاد کنند، از روش برآورد

---

[1]. Parshall, Spray, Kalohn & Davey

[2]. Expected a Posteriori Estimation (EAP)

مبتنی بر بیزین اوون[۱](۱۹۷۵) برای میانگین پسین[۲]استفاده شد، ولی پس از ایجاد حداقل دو پاسخ صحیح و غلط در الگوی پاسخ، از روش بیشینه درست‌نمایی بهره گرفته شد؛ ۵- طول آزمون‌ها یا قاعده توقف آزمون به دو شیوه طول ثابت آزمون[۳](در این روش تعداد از پیش تعیین شده سؤال برای همه آزمودنی‌ها اجرا شد. در اجراهای شبیه‌سازی شده نیز همانند آزمون تولیموی مداد کاغذی ۱٤۰ سؤال اجرا شد) (چویی، گریدی و داد[٤] ۲۰۱۰) و خطای استاندارد ثابت[٥](SE) (بر اساس این روش آزمون زمانی متوقف شد که خطای استاندارد برآورد توانایی کمتر از مقدار خطای استاندارد از پیش تعیین‌شده باشد) (داد، کوک و دی آیالا[٦] ۱۹۹۳؛ بوید، داد و چویی[۷] ۲۰۱۰) در نظر گرفته شده است. در این مطالعه، سه مقدار خطای استاندارد در نظر گرفته شد: SE کمتر از ۰/۳۸۵ (مشابه ضریب اعتبار[۸]۰/۸۵)، SE کمتر از ۰/۳۱۵ (مشابه ضریب اعتبار ۰/۹۰) و SE کمتر از ۰/۲۲ (مشابه ضریب اعتبار ۰/۹۵) (بابکوک و وایس[۹]؛ ۲۰۰۹)؛ ٦- به منظور کنترل مواجهه بیش از حد سؤال، یک بار شبیه‌سازی «بدون کنترل مواجهه» و یک بار با کنترل مواجهه سؤال به روش سیمپسون-هتر با «نرخ مواجهه هدف»[۱۰]برابر با $\frac{1}{3}$ انجام گرفت.

گام دوم، ایجاد جامعه آزمون‌دهندگان (جامعه و نمونه مطالعه)

جامعه: دو عامل مهمی که در هر دو CAT شبیه‌سازی شده و عملیاتی در نظرگرفته شد، عبارت از جمعیت آزمودنی‌های هدف و پهنای «bin» بودند. در شبیه‌سازی CAT لازم است که در مورد جمعیت آزمودنی‌های هدف اصلی که آزمون برای آنها در آینده ساخته خواهد شد، اطلاعاتی دقیق وجود داشته باشد (ریکیسی، ۲۰۱۰). بنابراین، در این پژوهش نیز از توزیع عملکرد جمعیت مشاهده شده

---

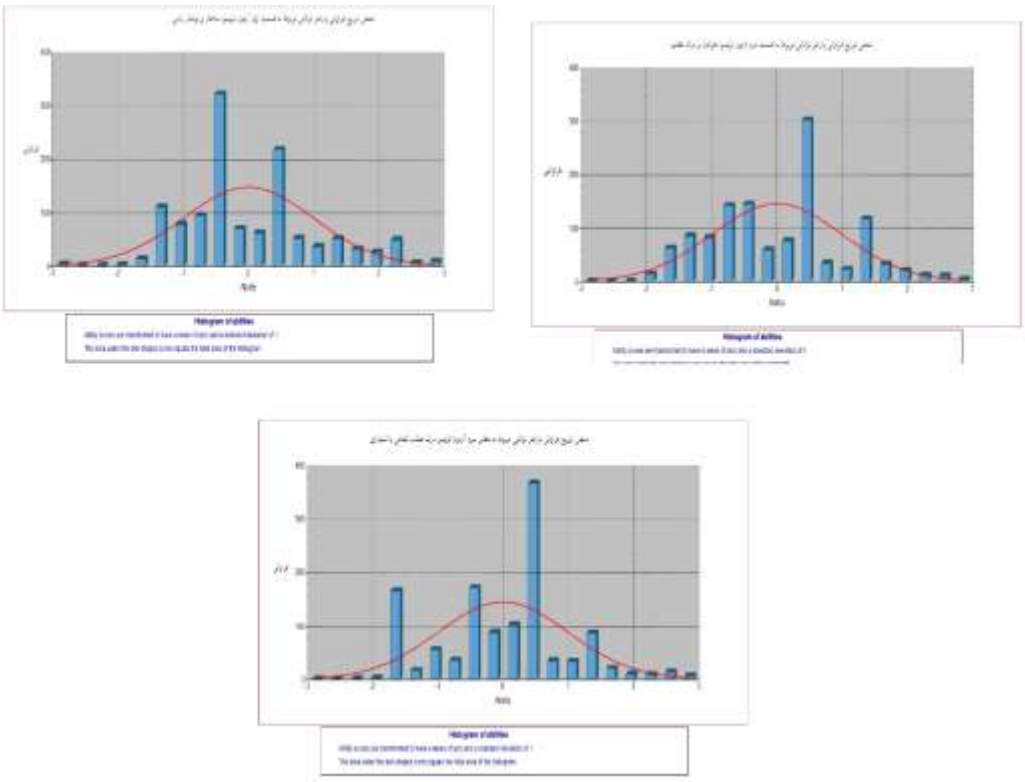
[1]. Owen

[۲]. توزیع پیشین توانایی با میانگین صفر و انحراف استاندارد یک در نظر گرفته شد.

[3]. fixed length of test
[4]. Choi, Grady & Dodd
[5]. fixed standard error
[6]. Dodd, Koch & De Ayala
[7]. Boyd, Dodd & Choi
[8]. reliability
[9]. Babcock & Weiss
[10]. target exposure rate

آزمودنی‌های تولیمو مداد-کاغذی به‌عنوان توزیع هدف برای CAT شبیه‌سازی شده و اجرای CAT عملیاتی استفاده شد. در مورد پهناهای «bin»، پهنای ۰/۲ در اجرای CAT عملیاتی و شبیه‌سازی شده وارد شد. پس از بررسی توزیع جمعیت آزمودنی‌هایی که در آزمون تولیموی واقعی سازمان سنجش آموزش کشور شرکت کردند، نوع توزیع توانایی آنها با اجراهای مداد کاغذی ۱۰ دوره آزمون تعیین شد (دوره ۱۱٤ تا ۱۲۳، به‌صورت نمونه‌گیری هدفمند انتخاب شدند، و توزیع توانایی همه ۱۰ دوره در برنامه CAT عملیاتی وارد شد. به‌دلیل محدود بودن حجم این مقاله، تنها نمایش تصویری برآورد توانایی دوره ۱۲۳ به‌عنوان نماینده سایر دوره‌ها انتخاب شد، زیرا بیشترین تعداد آزمودنی نسبت به سایر دوره‌ها، در آن شرکت داشتند و همچنین، میانگین و انحراف معیار توزیع پارامتر توانایی این دوره تقریباً مشابه سایر ۹ دوره دیگر بود. به‌همین‌دلیل، تنها توزیع توانایی این مرحله نمایش داده‌شد). توزیع توانایی واقعی آزمودنی‌ها (توزیع پیشین تتا) هم در برنامه CAT عملیاتی و هم برنامه CAT شبیه‌سازی برای قابلیت بیشتر برای مقایسه‌پذیری وارد شد. همان‌طور که نمودار (۱) نشان می‌دهد، توزیع توانایی در بخش اول آزمون که مربوط به ساختار و نوشتار زبانی است، در پایین توزیع تراکم بیشتری دارد، ولی در قسمت دوم و سوم آزمون، یعنی به ترتیب، خواندن و درک مطلب شفاهی، در بالای توزیع تراکم بیشتری به چشم می‌خورد.

نمونه: دو توزیع حجم نمونه در این پژوهش به کار رفت: به عبارت دیگر، با یک نمونه CAT عملیاتی اجرا شد و همچنین، خزانه‌های سؤال بهینه نیز برای CAT شبیه‌سازی شده طراحی شدند، و با نمونه دیگری عملکرد کلی CAT شبیه‌سازی‌شده و عملیاتی ارزیابی شدند. الف) ۱۲۰۰۰ θ از توزیع جامعه هدف نمونه‌گیری شدند و به‌عنوان توانایی واقعی آزمودنی‌ها وارد اجرای CAT عملیاتی و CAT شبیه‌سازی‌شده شدند. ب) ٦٥ مقدار ثابت θ در دامنه ٤- تا ٤ با فاصله ۰/۱۲۵ انتخاب شدند (یعنی، $(-4, -3.875, \ldots, 3.875, 4)$) و با درنظر گرفتن ۵۰۰ آزمودنی در هر سطح θ، ۳۲۵۰۰ آزمودنی به صورت مفروض انتخاب شدند؛ زیرا محاسبه ملاک‌های مشروط تتا به تعداد آزمودنی در هر سطح تتا وابسته بود، به‌همین دلیل، در شبیه‌سازی یک توزیع یکنواخت تعبیه شد.



نمودار (۱) توزیع برآورد توانایی آزمودنی‌ها شرکت‌کننده در آزمون تولیمو دوره ۱۲۳

گام سوم، ایجاد پارامترهای سؤال: با توجه به ملاک‌های شروع آزمون، نوع برآورد توانایی آزمودنی‌ها، نقطه توقف آزمون و شیوه نمره‌دهی به پاسخ‌ها، سؤال‌ها و پارامترهای آنها در برنامه شبیه‌سازی شده تولید شدند. البته مدل IRT که فرض شد سؤال‌ها بر اساس آن مدرج شدند، مدل دو و سه‌پارامتری لوجستیک بود. پارامترهای a، b و c از طریق سه روش طراحی خزانه سؤال (یعنی، R، MRP و MTI) تولید شدند. این سه روش براساس ایده‌های گو (۲۰۰۷)، مک‌برید و وایس (۱۹۷۶)، برای ایجاد خصوصیات بهینه سؤال در مدل سه پارامتری به‌وجود آمد. در این پژوهش با اندکی تغییر در روش‌های پیشین ایجاد سؤال، از آنها استفاده شد. روش اول، روش تصادفی (R)؛ روش دوم، روش آمیخته تصادفی و پیش‌بینی (MRP) و روش سوم، روش کمینه آگاهی تست (MTI) نامیده می‌شود.

گام چهارم، ایجاد داده‌های پاسخ: در برنامه شبیه‌سازی شده CAT، پاسخ‌های آزمودنی‌ها به دنبال هر سؤالی که بر طبق مدل مفروض و شیوه برآورد پارامتر توانایی ایجاد شد، تولید شد. از آنجاکه، روش برآورد توانایی آزمودنی‌ها در CAT شبیه‌سازی شده و عملیاتی تولیمو، روش برآورد بیشینه درست‌نمایی (ML) و پسین مورد انتظار بود، به دلیل آنکه در روش (ML)، تا زمانی که آزمودنی پاسخ درست یا غلط به سؤال‌ها نداده است (یعنی، نمره ۱ یا ۰ دردسترس نباشد)، برآورد بیشینه درست‌نمایی تعیین نمی‌شود، بنابراین تا زمانی‌که هیچ پاسخ درست یا غلطی در الگوی پاسخ آزمودنی‌ها دیده نشده بود، به شیوه زیر عمل شد:

چون (Θ) واقعی اولیه آزمودنی در شبیه‌سازی معلوم بود، (P_{ij}) بعد از هر سؤالی که برای آزمودنی اجرا شد، قابل محاسبه بود. بنابراین، چون در ابتدا در الگوی پاسخ، نمره ۰ یا ۱ موجود نبود، به این نحو عمل شد، ابتدا عدد تصادفی (m_{ij}) از یک توزیع یکنواخت ($U(0,1)$) استخراج شد و با مقدار (P_{ij}) به دست آمده مقایسه شد. اگر (m_{ij}) برابر با یا کمتر از (P_{ij}) بود، بنابراین، پاسخ برابر با عدد ۱ می‌شد، در غیر این صورت، پاسخ برابر با صفر در نظر گرفته می‌شد (ریکیسی، ۲۰۰۳؛ ۲۰۰۷؛ ۲۰۰۹؛ ۲۰۱۰) مقدار (P_{ij}) بر اساس معادله (۱) به دست می‌آید:

$$P_i(u_{ij}=1|\theta_j,b_i)=\frac{1}{1+\exp[-1.7(\theta_j-b_i)]} \quad (۱)$$

$$m_{ij} \square U(0,1)$$

اگر

$$m_{ij} \le P_{ij} \Rightarrow 1$$
$$m_{ij} > P_{ij} \Rightarrow 0$$

پس از پاسخ اول که بر طبق ساختار بالا، آزمودنی نمره ۰ یا ۱ گرفت، چون هنوز پاسخ درست یا غلط دیگری در الگوی پاسخ دیده نمی‌شد، برآورد بیشینه درست‌نمایی نامتناهی بود و تعیین نمی‌شد. بنابراین در این روش اگر پاسخ اول آزمودنی درست بود به برآورد جدید ($\hat{\theta}$) ۰/۷۰ اضافه می‌شد و اگر غلط بود، ۰/۷۰ از برآورد جدید ($\hat{\theta}$) کم می‌شد. بر این اساس، در سؤال بعدی پارامتر ($b=\hat{\theta}$) بود، تا میزان آگاهی سؤال را بیشینه کند، پارامتر b سؤال بعدی بر اساس پاسخ سؤال قبل باید برابر با دو مقدار $b=0.70$ یا $b=-0.70$ فرض می‌شد. بنابراین، بعد از زمانی‌که در الگوی پاسخ، یک پاسخ درست یا غلط دیده شد از برآورد بیشینه درست‌نمایی برای برآورد توانایی استفاده شد.

گام پنجم، طراحی خزانه سؤال بهینه: در این پژوهش به‌منظور طراحی خزانه سؤال بهینه برای CAT شبیه‌سازی‌شده از روش شبیه‌سازی مونت‌کارلو ریکیسی (۲۰۰۳) استفاده شد. همچنین، به منظور تعیین مجموعه‌ای از ویژگی‌های محتوایی آزمون، از روش برنامه‌نویسی اعداد صحیح یا برنامه‌نویسی خطی[1](WDM) استفاده گردید. در مجموع به منظور شبیه‌سازی خزانه سؤال برای برنامه CAT تولیمو، دو مرحله مهم انجام گرفته است: در مرحله اول با به کارگیری روش مونت‌کارلو اکتشافی ریکیسی (۲۰۰۳)، مشخصات خزانه‌های سؤال ایدئال شامل اندازه بهینه و پارامترهای آماری مربوط به خزانه سؤال بهینه تعیین شد. در مرحله دوم بر اساس روش

[1]. linear integer programming

برنامه‌نویسی اعداد صحیح (WDM) (استوکینگ و سوانسون، ۱۹۹۳)، صفات محتوایی آزمون، مشخص و در شبیه‌سازی وارد شد و بدین ترتیب مدل طرح بهینه خزانه سؤال طراحی شد. بدین صورت که، ۱۲۰۰۰ آزمودنی به‌طور تصادفی از جامعه هدف توانایی برای شبیه‌سازی انتخاب شدند و بر این اساس، ویژگی‌های خزانه سؤال بهینه برای آزمون تولیمو استخراج شد. پژوهش‌های متعدد نشان داده است که این روش به‌خوبی عمل می‌کند (ریکیسی و هی، ۲۰۰٤؛ ریکیسی و هی، ۲۰۰۵؛ گو، ۲۰۰۷). این ایده اصلی پشت این روش است تا از «bin»هایی استفاده شوند که پهنای معینی روی مقیاس پارامتر «b» ایجاد می‌کند. به‌طوری که، مجموع سؤال‌ها در هر b-bin، محاسبه و از مکانیسم تئوری «اجتماع»[1] برای تعیین تعداد کلی سؤال‌ها استفاده می‌شود. برای افزایش دقت، در این پژوهش پهنای b-bin برابر با ۰/۲ انتخاب شد. روش کار در این پژوهش به این صورت بود که ابتدا، یک خزانه سؤال به قسمت‌های کوچک‌تری تقسیم شد و این تقسیم‌بندی‌ها بر اساس صفات غیر آماری همچون سطوح و قیود محتوایی مبتنی بودند. سپس شبیه‌سازی با یک آزمودنی شروع شد که به‌طور تصادفی از جامعه هدف، انتخاب و CAT روی او اجرا شد. هر سؤالی که اجرا می‌شد، انتخاب آن به‌صورت بهینه بود، به‌طوری که همه ویژگی‌های آماری و غیر آماری یک سؤال بهینه را داشته باشد. سؤال‌هایی که اجرا می‌شدند، درون «bin»هایی مرتب و منظم می‌شدند و تعداد آنها محاسبه می‌شد. در مرحله بعد، روش مشابهی برای آزمودنی‌های دیگر اجرا شد. از آنجا که سؤال‌هایی که برای یک نفر انتخاب می‌شد، می‌توانست برای اشخاص دیگری نیز انتخاب شود، پس خزانه سؤال بهینه، اجتماعی از مجموعه سؤال‌هایی بود که برای هر کدام از افراد انتخاب می‌شد. با استفاده از تعداد زیادی از آزمودنی‌های جامعه هدف، این انتظار وجود داشت که با افزایش تعداد آزمودنی‌ها، تعداد سؤال‌هایی که باید به خزانه اضافه شود، کمتر شود. در پایان، اندازه خزانه سؤال با مهیا کردن تمام ملزومات برای همه آزمودنی‌ها، به سطح مجانب[2] رسید. همچنین، در این پژوهش روش اکتشافی ریکیسی با استفاده از مدل‌های R، MRP و MTI به مدل‌های دو و سه‌پارامتری نیز تعمیم داده شد (ریکیسی، ۲۰۱۰؛ مقدسین و همکاران، ۱۳۹٤). در مدل‌های دو و سه‌پارامتری نیز از

---

1. union
2. asymptote

ایده «bin»ها استفاده شد، با این تفاوت که پهنای معینی روی مقیاس پارامتر «b» و پهنای دیگری روی پارامتر «a» ایجاد شد. پهنای پارامتر a بر اساس تغییرات میزان آگاهی سؤال‌ها نسبت به تابع درجه دوم پارامتر a مشخص شد. سپس، مجموع سؤال‌ها در هر ab-bin محاسبه و از مکانیسم تئوری «اجتماع» برای تعیین تعداد کلی سؤال‌ها استفاده شد. پارامترهای بهینه سؤال‌ها نیز بر این اساس مشخص شد و در پایان در یک مخزن نگه‌داری شدند. به منظور کنترل مواجهه بیش از حد سؤال‌ها، دو خزانه سؤال شبیه‌سازی شدند، در یکی از شبیه‌سازی کنترل مواجهه اعمال نشد و در شبیه‌سازی دیگر کنترل مواجهه اعمال شد.

گام ششم، «اصلاح پس از شبیه‌سازی»[1]: تعداد ۲۰ تکرار برای هر ترکیبی از روش‌ها و متغیرهای کنترل انجام شد، تا جایی که برآورد نسبتاً ثابتی از خزانه سؤال بهینه به دست آید. پیش از اصلاح پس از شبیه‌سازی، الگوها و تعداد مواجهه سؤال در ۲۰ تکرار میانگین‌گیری شد.

بنابراین، در پژوهش حاضر چهار عامل در CAT شبیه‌سازی شده دستکاری شد، ۱- روش برآورد پارامتر توانایی آزمودنی‌ها (دو روش)؛ ۲- ملاک توقف آزمون (چهار روش)؛ ۳- روش ایجاد سؤال بهینه (سه روش R، MRP، MTI)؛ ٤- عامل کنترل مواجهه سؤال سیمپسون-هتر (دو روش، عدم کنترل مواجهه سؤال، کنترل مواجهه با روش سیمپسون-هتر). اگر تمام عوامل با یکدیگر تقاطع پیدا کنند، (۲*٤*۳*۲ روش) ایجاد می‌شود، یعنی ٤۸ شیوه مطالعه CAT شبیه‌سازی شده. از آنجایی که در این مقاله، امکان بررسی و گزارش تمام حالات ممکن وجود ندارد، به جای روش متقاطع، از روش آشیانه‌ای استفاده می‌شود، یعنی ابتدا، بهترین روش برآورد و خاتمه سؤال (تقاطع این دو مؤلفه با یکدیگر) از لحاظ دقت اندازه‌گیری مشخص می‌شود، سپس، تنها دو روش ایجاد سؤال بهینه و کنترل مواجهه با توجه به روش برآورد پارامتر و ملاک توقف بهینه، مورد مطالعه قرار می‌گیرد.

۲- روش‌های ارزیابی دقت برآورد پارامتر توانایی: بدین منظور از شاخص‌های زیر استفاده شد. همه مقادیر زیر به‌طور جداگانه برای ۲۵ تکرار محاسبه شد (برای

---

[1]. post-simulation

کاهش واریانس نمونه‌گیری، به پیشنهاد هارول و همکاران، ۱۹۹۹)، سپس میانگین ۲۵ تکرار در بخش نتایج گزارش شد:

۱. ضریب راستی‌آزمایی[1] که با محاسبه همبستگی گشتاوری پیرسون بین برآوردهای پارامتر توانایی در سنجش انطباقی و اجرای مداد-کاغذی محاسبه می‌شود (یوری[2] ۱۹۷۰؛ وال و ویس[3] ۱۹۷۵، به نقل از لیانگ، چانگ و هاو[4] ۲۰۰۲). با بالا رفتن این ضریب، میزان اعتماد یا پایایی نیز بالاتر می‌رود.

۲. اریب[5] که با محاسبه میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی در همه آزمودنی‌ها محاسبه می‌شود.

۳. ریشه میانگین مجذور خطا (RMSE)، که با محاسبه ریشه دوم مجذور میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی در همه آزمودنی‌ها محاسبه می‌شود.

٤. متوسط قدر مطلق تفاوت[6](AAD)، که با محاسبه قدرمطلق میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی در همه آزمودنی‌ها محاسبه می‌شود (بوید و همکاران[7] ۲۰۱۳؛ داویس[8] ۲۰۰۲).

۳- روش‌های ارزیابی عملکرد CAT شبیه‌سازی شده: در این پژوهش عملکرد CAT شبیه‌سازی شده بر اساس مجموعه‌ای از ملاک‌های تجربی، ارزیابی و با آزمون CAT عملیاتی مقایسه شد (چانگ و ییِنگ[9] ۱۹۹۹؛ ریکیسی و هی، ۲۰۰۵):

۱. آگاهی شرطی آزمون:[10]آگاهی آزمون در هر یک از سطوح توانایی برابر است با مجموع کل آگاهی هر یک از سؤال‌های آزمون در آن سطح توانایی. آگاهی آزمون، به‌عنوان شاخص کارایی[11]آزمون در نقاط مختلف توانایی در نظر گرفته

---

1. fidelity coefficient
2. Urry
3. Vale &Weiss
4. Leung, Chang & Hau
5. bias
6. average absolute difference
7. Boyd et al
8. Davis
9. Chang&Ying
10. Conditional test information
11. Efficacy index

می‌شود. هرچه میزان آگاهی آزمون در یک سطح توانایی بیشتر باشد، کارایی آزمون در آن سطح نسبت به سایر سطوح توانایی نیز بیشتر است:

$$I(\theta_j) = \sum_{i=1}^{I} a_{ij}^2 \frac{(P_{ij} - c_i)^2 . q_{ij}}{(1 - c_i)^2 . p_{ij}} \qquad (۲)$$

۲. خطای استاندارد شرطی اندازه‌گیری[1](CSEM): این شاخص میزان خطای اندازه‌گیری برآورد توانایی را در هر یک از سطوح توانایی واقعی (θ_j) محاسبه می‌کند:

$$SEM(\theta_j) = \sqrt{\frac{1}{N_i} \sum_{i=1}^{N_i} (\hat{\theta}_{ij} - \bar{\theta}_{ij})^2} \qquad (۳)$$

اگر θ_j، توانایی j ام (j=۱، ۲، ... ،٦٥) در پیوستار ٤- تا ٤+ (یعنی؛ ٤+، ۳/۸۷٥+، ... ۳/۸۷٥-، ٤-=θ_j) را نشان دهد، i هر یک از آزمودنی‌ها در θ_j و ۵۰۰=N_i، تعداد کل تکرارهای CAT اجرا شده در θ_j است. $\hat{\theta}_{ij}$ ($\hat{\theta}_{ij}$=۱، ۲،...،۵۰۰) برآورد θ_{ij} و $\bar{\theta}_{ij} = \frac{1}{N_i}\sum_{i=1}^{500} \hat{\theta}_{ij}$، میانگین ۵۰۰ برآورد θ_{ij} ($\hat{\theta}_{ij}$) در θ_j است.

۳. اریب و میانگین مجذور خطا[2](MSE):

$$Bias = \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i) \qquad (٤)$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2 \qquad (٥)$$

٤. اریب شرطی[3]و میانگین مجذور خطای شرطی[4](CMSE):

---

[1]. Conditional standard error of measurement
[2]. Bias and mean square error
[3]. Conditional Bias
[4]. Conditional mean square error

$$Conditional\ Bias(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j) \quad (٦)$$

$$CMSE(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j)^2 \quad (٧)$$

۵. کجی توزیع نرخ مواجهه سؤال[1]: آماره کای دو که توسط چانگ و ییِنگ (۱۹۹۹) ارائه شده، برای اندازه‌گیری میزان کجی توزیع مواجهه سؤال به کار رفته و برابر است با:

$$X^2 = \sum_{i=1}^{n}\frac{(r_i - \frac{L}{n})^2}{\frac{L}{n}} \quad (٨)$$

٦. درصد سؤال‌های بیش مواجهه شده[2]: نرخ مواجهه یک سؤال را می‌توان به عنوان نسب تعداد دفعات اجراهای سؤال به تعداد کل آزمودنی‌ها در نظر گرفت. (هَو و چانگ[3]، ۲۰۰۱).

۷. درصد سؤال‌های کم‌مواجهه شده[4]: نرخ کم‌مواجهه شدن یک سؤال بدین معناست که یک سؤال به‌ندرت در برنامه CAT مورد استفاده قرار گیرد. خزانه سؤالی که سؤال‌های بسیار زیادی با نرخ مواجهه خیلی پایینی دارد، دارای فایده کمی است. (هَو و چانگ، ۲۰۰۱، ریکیسی، ۲۰۰۹).

۸. نرخ همپوشی آزمون[5]: نرخ همپوشی آزمون، عبارت است از تعداد مورد انتظار سؤال‌های مشترکی که به دو آزمودنی که به طور تصادفی نمونه‌گیری شدند، ارائه می‌شود، تقسیم بر طول مورد انتظار آزمون[6]: به‌طور ایدئال، تعداد سؤال‌های مشترک

---

1. Skewness of item exposure rate distribution
2. Percentage of overexposed items
3. Hau & Chang
4. Percentage of underexposed items
5. Test overlap rate
6. Expected test length

بین دو آزمودنی که به‌طور تصادفی نمونه‌گیری شدند، باید حداقل باشد (چانگ و ییننگ، ۱۹۹۹؛ چن، آنکنمان، اسپری[1] ۱۹۹۹):

$$\bar{T} = \frac{\sum_{i=1}^{n} \binom{m_i}{2}}{L\binom{N}{2}} = \frac{\sum_{i=1}^{n} m_i(m_i - 1)}{LN(N-1)} \qquad (۹)$$

٤- طراحی برنامه CAT عملیاتی برای آزمون تولیمو

به‌منظور اینکه، ویژگی‌های بهینه طراحی شده برای برنامه CAT آزمون تولیمو از لحاظ کاربردی مفید و واقع‌بینانه باشد، اطلاعات پیشین آزمون تولیموی واقعی - علاوه بر طراحی مطالعه شبیه‌سازی‌شده - در مطالعه CAT عملیاتی نیز وارد تجزیه‌وتحلیل شدند. اطلاعات پیشین شامل، توزیع‌های پارامترهای سؤال‌های واقعی آزمون تولیمو، رابطه بین پارامترهای سؤال، اعتبار تست، و برآوردهای توانایی آزمودنی‌ها است. بر این اساس، در این پژوهش، برای طراحی CAT آزمون تولیمو، ابتدا سؤال‌های موجود مربوط به خزانه سؤال واقعی آزمون تولیمو مورد تحلیل روان‌سنجی با مدل IRT مناسب قرار گرفتند و اطلاعات وارد خزانه سؤال عملیاتی شدند. همچنین، لازم به ذکر است که به دلیل حجم زیاد آزمون‌های اجرا شده که تاکنون ۱۳۵ دوره از آزمون تولیمو به صورت مداد کاغذی تا بهمن ماه ۱۳۹۷ اجرا شده است، از ۱۳۵ دوره، ۱۰ دوره از جدیدترین دوره‌های اجرایی آزمون به صورت هدفمند انتخاب شد و در این پژوهش سؤال‌های آنها بر اساس مدل دو و سه‌پارامتری لوجستیک مورد تجزیه و تحلیل قرار گرفتند. تعداد سؤال‌های موجود، ۱٤۰۰ سؤال بود (۱۰ دوره آزمون تولیمو، در بخش ساختار و نوشتار زبانی (٤۰ سؤال)، در بخش خواندن و درک مطلب (۵۰ سؤال) و در بخش درک مطلب شفاهی و شنیداری (۵۰ سؤال)). لازم به ذکر است، سؤال‌های مربوط به بخش خواندن و درک مطلب به این صورت بودند که تعدادی سؤال (برای مثال، ۸ تا ۱۲ سؤال) مربوط به یک متن می‌شدند. همچنین، در هر دوره از آزمون اغلب چهار یا پنج متن (Reading) به همراه مجموعه‌ای از سؤال‌ها که همه به یک متن مربوط بودند، ارائه می‌شد؛ بنابراین،

[1]. Chen & Ankenmann & Spray

ابتدا باید در مورد استقلال موضعی[1](LII) داده‌های مربوط به سؤال‌های این بخش از آزمون تولیمو اطمینان حاصل می‌شد. به‌همین دلیل در این پژوهش از روش مقادیر بحرانی آماره Q_3 ین[2]استفاده شد. ین (۱۹۸٤) آماره Q_3 را برای تعیین وابستگی موضعی برای مدل دو و سه‌پارامتری لوجستیک ارائه داد. این آماره به باقی‌مانده سؤال‌ها (تفاوت بین نمره واقعی و نمره برآورده شده برای سؤال برحسب پارامتر توانایی) وابسته است:

$$d_i = X_i - E(X_i \mid \hat{\theta}) \qquad (۱۰)$$

همبستگی پیرسون بین همه جفت باقی‌مانده‌های سؤال‌ها (برای همه آزمودنی‌ها) با معادله ۱۱ محاسبه می‌شود:

$$Q_{3,ij} = r_{d_i d_j} \qquad (۱۱)$$

مقدار d_i و d_j باقی‌مانده سؤال i و j به‌ترتیب هستند. چن و تیسن[3](۱۹۹۷) مقدار بحرانی آماره Q_3 را برای بررسی وابستگی موضعی برابر با قدرمطلق مقدار ۰/۲ اعلام کرده‌اند. امّا، درمطالعه‌ای که به شیوه شبیه‌سازی شده بوت‌استرپ[4]انجام گرفت، علاوه بر مقدار ثابت بالا، مقادیر دقیق این آماره را به تعداد سؤال‌ها، حجم آزمودنی‌ها و تعداد طبقات پاسخ نیز مرتبط می‌دانند. همچنین، در این مطالعه مقدار استقلال موضعی را به متوسط همبستگی مشاهده شده بین باقی‌مانده‌ها مرتبط می‌دانند نه فقط به توزیع یکنواخت و ثابت یک مقدار باقی‌مانده (کریستنسن، مکرانکی و هورتن[5]، ۲۰۱۷). ماریس[6](۲۰۱۳) نیز بیان کرد، زمانی که سؤال‌های یک آزمون کمتر از ۲۰ سؤال باشد، یک مقدار قدر مطلق ثابت مناسب نیست ولی زمانی که بیشتر از ۲۰ باشد (در مطالعه حاضر ۵۰ سؤال) این مقدار دقیق است. او نیز پیشنهاد کرده است در آزمون‌هایی با ۲۰ سؤال یا کمتر برای بررسی مقادیر همبستگی بین باقی‌مانده‌ها باید به

---

1. local item independence
2. Yen
3. Chen & Thissen
4. bootstrapping
5. Christensen, Makransky & Horton
6. Marais

متوسط این توزیع توجه کرد. بنابراین، بر اساس فرض صفر (عدم وابستگی موضعی)، متوسط همبستگی باقی‌مانده‌ها منفی است و به‌صورت ایدئال همبستگی مشاهده شده بین باقی‌مانده‌ها در مجموعه داده‌ها باید بر اساس متوسط همبستگی محاسبه شود؛ که مقدار متوسط همبستگی‌های مشاهده شده بر اساس فرمول زیر محاسبه می‌شود:

$$\bar{Q}_3 = \binom{1}{2}^{-1} \sum_{i>j} Q_{3,ij} \qquad (۱۲)$$

مقادیر Q_3 و مقدار متوسط آن در پژوهش حاضر برای هر ۱۰ دوره به صورت مجزا با استفاده از بسته نرم افزار "haven" در برنامه R محاسبه شد. مقدار میانه Q_3 از ۰/۰۸۱٤- تا ۰/۰۰۲۳- متغیر بود و میانگین Q_3 از ۰/۰۰۹۲۵- تا ۰/۰۰۵٤۸- متغیر بود. از آنجا که در این مقاله، فقط نتایج مربوط به تحلیل سؤال‌های دوره ۱۲۳ به‌عنوان نماینده سایر دوره‌ها ارائه شده است و به دلیل محدودیت جا امکان ارائه نتایج مربوط به همه دوره‌ها وجود نداشت، بنابراین، فقط جدول و نمودار مربوط به بررسی استقلال موضعی دوره ۱۲۳ در بخش نتایج ارائه خواهد شد. میانه Q_3 در این دوره برابر با ۰/۰۱۹۰- و میانگین آن برابر با ۰/۰۰۸۷۵- به‌دست آمد.

٤-۱- نحوه نمونه‌گیری از دوره‌های آزمون تولیمو مداد-کاغذی و برآورد توانایی شرکت‌کنندگان آزمون تولیمو

از آنجایی‌که به‌طور رایج در سازمان سنجش آموزش کشور در طول یک سال اغلب ۱۰ دوره آزمون تولیمو برگزار می‌شود، نتایج تحلیل سؤال‌های همه ۱۰ دوره محاسبه و در MYSQL خزانه سؤال عملیاتی وارد شد. به دلیل آنکه هدف از اجرای این پژوهش، طراحی و اجرای آزمون تولیمو به شکل سنجش انطباقی کامپیوتری است، مطالعه شبیه‌سازی برای CAT تولیمو، نیز برای یک دوره یک ساله (یعنی، ۱۰ دوره اجرای آزمون تولیمو) انجام گرفته است. در جدول (۱) اطلاعات مربوط به تعداد شرکت‌کنندگان این ۱۰ دوره، ارائه شده است و نشان می‌دهد که ۱۱۹۰۳ نفر در این ۱۰ دوره شرکت کرده‌اند. ما اطلاعی از تکراری بودن آزمودنی‌ها نداشتیم، بنابراین فرض را بر استقلال آزمودنی‌ها در ۱۰ دوره قرار دادیم (در صورتی‌که این امکان وجود دارد که آزمودنی‌ها در چند دوره برگزار شده در یک سال شرکت کنند). در مجموع برای مشابهت اجرای شبیه‌سازی شده CAT با اجراهای واقعی مداد-کاغذی،

در طراحی CAT شبیه‌سازی‌شده فرض را بر این قرار دادیم که در یک سال حدوداً ۱۲۰۰۰ شرکت کننده در آزمون‌های تولیمو شرکت خواهند کرد. به‌دلیل آنکه، هر برنامه آزمون CAT باید هر سال به‌روزرسانی شود، این شبیه‌سازی برای یک دوره یک ساله انجام می‌گیرد.

جدول (۱) اطلاعات مربوط به تعداد شرکت‌کنندگان در آزمون‌های تولیمو

تعداد / دوره	زمان برگزاری	تعداد شرکت‌کنندگان	تعداد سؤال‌های آزمون
دوره ۱۱٤	دی ماه ۱۳۹۵	۱۳۵٦ نفر	۱٤۰
دوره ۱۱۵	بهمن ماه ۱۳۹۵	۵۰۵ نفر	۱٤۰
دوره ۱۱٦	اسفندماه ۱۳۹۵	۲۲۰۹ نفر	۱٤۰
دوره ۱۱۷	خرداد ماه ۱۳۹٦	۱۹۱۳ نفر	۱٤۰
دوره ۱۱۸	تیر ماه ۱۳۹٦	۵۳۹ نفر	۱٤۰
دوره ۱۱۹	مرداد ماه ۱۳۹٦	۱۷۲۹ نفر	۱٤۰
دوره ۱۲۰	شهریور ماه ۱۳۹٦	۸۵۲ نفر	۱٤۰
دوره ۱۲۱	مهر ماه ۱۳۹٦	۱۰۷۲ نفر	۱٤۰
دوره ۱۲۲	آبان ماه ۱۳۹٦	٤۹۵ نفر	۱٤۰
دوره ۱۲۳	آذر ماه ۱۳۹٦	۱۲۳۳ نفر	۱٤۰
جمع کل ۱۰ دوره	-	**۱۱۹۰۳**	۱٤۰۰

٤-۲- اطلاعات پیشین مربوط به خزانه سؤال واقعی آزمون تولیمو

در این قسمت، سؤال‌های هر بخش از آزمون تولیمو بر اساس مدل دو یا سه‌پارامتری لوجستیک به صورت جداگانه، مورد تحلیل روان‌سنجی قرار گرفتند. متوسط نتایج تحلیل در همه ۱۰ دوره برای هر بخش، در زیر ارائه شده است. همچنین، توانایی آزمونی‌ها بر اساس دو روش بیشینه درست‌نمایی و پسین مورد انتظار برآورد شد.

نتایج تحلیل روی ٤۰۰ سؤال مربوط به مبحث ساختار و نوشتار زبانی، نشان داد که بین پارامترهای a و b سؤال‌ها همبستگی معنی‌دار و منفی وجود دارد (

($r=-0.41, p=0.009$). توزیع پارامتر a، نرمال با میانگین ۰/۶۳ و انحراف استاندارد ۰/۲۵ است، نرمال بودن این توزیع با آزمون کولموگروف-اسمیرنف[1] ($KS=0.58, p=0.89$)تأیید شد. همچنین، پارامتر c نیز از توزیع نرمال $c \sim N(0.07, 0.022)$ پیروی کرد. این توزیع بهتر از توزیع‌های دیگر پارامتر c را توصیف می‌کرد. برای این‌که الگوی رابطه پارامترهای a و b در سؤال‌های خزانه سؤال عملیاتی و شبیه‌سازی‌شده به صورت دقیق‌تری مشخص شود، نخست، همه سؤال‌ها بر اساس مقادیر پارامتر b خود به سه گروه تقسیم شدند. پارامتر b در بخش ساختار و نوشتار زبانی دارای پراکندگی بسیار کمی بود و از ۱/۷۰- تا ۱/۷۰ توزیع شده بود (زیرا سؤال‌های این بخش، دشواری متوسطی داشتند)، بنابراین، گروه‌بندی (بر مبنای بهترین نقطه برش برای تفکیک توانایی آزمودنی‌ها انجام گرفت) دامنه پارامتر b در این بخش آزمون بسیار محدود بود (گروه اول؛ ۱/۷۰- تا ۰/۵۰-، گروه دوم؛ ۰/۵۰- تا ۰ و گروه سوم؛ ۰ تا ۱/۷۰). سپس، همبستگی بین پارامترهای a و b برای هر سه گروه با استفاده از نرم‌افزار SPSS-23 محاسبه شد. نتایج نشان داد که فقط در گروه سوم سؤال‌ها، یعنی، گروهی که پارامتر b آنها بالا است (سؤال‌های با مقادیر بالاتر از ۰ تا ۱/۷۰) بین پارامترهای a و b از لحاظ آماری همبستگی معنی‌دار منفی وجود داشت ($r=-0.58, p=0.002$). با بالارفتن سطح دشواری سؤال در این بخش آزمون تولیمو، ضریب تشخیص سؤال پایین‌تر آمده است. امّا در گروه اول همبستگی بین پارامتر a و b برابر با ($r=0.28, p=0.11$) و در گروه دوم ($r=0.10, p=0.77$) بود، که هیچ‌یک از نظر آماری معنی‌دار نبودند. بنابراین، تنها در گروه سوم یک رگرسیون ساده برای پیش‌بینی a توسط b محاسبه شد. معادله رگرسیون در گروه سوم برابر با $a_i = 0.70 - 0.24 b_i + e_i$ به‌دست آمد که e_i یک عنصر تصادفی‌ای بود که از توزیع نرمال $N(0, \sigma_e^2)$ پیروی می‌کند. در این توزیع σ_e^2 از طریق فرمول زیر که بر اساس ایده مک‌برد و وایس (۱۹۷۶) ایجاد شده، محاسبه شد: $\sigma_e = S_a \sqrt{(1-r_{ab}^2)} = 0.21\sqrt{(1-(-0.41)^2)} = 0.1915$.

---

[1]. Klmogorov – Smirnov

نتایج تحلیل روی ۵۰۰ سؤال مربوط به مبحث خواندن و درک مطلب، نشان داد که بین پارامترهای a و b سؤال‌ها همبستگی معنی‌داری وجود ندارد ($r=0.09, p=0.54$). توزیع پارامتر a، نرمال با میانگین ۱/۱۹ و انحراف استاندارد ۰/٦٤ است، نرمال بودن این توزیع از طریق آزمون کولموگروف- اسمیرنف ($KS=0.93, p=0.36$)تأیید شد. همچنین، پارامتر c از توزیع بتا[1] $c \sim N(0.21, 0.099)$ پیروی کرد. این توزیع بهتر از توزیع‌های دیگر پارامتر c را توصیف می‌کرد. ابتدا، همه سؤال‌ها بر اساس مقادیر پارامتر b آنها به سه گروه تقسیم‌بندی شدند. در پایین توزیع، پارامتر b سؤال‌ها دارای پراکندگی بسیار کمی بود و در کل پراکندگی پارامتر b در بخش خواندن و درک مطلب از ۰/۵۵- تا ۲/۵۷ توزیع شده بود و توزیع پارامتر b دارای کجی منفی زیادی بود. بنابراین، گروه‌بندی دامنه پارامتر b در این بخش آزمون بسیار متفاوت بود و متقارن نبود (گروه اول؛ ۰/۵۵- تا ۰/۵۵، گروه دوم؛ ۰/۵۵ تا ۱/۵ و گروه سوم؛ ۱/۵ تا ۳/۰). سپس، همبستگی بین پارامترهای a و b برای هر سه گروه محاسبه شد. نتایج نشان داد که، در گروه‌های اول ($r=0.37, p=0.03$) و دوم ($r=0.35, p=0.04$) بین پارامتر a و b رابطه معنی‌دار و مثبتی وجود داشت. امّا در گروه سوم هیچ رابطه‌ای ($r=0.18, p=0.60$) بین پارامتر a و b وجود نداشت. بنابراین، در گروه اول و دوم یک رگرسیون ساده برای پیش‌بینی a توسط b محاسبه شد. معادله رگرسیون برای گروه اول برابر با $a_i=0.96+0.18b_i+e_i$ بود. در این توزیع σ_e^2 برابر با: $\sigma_e = S_a\sqrt{(1-r_{ab}^2)} = 0.20\sqrt{(1-(0.37)^2)} = 0.1858$ به‌دست آمد. معادله رگرسیون برای گروه دوم برابر با $a_i=0.51+0.91b_i+e_i$ بود و σ_e^2 برابر با: $\sigma_e = S_a\sqrt{(1-r_{ab}^2)} = 0.71\sqrt{(1-(0.35)^2)} = 0.6650$ به‌دست آمد.

نتایج تحلیل روی ۵۰۰ سؤال مربوط به مبحث درک مطلب شفاهی و شنیداری، نشان داد که بین پارامترهای a و b سؤال‌ها مربوط به این بخش آزمون تولیمو همبستگی معنی‌دار و منفی وجود دارد ($r=-0.49, p=0.0001$)، توزیع پارامتر a ، نرمال با میانگین ۰/٥٤ و انحراف استاندارد ۰/٤١ است ($KS=0.82, p=0.52$). همچنین، پارامتر c سؤال‌ها از توزیع بتا

[1]. Beta

$c \sim N(0.08, 0.012)$ پیروی کرد. ابتدا، همه سؤال‌ها بر اساس مقادیر پارامتر b آنها به سه گروه تقسیم شدند. پارامتر b در بخش درک مطلب شفاهی یا شنیداری از پراکندگی قابل توجهی در قسمت راست (بالای) توزیع برخوردار بود. یعنی، مقادیر پارامتر b از ۰/٦۰- تا ٤/۰۰ توزیع شده است، توزیع پارامتر b دارای کجی منفی بسیار زیادی بود. بنابراین، گروه‌بندی دامنه پارامتر b در این بخش آزمون بسیار متفاوت است و متقارن نیست (گروه اول؛ ۰/٦۰- تا ۱/۰۰، گروه دوم؛ ۱/۰۰ تا ۱/۵ و گروه سوم؛ ۱/۵ تا ٤/۰۰). سپس، همبستگی بین پارامترهای a و b برای هر سه گروه محاسبه شد. نتایج نشان داد که، در گروه اول همبستگی بین پارامتر a و b مثبت و معنی‌دار بود ($r = 0.60, p = 0.01$)، یعنی هرچه ضریب دشواری افزایش یافته، ضریب تشخیص نیز افزایش یافته است. در گروه سوم همبستگی بین پارامتر a و پارامتر b، منفی و معنی‌دار بود ($r = -0.74, p = 0.0001$)، یعنی هرچه ضریب دشواری افزایش یافته، ضریب تشخیص کاهش پیدا کرده است. امّا در گروه دوم، هیچ رابطه معنی‌داری بین پارامتر a و b وجود نداشت ($r = -0.12, p = 0.74$). بنابراین، در گروه اول و سوم یک رگرسیون ساده برای پیش‌بینی پارامتر a توسط b محاسبه شد. معادله رگرسیون برای گروه اول برابر با $a_i = 0.39 + 0.72 b_i + e_i$ بود. مقدار σ_e^2 برابر با: $\sigma_e = S_a \sqrt{(1 - r_{ab}^2)} = 0.50 \sqrt{(1 - (0.60)^2)} = 0.40$ محاسبه شد. معادله رگرسیون برای گروه سوم برابر با $a_i = 0.69 - 0.10 b_i + e_i$ بود. مقدار σ_e^2 برابر با : $\sigma_e = S_a \sqrt{(1 - r_{ab}^2)} = 0.27 \sqrt{(1 - (-0.74)^2)} = 0.1816$ به‌دست آمد.

۵- تولید خزانه سؤال بهینه در آزمون CAT شبیه‌سازی شده، بر اساس اطلاعات به دست آمده از تحلیل سؤال‌های واقعی آزمون تولیمو مداد-کاغذی

الف) روش تصادفی (R): در موقعیتی از آزمون‌های CAT عملیاتی که در آن پارامترهای a و b سؤال‌ها از لحاظ آماری با یکدیگر همبستگی معنی‌داری نداشتند و مستقل از یکدیگر بودند، این روش تولید سؤال، برای شبیه‌سازی خزانه سؤال بهینه تولیمو، به‌کار رفت.

برای ایجاد یک سؤال بهینه، با استفاده از روش R، گام‌های زیر دنبال شد:

۱. پارامتر a_i و c_i از طریق توزیع‌های هدف خود در خزانه سؤال موجود آزمون تولیمو ایجاد شدند. در این پژوهش، در بخش اول آزمون تولیمو، توزیع a_i از

توزیع نرمال $a \sim N(0.63, 0.25)$ و توزیع پارامتر c نیز از توزیع نرمال $c \sim N(0.07, 0.022)$ پیروی کرد. در بخش دوم آزمون تولیمو، توزیع a_i از توزیع نرمال $a \sim N(1.19, 0.64)$ و توزیع پارامتر c از توزیع بتا $c \sim N(0.21, 0.099)$ پیروی کرد. در بخش سوم آزمون تولیمو، توزیع a_i از توزیع نرمال $a \sim N(0.45, 0.41)$ و توزیع پارامتر c از توزیع بتا $c \sim N(0.08, 0.012)$ پیروی کرد.

۲. باتوجه به اینکه هم پارامتر a_i وهم پارامتر c_i در سه بخش آزمون بر اساس گام بالا معلوم بودند، b_i با استفاده از معادله (۱٤) محاسبه شد، درصورتی‌که $\theta_{\max} = (\hat{\theta}_{ij})$ برآورد جدید توانایی باشد، بر اساس معادله (۱۳):.

$$\theta_{\max} = b_i + \frac{1}{Da_i} ln[0.5(1+\sqrt{1+8c_i})]$$

(۱۳)

در نتیجه

$$b_i = \hat{\theta}_{ij} - \frac{1}{Da_i} \ln \frac{1+\sqrt{1+8c_i}}{2}$$

(۱٤)

ب) روش آمیخته تصادفی و پیش‌بینی (MRP): همان‌طور که نام این روش نیز اشاره دارد، MRP، یک روش آمیخته است. قسمت روش تصادفی R آن که در بالا توصیف شد. قسمت روش پیش‌بینی (P)، در اصل، عقیده مک‌برید و وایس (۱۹۷٦)، را دنبال می‌کند که در این شیوه، خزانه سؤال «کاملی» به همراه پارامترهای سؤال بهینه براساس رگرسیون پارامترهای a_i روی پارامترهای b_i، شبیه‌سازی می‌شود. روش P براین واقعیت استوار است که پارامترهای a و b به‌طور معنی‌داری با یکدیگر همبسته‌اند (چانگ و ون‌درلیندن، ۲۰۰۳؛ ون‌درلیندن، اسکرامز و اسچنیپکا، ۱۹۹۹). یعنی، واریانس پارامتر a با افزایش پارامتر b، افزایش می‌یابد که این مشخص می‌کند با استفاده از تبدیلات لگاریتمی، پارامترهای a به‌طور خطی با پارامترهای b مرتبط می‌شوند (گو و ریکیسی، ۲۰۰۷). برای مدل‌یابی کردن این روابط، پارامتر a برای یک سؤال شبیه‌سازی شده برابر با تابع رگرسیونی تبدیل لگاریتمی پارامتر a (a') روی

پارامتر b است (ریکیسی، ۲۰۰٤). که $\varepsilon_i \sim N(0,\sigma^2)$ دارای توزیع نرمال است. با اضافه کردن یک عبارت خطا در تابع رگرسیونی، پراکندگی در پارامترهای a در روش برآورد خزانه سؤال، به‌وجود می‌آید.

$$a' = \log(a_i) = B_0 + B_1 b_i + \varepsilon_i \quad (۱۵)$$

$$a = \exp(a') \quad (۱٦)$$

برای ایجاد یک سؤال بهینه با استفاده از رویکرد MRP گام‌های زیر دنبال شد:

در بخش اول از آزمون تولیمو، برای هر سؤال، اگر مقدار b_i آن، که می‌توانست به وسیله $\hat{\theta}$ در هر مرحله از اجرای آزمون تقریب زده شود، پایین‌تر از تتا صفر بود، یعنی جزء گروه‌هایی بود که در آن همبستگی معنی‌داری بین پارامتر a و b وجود نداشت، برای ایجاد ویژگی‌های سؤال بهینه از روش R که در بالا توصیف شد، استفاده شد. در غیر این‌صورت، اگر مقدار b_i برابر یا بالاتر از صفر بود، یعنی، برای گروهی که همبستگی معنی‌داری بین پارامتر a و b وجود داشت، روش‌های زیر به‌کار رفت:

۱. پارامتر c_i از طریق توزیع هدف، یعنی توزیع نرمال $c \sim N(0.07, 0.022)$ ایجاد شد.

۲. پارامتر a_i از طریق $a_i = 0.70 - 0.24 b_i + e_i$ ، که b_i می‌توانست در هر گام انتخاب سؤال با $\hat{\theta}$ به دست آید و e_i از توزیع نرمال ($N(0, 0.1915^2)$) پیروی می‌کرد، ایجاد شد.

۳. پارامتر b_i دوباره با معادله (۱٤) محاسبه شد.

در بخش دوم از آزمون تولیمو، برای هر سؤال، اگر مقدار b_i آن، که می‌توانست به وسیله $\hat{\theta}$ در هر مرحله از اجرای آزمون تقریب زده شود، بالاتر از تتا ۱/۵ بود، یعنی جزء گروه‌هایی بود که در آن همبستگی معنی‌داری بین پارامتر a و b وجود نداشت، برای ایجاد ویژگی‌های سؤال بهینه از روش R که در بالا توصیف شد،

استفاده می‌شد. در غیر این‌صورت، اگر مقدار b_i برابر یا پایین‌تر از ۱/۵ بود، یعنی، برای گروه‌هایی که همبستگی معنی‌داری بین پارامتر a و b وجود داشت، روش‌های زیر به‌کار رفت:

۱. پارامتر c_i از طریق توزیع هدف، یعنی توزیع بتا $c \sim N(0.21, 0.099)$ ایجاد شد.

۲. پارامتر a_i در گروه اول از طریق $a_i = 0.96 + 0.18b_i + e_i$ ، و در گروه دوم از طریق $a_i = 0.51 + 0.91b_i + e_i$ که b_i می‌توانست در هر گام انتخاب سؤال از طریق $\hat{\theta}$ به دست آید و در گروه اول، e_i از توزیع نرمال ($N(0, 0.1858^2)$) و در گروه دوم، e_i از توزیع نرمال ($N(0, 0.6650^2)$) پیروی می‌کرد، ایجاد شد.

۳. پارامتر b_i دوباره با معادله (۱٤) محاسبه شد.

در بخش سوم از آزمون تولیمو، برای هر سؤال، اگر مقدار b_i آن، که می‌توانست به وسیله $\hat{\theta}$ در هر مرحله از اجرای آزمون تقریب زده شود، بین تتا ۱/۰ تا ۱/۵ بود، یعنی جزء گروه‌هایی بود که در آن همبستگی معنی‌داری بین پارامتر a و b وجود نداشت، برای ایجاد ویژگی‌های سؤال بهینه از روش R که در بالا توصیف شد، استفاده می‌شد. در غیر این‌صورت، اگر مقدار b_i برابر یا بالاتر از ۱/۵ بود یا برابر یا پایین‌تر از ۱/۰ بود، یعنی، برای گروه‌هایی که همبستگی معنی‌داری بین پارامتر a و b وجود داشت، روش‌های زیر به‌کار رفت:

۱. پارامتر c_i از طریق توزیع هدف، یعنی توزیع بتا $c \sim N(0.08, 0.012)$ ایجاد شد.

۲. پارامتر a_i در گروه اول از طریق $a_i = 0.39 + 0.72b_i + e_i$ ، و در گروه سوم از طریق $a_i = 0.69 - 0.10b_i + e_i$ که b_i می‌توانست در هر گام انتخاب سؤال با $\hat{\theta}$ به دست آید و در گروه اول، e_i از توزیع نرمال ($N(0, 0.40^2)$) و در گروه سوم، e_i از توزیع نرمال ($N(0, 0.1816^2)$) پیروی می‌کرد، ایجاد شد.

۳. پارامتر b_i دوباره با معادله (۱٤) محاسبه شد.

ج) روش حداقل آگاهی آزمون (MTI)

روش MTI بر این فرض است که وقتی آزمون‌های CAT مجزایی که از خزانه سؤال سرهم می‌شود، عیناً بتواند میزان آگاهی کافی مورد نظر را برای اندازه‌گیری توانایی آزمودنی فراهم کند، خزانه سؤال، بهینه محسوب می‌شود. اگر آزمون میزان آگاهی بیشتری بتواند ایجاد کند، آزمون با دقت بیشتری می‌تواند سطح توانایی آزمودنی را برآورد کند. با این وجود، برای ساخت آزمونی که میزان آگاهی بیشتری بتواند ایجاد کند، به سؤال‌های با ضرایب تشخیص بالا نیاز داریم. در صورتی که معمولاً ساخت این نوع سؤال‌ها گران و دشوار است، مخصوصاً اگر سؤال‌ها آسان باشد، این دشواری دو چندان می‌شود. روش MTI این اطمینان را به ما می‌دهد که آزمون‌ها دارای دقت کافی برای برآورد توانایی هستند، ولی شامل سؤال‌های با ضرایب تشخیص بسیار بالا نیستند. در روش MTI، یک مقدار آگاهی هدف برروی دامنه‌ای از مقیاس θ قرار داده می‌شود. هر سؤالی که برای آزمودنی اجرا می‌شود در مقدار آگاهی هدف آزمون، سهیم است. برای اجرای رویکرد حداقل آگاهی تست، مرحله اصلی، تعیین آگاهی هدف تست است. براساس اطلاعات پیشین در مورد تست از روی خزانه عملیاتی موجود و توزیع برآوردهای توانایی، حداقل آگاهی هدف تست می‌تواند با دو معادله (۱۷) و (۱۸) تعیین شود:

$$S_e = S_O\sqrt{1-r_{xx'}} \qquad (۱۷)$$

$$I_{\hat{\theta}} = \frac{1}{S_e^2} \qquad (۱۸)$$

S_O؛ انحراف استاندارد برآوردهای توانایی، S_e ؛ خطای استاندارد برآورد، و $I_{\hat{\theta}}$ ؛ آگاهی تست را نشان می‌دهد. زمانی‌که $I_{\hat{\theta}}$ معلوم شد، پس از آن آگاهی مورد انتظاری که هر سؤال باید فراهم کند، می‌تواند با تقسیم $I_{\hat{\theta}}$ بر طول تست به دست آید. با توجه به این واقعیت که آگاهی واقعی‌ای که یک سؤال می‌تواند ایجاد کند، مشروط به برآورد جدید توانایی، ممکن است کاملاً مطابق با آگاهی مورد انتظار

نباشد، بنابراین، آگاهی هدف سؤال باید هر بار پس از اینکه یک سؤال اجرا می‌شود، به‌روز شود. معادله (۱۹) برای به‌روز کردن آگاهی هدف سؤال به‌کار می‌رود. T، آگاهی تست را نشان می‌دهد و L، طول تست را نشان می‌دهد.

$$I_i = \frac{T_{target} - T_{admin}}{L_{target} - L_{admin}} \qquad (۱۹)$$

در این پژوهش، با استفاده از تحلیل روی داده‌های مربوط به خزانه سؤال موجود آزمون تولیمو (۱۰ دوره)، حداقل آگاهی هدف آزمون برای بخش اول، تقریباً برابر با ۱۲/۱، برای بخش دوم برابر با ۲۹/۳ و برای بخش سوم برابر با ۱۵/۱ به دست آمد. در این پژوهش، آگاهی هدف آزمون با توجه به سطوح متفاوت توانایی محاسبه شد. برای بخش اول آزمون تولیمو، برای آزمودنی‌هایی با توانایی بین ۱/۷۰- و ۰/۵۰-، یا ۰ تا ۱/۷۰ آگاهی هدف آزمون برابر با ۱۰/۱۲ به دست آمد، $S_e = 0.3142$ و $S_e^2 = (0.3142)^2 = 0.09873$ بنابراین، $I_{\hat{\theta}} = \frac{1}{0.09873} = 10.12$. همچنین، برای آزمودنی‌هایی با توانایی‌های واقعی بین ۰/۵۰- تا ۰ ، آگاهی هدف آزمون برابر با ۱۲/۷۰ به دست آمد، $S_e = 0.2805$ و $S_e^2 = (0.2805)^2 = 0.07869$. بنابراین، $I_{\hat{\theta}} = \frac{1}{0.07869} = 12.7$ و برای بقیه آزمودنی‌ها، آگاهی هدف آزمون برابر با ۸/۱۲ است، ($S_e = 0.3508$ و $S_e^2 = (0.3508)^2 = 0.1231$ بنابراین، $I_{\hat{\theta}} = \frac{1}{0.1231} = 8.12$ هنگامی که آگاهی آزمون معلوم شد، آگاهی مورد انتظاری که هر سؤال باید ایجاد کند، از طریق معادله (۲۰) به‌دست آمد. لازم به ذکر است که دو نمره برش ۱/٦۲٤٥- و ۱/۰۸۸۲۳۵ به همراه نمره ۵۳۹۷/.-، سه نمره برشی بودند که در این پژوهش برای تعیین جایگاه آزمودنی‌ها در سطوح متفاوت توانایی آزمون تولیمو، بخش ساختار و نوشتار زبانی به کار رفتند. این نمره‌های برش از مطالعات گوناگون در مورد توانایی و مهارت زبان و ساختار نوشتاری، مانند مطالعه هی و ریکیسی، ۲۰۱۰؛ گرفته شدند. همین روند برای دو بخش دیگر آزمون انجام گرفت (به جدول (۲) مراجعه شود).

در روش MTI سؤال‌ها در سه گام ایجاد شدند (ریکیسی، ۲۰۰٤، گو و ریکیسی، ۲۰۰۷؛ هی و ریکیسی، ۲۰۱۰):

در بخش اول آزمون تولیمو:

۱. پارامتر c_i با توزیع هدف، یعنی توزیع نرمال $c \sim N(0.07, 0.022)$ ایجاد شد.

۲. پارامتر a_i با معادله (۲۰) ایجاد شد.

$$a_i = \sqrt{\frac{8(1-c_i)^2 I_i}{D^2[1-20c_i-8c_i^2+(1+8c_i)^{3/2}]}} \quad (۲۰)$$

که در واقع، از سازمان‌دهی دوباره معادله (۲۱) به‌دست آمده است. البته M_i می‌تواند به جای I_i در معادله (۲۰) قرار گیرد:

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1-20c_i-8c_i^2+(1+8c_i)^{3/2}] \quad (۲۱)$$

۳. با توجه به اینکه هم پارامتر a_i و هم پارامتر c_i معلوم بودند، پارامتر b_i نیز با استفاده از معادلات ۱۳ و ۱٤ محاسبه شد.

به‌منظور جلوگیری از طولانی شدن این بخش، سایر ویژگی‌های مربوط به آگاهی آزمون در جدول (۲) آورده شده است:

جدول (۲) اطلاعات پیشین مربوط به میزان آگاهی تست در خزانه سؤال عملیاتی مربوط به سه بخش آزمون تولیمو

آماره / دامنه توانایی	ساختار و نوشتار زبانی			
	S_e	S_e^2	$I_{\hat{\theta}}$	$I_{\hat{\theta}}$
۰/۵۰- تا ۰	۰/۲۸۰۵	۰/۰۷۸۶۹	$I_{\hat{\theta}}=\frac{1}{0.07869}$	۱۲/۷۰
۱/۷۰- تا ۰/۵۰- & ۰ تا ۱/۷۰	۰/۳۱۴۲	۰/۰۹۸۷۳	$I_{\hat{\theta}}=\frac{1}{0.09873}$	۱۰/۱۲
٤- تا ۱/۷۰- & ۱/۷۰ تا ٤	۰/۳۵۰۸	۰/۱۲۳۱	$I_{\hat{\theta}}=\frac{1}{0.1231}$	۸/۱۲
آماره / دامنه توانایی	خواندن و درک مطلب			
	S_e	S_e^2	$I_{\hat{\theta}}$	$I_{\hat{\theta}}$
۰/۵۵ تا ۱/۵	۰/۱۸۳۱	۰/۰۳۳۵٤	$I_{\hat{\theta}}=\frac{1}{0.03354}$	۲۹/۸۱
۰/۵۵- تا ۰/۵۵ & ۱/۵ تا ۳/۰	۰/۲۱۷۵	۰/۰٤۷۳٤	$I_{\hat{\theta}}=\frac{1}{0.04734}$	۲۱/۱۲
٤- تا ۰/۵۵- & ۳ تا ٤	۰/۳۱۲۹	۰/۰۹۷۹٤	$I_{\hat{\theta}}=\frac{1}{0.09794}$	۱۰/۲۱
آماره / دامنه توانایی	درک مطلب شفاهی و شنیداری			
	S_e	S_e^2	$I_{\hat{\theta}}$	$I_{\hat{\theta}}$
۱/۰ تا ۱/۵	۰/۲۵۷۳	۰/۰٦٦۲۲	$I_{\hat{\theta}}=\frac{1}{0.06622}$	۱۵/۱
۰/٦۰- تا ۱/۰۰ & ۱/۵ تا ٤	۰/۲۸۵۷	۰/۰۸۱٦۳	$I_{\hat{\theta}}=\frac{1}{0.08163}$	۱۲/۲۵
۰/٦۰- تا ٤-	۰/۳۷۲٦	۰/۱۳۸۸	$I_{\hat{\theta}}=\frac{1}{0.1388}$	۷/۲

یافته‌ها

قسمت یافته‌ها، در دو بخش ارائه شده است: در بخش اول، ابتدا سؤال‌های مربوط به خزانه سؤال موجود آزمون تولیمو مورد تجزیه و تحلیل روان‌سنجی قرار گرفت. زیرا

یکی از شروط اجرای آزمون به شکل سنجش انطباقی کامپیوتری، مدرج کردن سؤال‌های آزمون بر اساس نظریه سؤال پاسخ است. از این‌رو، در این پژوهش نیز ابتدا سؤال‌های موجود درخزانه سؤال آزمون تولیمو در سازمان سنجش آموزش کشور، مورد تحلیل قرار گرفته است. از آنجایی‌که ارائه تمام ویژگی‌های سؤال‌های ۱۰ دوره به‌دلیل محدودیت صفحات مقاله امکان‌پذیر نیست، همان‌طور که در قسمت روش‌شناسی نیز بیان شد، دوره ۱۲۳ به نمایندگی از سایر دوره‌ها در این قسمت ارائه شده است. در بخش دوم، نتایج مربوط به مطالعه شبیه‌سازی شده CAT بهینه و طراحی CAT عملیاتی با درنظرگرفتن روش برآورد پارامتر توانایی، ملاک توقف آزمون، روش طراحی خزانه سؤال و روش کنترل مواجهه سؤال ارائه شده است. همچنین، نتایج مربوط به مقایسه‌های انجام گرفته برای دقت برآورد پارامتر توانایی در اجراهای مداد-کاغذی و سنجش انطباقی کامپیوتری مورد بررسی قرار گرفته است. بخش دوم از سه زیربخش عمده تشکیل شده است. در زیربخش اول، آزمون CAT عملیاتی بر اساس ویژگی‌های از پیش تعیین‌شده در بخش روش‌شناسی طراحی شده است. همچنین، به‌منظور استخراج بهترین روش برآورد پارامتر توانایی و ملاک توقف آزمون، در همین بخش، نتایج مربوط به دقت برآورد توانایی آزمون‌ها بر اساس روش برآورد بیشینه درست‌نمایی (ML) و پسین مورد انتظار (EAP) و ملاک توقف طول ثابت آزمون (اجرای ۱٤۰ سؤال برای همه آزمودنی‌ها) و خطای استاندارد ثابت آزمون (۰/۳۸۵، ۰/۳۱۵ و ۰/۲۲) یعنی ۸ حالت ممکن (۲*٤)، ارائه شده است، تا بهترین روش برآورد و ملاک توقف آزمون مشخص شود. دقت برآورد با ملاک‌هایی چون، مقادیرهمبستگی، اریب، RMSE و ADD توانایی در اجراهای مداد-کاغذی و سنجش انطباقی کامپیوتری ارزیابی شده است. در زیربخش دوم، بر اساس بهترین روش برآورد پارامتر توانایی و ملاک توقف آزمون که در زیربخش اول استخراج شد، خزانه‌های سؤال بهینه شبیه‌سازی شده بر اساس روش مونت‌کارلو ریکیسی (۲۰۰۳) مبتنی بر مدل دو و سه‌پارامتری و برنامه‌نویسی خطی WDM طراحی شده است (لازم به ذکر است که امکان مقایسه تمام حالت‌های ممکن روش برآورد پارامتر توانایی و ملاک توقف آزمون در سه روش طراحی خزانه سؤال بهینه به دلیل محدودیت صفحات در مقاله فعلی وجود نداشت، در گزارش‌های دیگر به این موضوع خواهیم پرداخت). در این زیربخش ویژگی‌هایی همچون؛ اندازه خزانه سؤال، ویژگی‌های آماری و غیر آماری خزانه‌ها، قیود محتوایی بررسی شده است. در

زیربخش سوم، عملکرد CATهایی که بر اساس خزانه‌های سؤال بهینه شبیه‌سازی شده با عملکرد CAT که بر اساس خزانه سؤال عملیاتی طراحی شدند، مقایسه شده است، این عملکردها از طریق ملاک‌های ارزیابی یک CAT بهینه شامل، میانگین آگاهی آزمون در هر سطح θ، سوگیری یا بایاس، میانگین مجذور خطا (MSE)، به منظور سنجش دقت برآورد θ، میزان چولگی یا کجی توزیع نرخ مواجهه سؤال، درصد سؤال‌های بیش از حد مواجهه شده، نرخ همپوشی سؤال و درصد سؤال‌های کم‌مواجهه شده، به منظور محاسبه شاخص امنیت آزمون، مورد ارزیابی قرار گرفته‌اند (چانگ و ینگ، ۱۹۹۹ ؛ ریکیسی ۲۰۰۵). در این قسمت، ذکر چندین نکته قبل از گزارش نتایج ضرورت دارد. از این پس به CATهایی که بر اساس خزانه‌های سؤال بهینه مبتنی بر روش bin-and-union و برنامه‌نویسی خطی ایجاد می‌شود، بر اساس اختصار ROP (range-optimal item pool) گفته می‌شود. این اختصار از پژوهش ریکیسی و هی (۲۰۰۹b) گرفته شده است. همچنین، به CATهایی که براساس خزانه‌های سؤال عملیاتی طراحی شده‌اند، به صورت اختصار، OP (operational item pool) گفته می‌شود، این اصطلاح از پژوهش گو و ریکیسی (۲۰۰۷) گرفته شده است. در مطالعه CAT شبیه‌سازی شده، خزانه سؤال بهینه، با دستکاری دو عامل ساخته شده‌اند. این دو عامل عبارت‌اند از: روش ایجاد سؤال بهینه (R، MRP، MTI) و کنترل یا عدم کنترل مواجهه بیش از حد سؤال در CAT، با دستکاری این دو عامل ٦ خزانه سؤال بهینه یا (ROP_1, ROP_2, ROP_3, …, ROP_6) ایجاد شده است. همچنین، از آنجا که پهنای b-bin ها، یکی از عوامل مهم در دقت اندازه‌گیری خزانه ایجاد شده، است و بر ویژگی‌هایی همچون تعداد سؤال و توزیع پارامترهای سؤال‌ها نیز تأثیر می‌گذارد، بنابراین، پهنای ۰/۲ که معادل ۰/۹۹ دقت اندازه‌گیری دارد، با یک میزان ثابت در دامنه تغییر پارامتر a یا تغییر در بیشینه آگاهی برابر با ۰/٤، وارد تحلیل شده‌اند. در همه CAT های شبیه‌سازی شده تعادل محتوایی وارد شده است. بنابراین، در زیربخش دوم و سوم، نتایج به دست آمده از مطالعه شبیه‌سازی صورت گرفته، در ٦ طرح متنوع گزارش شده است، که نتایج حاصل از مطالعه شبیه‌سازی شده CAT، با نتایج CAT طراحی‌شده بر اساس خزانه سؤال عملیاتی، مقایسه شده است.

جدول (۳) خلاصه طرح‌های مطالعه شبیه‌سازی شده را شرح می‌دهد:

جدول (۳) طرح‌های خزانه سؤال شبیه‌سازی شده

موقعیت condition	خزانه سؤال item pool	روش ایجاد سؤال item generation methods	پهنای b-bin	کنترل مواجهه Exposure control	تعادل محتوایی Content Balancing	تعداد سؤال
۱	ROP_1	R	۰/۲	without-Exposure control	Content Balancing	۱٤۰
۲	ROP_2	MRP	۰/۲	without - Exposure control	Content Balancing	۱٤۰
۳	ROP_3	MTI	۰/۲	without - Exposure control	Content Balancing	۱٤۰
٤	ROP_4	R	۰/۲	Exposure control	Content Balancing	۱٤۰
٥	ROP_5	MRP	۰/۲	Exposure control	Content Balancing	۱٤۰
٦	ROP_6	MTI	۰/۲	Exposure control	Content Balancing	۱٤۰

۱- تحلیل سؤال‌های آزمون تولیمو (دوره ۱۲۳) بر اساس روش IRT

الف) تحلیل سؤال‌های بخش ساختار و نوشتار زبانی

برای تجزیه‌وتحلیل سؤال‌های بخش ساختار و نوشتار زبانی بر اساس نظریه IRT به شیوهٔ زیر عمل شد: ابتدا مفروضه‌های اساسی و اولیه نظریه IRT یعنی مفروضه‌های تک‌بعدی بودن و استقلال موضعی بررسی شد. تک‌بعدی بودن به این معناست که مدل برای هر آزمودنی θ جداگانه‌ای را ارائه می‌دهد و با هر عامل دیگری که پاسخ به سؤال را تحت تاثیر قرار می‌دهد به‌عنوان خطای تصادفی یا عامل مزاحم که مخصوص آن سؤال است و در سایر سؤال‌ها مطرح نیست، برخورد می‌کند. نقض این مفروضه ممکن است به برآورد نادرست پارامتر یا خطای استاندارد بالا منجر شود. نتایج تحلیل بعدیت با نرم‌افزار TESTFACT و به همراه بزرگ‌ترین ریشه‌های مشخصه ماتریس همبستگی نشان می‌دهد که ارزش ویژه هر یک از عوامل استخراج شده بخش ساختار و نوشتار زبانی آزمون تولیمو به ترتیب ٤/۲٥، ۱/۷۲ و ۱/٦۰ است. به‌طوری‌که، عامل اول ۱۱/۸٦، عامل دوم ٤/۰۸ درصد و عامل سوم ۳/۰۱ درصد از واریانس را تبیین کرده است. از آنجا که عامل اول نسبت به دو عامل دیگر بیش از ۳

برابر قدرت تبیین‌کنندگی دارد، بنابراین این بخش از آزمون را تک‌بعدی در نظر خواهیم گرفت (همبلتون، ۱۹۹۱). همچنین، به‌منظور کسب اطمینان بیشتر از تک‌بعدی بودن این بخش آزمون، از نرم‌افزار NOHARM نیز استفاده شد. این برنامه سه شاخص برازش برای بررسی بعدیت آزمون ارائه می‌دهد. ریشه میانگین مجذورات باقی‌مانده‌ها[1] (RMSR)، مجموع مجذور باقی‌مانده‌ها[2] (SSR) و شاخص خوبی برازندگی تاناکا.[3] مقادیر کوچک‌تر ریشه میانگین مجذورات باقی‌مانده‌ها و همچنین مقادیر کوچک‌تر مجموع مجذور باقی‌مانده‌ها بیانگر برازش مدل با داده‌ها است. همچنین، در خصوص شاخص تاناکا، مقادیر ۰/۹۰ برازش قابل قبول، ۰/۹۵ بسیار خوب و ۱ برازش کامل را نشان می‌دهد (مینایی و فلسفی‌نژاد، ۱۳۸۹). همان‌طور که نتایج جدول (٤) نیز نشان می‌دهد، مدل تک‌بعدی بهترین برازش را با داده‌های این بخش دارد و نسبت به مدل چندبعدی به‌صرفه‌تر است. همبلتون (۱۹۹۱) ذکر می‌کند که چنانچه فرض تک‌بعدی بودن برقرار باشد، مفروضهٔ استقلال موضعی نیز برقرار است. با توجه به این مطلب در آزمون مذکور مفروضهٔ استقلال موضعی نیز برقرار است. تحلیل سؤال‌ها بر اساس مدل سه‌پارامتری، برازش نامناسبی با داده‌های تجربی نشان داد. نتایج تحلیل نشان داد که همه سؤال‌ها به جز سؤال‌های ۱۵، ۳٦، ۳۷، ۳۹ با مدل دوپارامتری برازش کامل داشتند و سطح معنی‌داری هیچ یک از سؤال‌ها کوچک‌تر از ۰/۰۱ نبود. با توجه به اینکه مدل دوپارامتری بهترین برازش را با داده‌های بخش اول آزمون تولیمو دارد، این برازش نشان می‌دهد که این مدل برای برآورد پارامترهای سؤال‌ها بسیار بهینه است و دقت عمل برآورد بسیار بالا است. لازم به ذکر است که در این بخش آزمون، اغلب سؤال‌ها ضریب دشواری منفی داشتند که نشان‌دهنده آسان بودن سؤال‌های این بخش آزمون برای آزمودنی‌ها است. در نمودار (۲) ماتریس منحنی‌های ویژگی (ICC) تمامی سؤال‌های آزمون ساختار و نوشتار زبانی در مدل دوپارامتری ارائه شده است. شیب ICC ها نشان‌دهنده پارمتر a است بنابراین سؤال‌ها با شیب تندتر مناسب‌تر هستند، زیرا تغییر در سطح صفت با احتمال پاسخ رابطه معنی‌داری دارد. سؤال‌های دشوارتر در ICC به سمت منتهی‌الیه بالای

---

[1]. Root mean square of residuals
[2]. Sum of squares of residuals
[3]. Tanaka index of goodness of fit indexes

مقیاس حرکت می‌کنند. همچنین می‌توان با استفاده از ICC تعیین کرد که این سؤال بیشترین میزان آگاهی را در کدام دامنه فراهم می کند. توجه کنید که سؤال‌ها از سمت چپ به راست شماره گذاری شده‌اند.

ب) تحلیل سؤال‌های بخش خواندن و درک مطلب

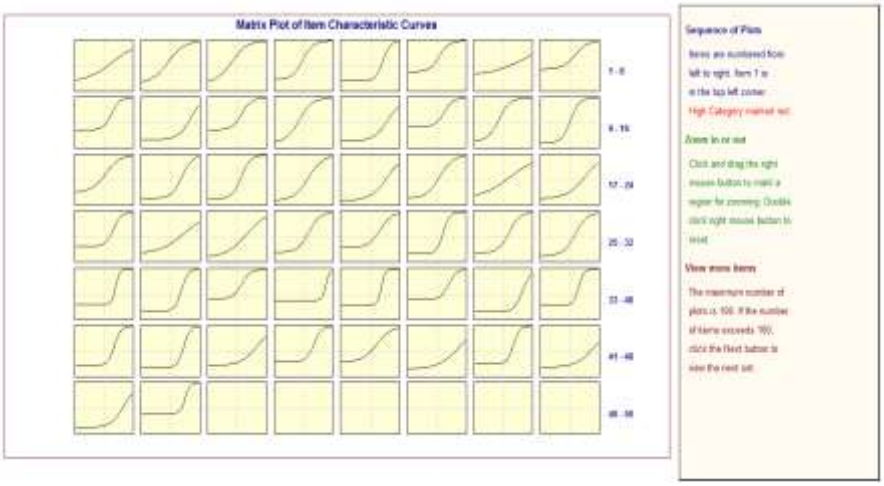
نتایج تحلیل توسط نرم‌افزار TESTFACT و به همراه بزرگ‌ترین ریشه‌های مشخصه ماتریس همبستگی برای بخش دوم آزمون تولیمو نشان می‌دهد که ارزش ویژه هر یک از عوامل استخراج شده بخش خواندن و درک مطلب آزمون تولیمو به ترتیب ٤/١٢، ١/٨٣ و ١/٧١ است، عامل اول ١٢/١٦، عامل دوم ٣/١٤ درصد و عامل سوم ٢/٩٩ درصد از واریانس را تبیین کرده است. از آنجا که عامل اول نسبت به دو عامل دیگر بیش از ۳ برابر قدرت تبیین‌کنندگی دارد، این بخش از آزمون را تک‌بعدی در نظر خواهیم گرفت. همچنین، طبق نتایج جدول (٤)، مدل تک‌بعدی بهترین برازش را با داده‌های این بخش نیز نشان می‌دهد. زیرا شاخص‌های SSR، RMSR و تاناکا برای مدل تک‌بعدی بهترین برازش را نشان می‌دهد. همچنین، همان‌طور که در بخش روش‌شناسی نیز بیان شد، به‌دلیل اینکه سؤال‌های این بخش به‌صورت Reading ارائه شده‌اند، برای اطمینان از استقلال موضعی داده‌های این بخش، تنها به تک‌بعدی بودن آن اکتفا نشده و برای این بخش آزمون، شاخص Q_3 محاسبه شد. نمودار (۸) و جدول (۵) نتایج مربوط به توزیع Q_3 (استقلال موضعی) مربوط به دوره ۱۲۳ را نشان می‌دهد. بنابراین از آنجایی‌که، میانه و میانگین توزیع Q_3 منفی است و همچنین، همه سؤال‌ها به غیر از یک سؤال مقدار Q_3 کمتر از قدرمطلق ۰/۲ دارد، بنابراین می‌توان بیان کرد که مفروضه استقلال موضعی در این بخش آزمون برقرار است و امکان مدرج‌سازی سؤال‌ها بر اساس مدل سه‌پارامتری لوجستیک وجود دارد و نیازی به تحلیل سؤال‌ها بر اساس مدل‌های testlet نیست. نتایج تحلیل نشان داد همه سؤال‌ها با مدل سه‌پارامتری برازش کامل دارند و سطح معنی‌داری هیچ یک از سؤال‌ها کوچک‌تر از ۰/۰۱ نیست. در این بخش آزمون، اغلب سؤال‌ها ضریب دشواری متوسط و بالا دارند که نشان‌دهنده دشواری نسبی سؤال‌های این بخش از آزمون است. نمودار (۳) منحنی‌های ویژگی (ICC) تمامی سؤال‌های آزمون خواندن و درک مطلب در مدل سه‌پارامتری را نشان می‌دهد.

ج) تحلیل سؤال‌های بخش درک مطلب شفاهی یا شنیداری

نتایج تحلیل توسط نرم‌افزار TESTFACT و به همراه بزرگ‌ترین ریشه‌های مشخصه ماتریس همبستگی نشان داد که ارزش ویژه هر یک از عوامل استخراج شده بخش درک مطلب شفاهی یا شنیداری آزمون تولیمو به ترتیب ۹/۸۵، ۳/٤٤ و ۲/٦٦ است، عامل اول ۲۰/۸٦، عامل دوم ٦/۰۲ درصد و عامل سوم ٤/۹۹ درصد از واریانس را تبیین کرده است. از آنجا که عامل اول نسبت به دو عامل دیگر بیش از ۳ برابر قدرت تبیین‌کنندگی دارد، این بخش از آزمون را تک‌بعدی در نظر خواهیم گرفت. همچنین، طبق نتایج جدول (٤)، شاخص‌های SSR، RMSR و تاناکا برای مدل تک‌بعدی بهترین برازش را نشان می‌دهد. بنابراین، مفروضهٔ استقلال موضعی نیز برقرار است. تحلیل نتایج با مدل دوپارامتری برازش مناسبی با داده‌های تجربی داشت. نتایج نشان داد همه سؤال‌ها به غیر از (سؤال‌های ۳، ۵، ۹، ۱۱، ۱٤، ۱۵، ۲۱، ۲۹، ۳۲، ۳۳، ۳۷، ٤۲، ٤۳، ٤۵، ٤۹ و ۵۰ که با هیچ مدلی برازش ندارند)، با مدل دوپارامتری برازش کامل دارند و سطح معنی‌داری هیچ یک از سؤال‌ها کوچک‌تر از ۰/۰۱ نیست. سؤال‌های این بخش آزمون، اغلب ضریب دشواری متوسط و بالا دارند که نشان‌دهنده دشواری نسبی سؤال‌های این بخش از آزمون برای شرکت‌کنندگان است. نمودار (٤) منحنی‌های ویژگی (ICC) تمامی سؤال‌های آزمون درک مطلب شفاهی یا شنیداری در مدل دوپارامتری را نشان می‌دهد.

نمودار (۲) ماتریس ICC مربوط به سؤال‌های بخش ساختار و نوشتار زبانی

نمودار (۲)، ویژگی‌های سؤال‌ها و ICC آنها را به صورت کلی نشان می‌دهد. نتایج مربوط به ICC سؤال‌های مربوط به سؤال ۱۰، ۱۵، ۳۶، ۳۷، ۳۹ میزان آگاهی بسیار پایینی دارد. سؤال‌های ۱۰ و ۳۹ نسبت به سایر سؤال‌ها دشوارتر است، سؤال‌های ۱۵ و ۳۷ نسبت به سایر سؤال‌ها آسان‌تر است و ضریب تشخیص پایینی دارند. سؤال ۳۶ دارای ضریب تشخیص پایینی در بیشتر طیف توانایی است. سؤال ۳۹، درجه دشواری خارج از دامنه طبیعی دارد. نتایج مربوط به منحنی آگاهی‌بخش ساختار و نوشتار زبانی نشان می‌دهد که آزمون در تتا برابر با ۰/۵- بیشترین آگاهی (نمودار (۵)) را ایجاد می‌کند. به نوعی بخش گرامر آزمون تولیمو، آسان است. البته برای آزمون ملاک‌محور-مهارتی این مقدار طبیعی است.

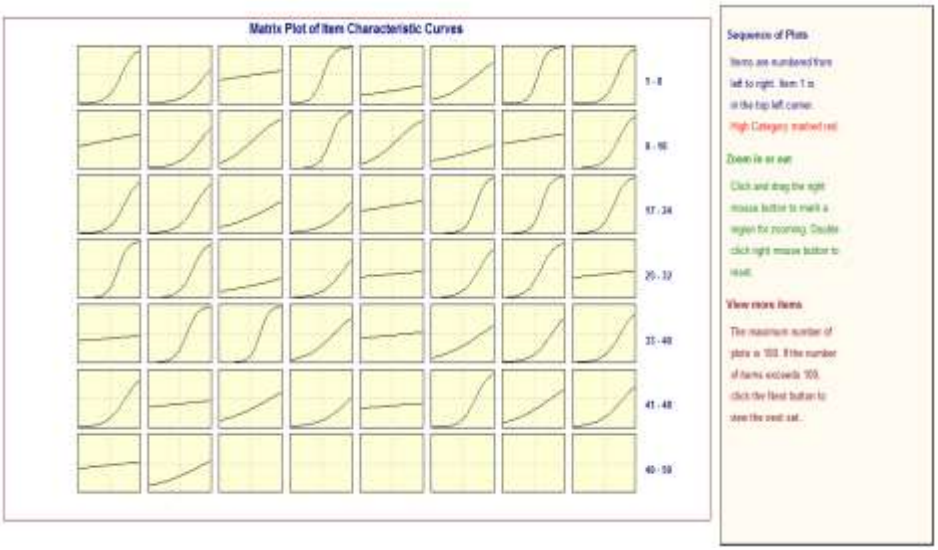


نمودار (۳) ماتریس ICC مربوط به سؤال‌های بخش خواندن و درک مطلب

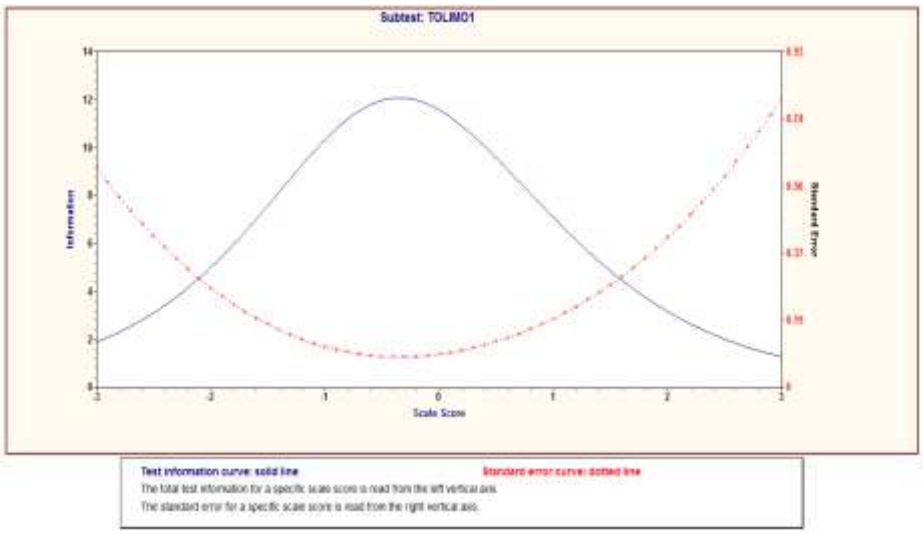
نمودار (۳)، ویژگی‌های سؤال‌ها و ICC آنها را به صورت کلی نشان می دهد. نتایج مربوط به ICC سؤال‌های مربوط به سؤال۱، ۷، ۸، ۱۰ و ۳۳ به‌صورت اجمالی بررسی می‌شود. سؤال‌های ۱، ۷ و ۸ و سؤال‌هایی که ICC مشابه با آنها دارند، نسبت به سایر سؤال‌ها دارای ضریب حدس بزرگ‌تری هستند. سؤال ۱۰ و سایر سؤال‌های مشابه با آن، نسبت به سایر سؤال‌ها دشوارتر است و ضریب حدس آن در دامنه پایین تتا و توانایی زیاد است. سؤال در دامنه بالای توانایی دارای میزانی آگاهی است. سؤال ۳٦ دارای ضریب تشخیص پایینی در بیشتر طیف توانایی است. سؤال ۳۳، ۳٦ و ۳۷ دارای میزان آگاهی بسیار زیادی در دامنه توانایی بالا است. در مجموع میزان آگاهی کل بخش خواندن و درک مطلب (نمودار (٦)) در توانایی برابر با ۱/۵ به اوج خود می‌رسد که نشان‌دهنده دشواری نسبی این بخش آزمون برای آزمودنی‌ها است. هیستوگرام توزیع توانایی (نمودار (۱)) نیز منحنی آگاهی کل این بخش را تأیید می‌کند.

نمودار (٤)، ویژگی‌های سؤال‌ها و ICC آنها را به صورت کلی نشان می دهد. نتایج مربوط به ICC سؤال‌هایی که با هیچ‌یک از مدل‌های تک، دو و سه‌پارامتری برازش نداشتند (۳، ۹، ۱۵، ۲۱، ۲۹، ۳۲، ۳۳، ۳۷، ٤۲، ٤۵ و ٤۹) مشابه یکدیگر است، بنابرین، یکی از سؤال‌های آن بررسی می‌شود. این مجموعه از سؤال‌های بخش درک مطلب شفاهی دارای ضریب تشخیص بسیار پایین و ضریب حدس بسیار بالا هستند.

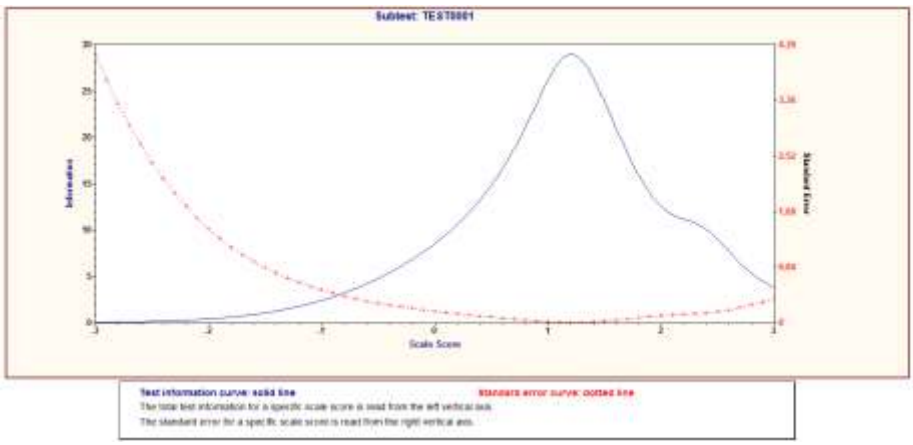
از این‌رو، این نوع سؤال‌ها از خزانه سؤال عملیاتی حذف شدند. امّا ICC سؤال‌های شبیه سؤال اول نشان می‌دهد که این سؤال‌ها به‌خوبی طراحی شدند و ضریب حدس آنها صفر است و ضریب تشخیص بسیار بالایی برای تفکیک افراد با توانایی بالا از افراد با توانایی پایین دارند. سؤال‌های شبیه سؤال ۵ دارای دشواری بسیار بالا و ضریب تشخیص متوسط است. این‌گونه سؤال‌ها نیز در ساخت خزانه سؤال برای سنجش انطباقی جایی ندارند. درمجموع، میزان آگاهی کل بخش درک مطلب شفاهی و شنیداری (نمودار (۷)) در توانایی برابر با ۱ به اوج خود می‌رسد که نشان‌دهنده دشواری نسبی این بخش آزمون برای آزمودنی‌ها است. هیستوگرام توزیع توانایی (نمودار (۱)) نیز منحنی آگاهی کل این بخش را تأیید می‌کند.



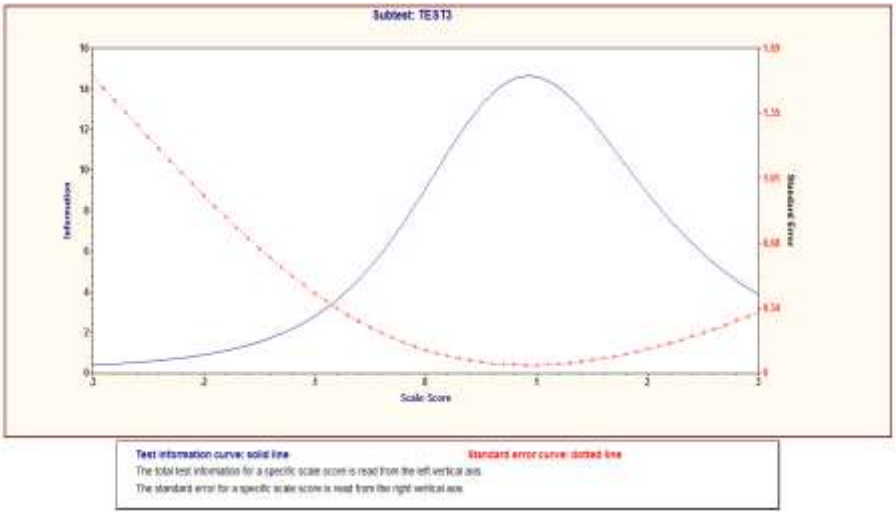
نمودار (٤) ماتریس ICC مربوط به سؤال‌ها بخش درک مطلب شفاهی و شنیداری



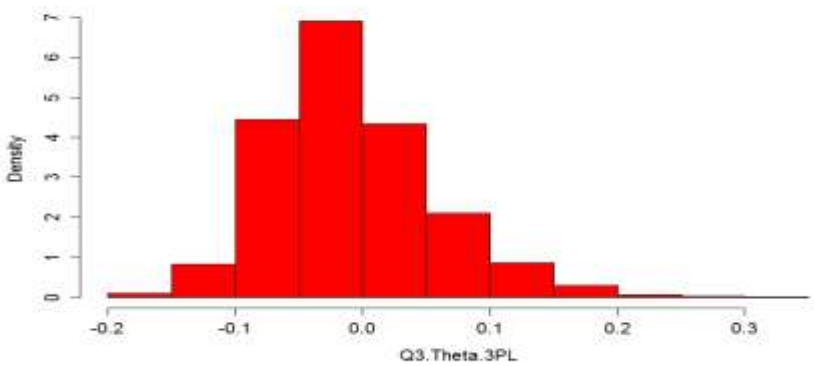
نمودار (۵) منحنی آگاهی مربوط به بخش ساختار و نوشتار زبانی



نمودار (٦) منحنی آگاهی مربوط به بخش خواندن و درک مطلب



نمودار (۷) منحنی آگاهی مربوط به بخش درک مطلب شفاهی و شنیداری



نمودار (۸) توزیع Q3 مربوط به بخش خواندن و درک مطلب

جدول (٤) تحلیل عاملی غیر خطی برای تعیین بعدیت زیربخش‌های آزمون تولیمو

ابعاد / شاخص برازش	تک بعدی			چندبعدی		
	ساختار و نوشتار زبانی	خواندن و درک مطلب	درک مطلب شنیداری	ساختار و نوشتار زبانی	خواندن و درک مطلب	درک مطلب شنیداری
SSR	۰/۰۰۱	۰/۰۰٤	۰/۰۰۲	۰/۰۰۷	۰/۰۰۹	۰/۰۰۸
RMSR	۰/۰۰۹	۰/۰۰۷	۰/۰۰۸	۰/۰۱	۰/۰۲	۰/۰۱
Tanaka	۰/۹۸۵	۰/۹٦۳	۰/۹۷۵	۰/۹۳۲	۰/۹٤۱	۰/۹۳۱

جدول (۵) نتایج مربوط به توزیع Q3 مربوط به بخش خواندن و درک مطلب

حداقل مقدار Q3	چارک اول Q3	میانه Q3	میانگین Q3	چارک سوم Q3	حداکثر مقدار Q3
-۰/۱۸۱	-۰/۰۵۳	-۰/۰۱۹۰	-۰/۰۰۸۷۵	۰/۰۲٦	۰/۲٤۱

جدول (٤)، نتایج مربوط به بعدیت هر سه بخش آزمون را نشان می‌دهد و نتایج گویای آن است که مدل تک‌بعدی بهترین برازش را با داده‌ها دارد. جدول (۵)، نتایج مربوط به استقلال موضعی در بخش خواندن و درک مطلب را نشان می‌دهد، نتایج گویای وجود استقلال موضعی در داده‌ها است.

۲- طراحی آزمون تولیمو به شیوه سنجش انطباقی کامپیوتری (مطالعه شبیه‌سازی‌شده، عملیاتی و نتایج مقایسه بین اجراهای مداد-کاغذی و سنجش انطباقی)

۲-۱- طراحی آزمون CAT عملیاتی تولیمو و ارزیابی بهترین روش برآورد پارامتر توانایی و ملاک توقف آزمون

در این بخش از یافته‌ها، آزمون CAT عملیاتی به‌شیوه زیر طراحی شد:

نخست، ویژگی‌های آماری خزانه سؤال عملیاتی CAT که شامل ۱٤۰۰ سؤال بود، (براساس نتایج مربوط به تحلیل سؤال‌های آزمون مداد-کاغذی) در سه محتوای اصلی و کلی (۱- ساختار و نوشتار زبانی، ۲- خواندن و درک مطلب، ۳- درک مطلب شفاهی و شنیداری) وارد MYSQL شدند. محتوای هر سؤال، بر اساس وزن

مشخصی که متخصصان موضوعی تعیین کرده بودند (٤ متخصص)، انتخاب شد (دو مؤلفه خزانه سؤال و تعادل محتوایی در این قسمت در نظر گرفته شد). سپس، بر اساس نوع روش برآورد توانایی آزمودنی‌ها و نوع ملاک توقف آزمون، ۸ طرح مختلف CAT عملیاتی طراحی شد. همچنین، برای حفظ امنیت آزمون، روش سیمپسون-هتر برای کنترل مواجهه سؤال در الگوریتم طراحی CAT عملیاتی درنظر گرفته شد. هر یک از طرح‌های CAT عملیاتی روی ۱۱۹۰۳ نفر اجرا شد. همچنین، برآورد پارامتر توانایی آزمون‌های مداد-کاغذی نیز برای همین تعداد آزمودنی در اجراهای واقعی در دسترس بود. در جدول (٦) نتایج مربوط به ارزیابی میزان دقت برآورد توانایی آزمودنی‌ها در ۸ طرح مختلف CAT، ارائه شده است. این ارزیابی بر اساس ٤ ملاک قرار دارد: ۱- ضریب راستی‌آزمایی یا همبستگی میان برآوردهای پارامتر توانایی در سنجش انطباقی و اجرای مداد-کاغذی؛ ۲- اریب یا محاسبه میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی؛ ۳- RMSE یا محاسبه ریشه دوم مجذور میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی. ٤- متوسط قدر مطلق تفاوت (ADD) یا محاسبه قدر مطلق میانگین تفاوت بین پارامتر توانایی برآورد شده و واقعی. همان‌طور که نتایج نشان می‌دهد روش برآورد توانایی MI و ملاک توقف طول ثابت آزمون (۱٤۰ سؤال)، بهترین شاخص‌های ارزیابی را نسبت به ۷ طرح دیگر نشان داد. بنابراین، با توجه به نتایج به‌دست آمده در این بخش از یافته‌ها، CAT عملیاتی با روش برآورد توانایی MI، ملاک توقف طول ثابت آزمون یعنی ۱٤۰ سؤال، روش کنترل مواجهه سیمپسون-هتر و یک خزانه سؤال موجود عملیاتی طراحی شد و مبنای مقایسه با مطالعات شبیه‌سازی‌شده CAT در دو زیربخش دیگر شد.

جدول (٦) مقادیر همبستگی، اریب، RMSE و ADD برآوردهای توانایی در آزمون تولیمو[1]

روش برآورد توانایی	ملاک توقف آزمون	مقدار ملاک	همبستگی	اریب	RMSE	ADD	میانگین سؤال‌های اجراشده
EAP	خطای استاندارد ثابت	۰/۳۸۵	۰/۹٤۳٤	-۰/۰۲۱٤	۰/۲۲۵۰	۰/۱۷۸٦	۱۱۹/۵۰
		۰/۳۱۵	۰/۹۵٤۳	-۰/۰۲۰۷	۰/۱۹۱۸	۰/۱۷٤۰	۱۲٤/٤۵
		۰/۲۲	۰/۹۵۷۲	-۰/۰۲۰۳	۰/۱۹۰۰	۰/۱۷۱۲	۱۳۵/۲۰
	طول ثابت آزمون	۱٤۰	۰/۹۷۵۵	-۰/۰۱۹٤	۰/۱۸٤۸	۰/۱٦۹۸	-
ML	خطای استاندارد ثابت	۰/۳۸۵	۰/۹۷۲٦	-۰/۰۱۳۲	۰/۱۷۹۷	۰/۱٤۸٦	۱۲۰/۱۲
		۰/۳۱۵	۰/۹۷۸۳	-۰/۰۱۲٤	۰/۱۷۹۳	۰/۱٤۵۳	۱۲٦/۵۵
		۰/۲۲	۰/۹۸۱٦	-۰/۰۱۲۱	۰/۱۷٦۳	۰/۱۳۵٤	۱۳۷/۱٤
	طول ثابت آزمون	۱٤۰	**۰/۹۸۹۹**	-۰/۰۱۱٤	**۰/۱٦٤٤**	**۰/۱۳۲٦**	-

۲-۲- طراحی CAT شبیه‌سازی شده آزمون تولیمو به همراه خزانه‌های سؤال بهینه، با در نظر گرفتن تعادل محتوایی: اجرا برای ۱۲۰۰۰ آزمودنی شبیه‌سازی‌شده

در این مرحله، برای شبیه‌سازی CAT بهینه، خزانه‌های سؤال بهینه در سه روش (R، MRP و MTI) با در نظر گرفتن عامل تعادل محتوایی و وزن‌های محتوایی از پیش تعیین شده برای آزمون تولیمو، طراحی شدند. به‌طوری که، محتوای آزمون CAT، بر اساس محتواهای دوره‌های جدید آزمون تولیمو کدگذاری شد و پس از توافق کامل میان ٤ متخصص موضوعی زبان انگلیسی، محتواها به کدهای معینی تبدیل شدند. محتواها به سه مجموعه اصلی (۱- ساختار و نوشتار زبانی، ۲- خواندن و درک مطلب، ۳- درک مطلب شفاهی و شنیداری) و به دنبال آن هر یک از مجموعه‌های اصلی به زیرمجموعه‌های معین (به ترتیب، ۲۰، ۲۵ و ۲۲) تقسیم‌بندی شدند. سپس با روش برنامه‌نویسی ریاضی، کدهای مربوط به هر یک از محتواها، وارد برنامه CAT شبیه‌سازی شده شدند. در این پژوهش از روش برنامه‌نویسی خطی (ریاضی) (WDM) برای تعیین محتواها و ایجاد تعادل محتوایی در خزانه‌های سؤال

[1]. همه مقادیر به‌دست آمده در جدول ازطریق میانگین ۲۵ تکرار محاسبه شده‌اند.

استفاده شد. با این روش، ابتدا یک خزانه سؤال بهینه شبیه‌سازی شده برای آزمون‌های سنجش انطباقی برای ۱۲۰۰۰ نفر سرهم شد تا خزانه سؤال کلی برای یک دوره یک ساله ساخته شود. در این روش، ابتدا پیش‌بینی جستجوی راه حل برای تست کامل صورت گرفت و هم‌زمان، هم قابل حل بودن و هم بهینه بودن تست در نظر گرفته شد. این روش جزء روش‌های شهودی حلّ مسائل سرهم کردن تست است. با کاربرد روش WDM به صراحت، ویژگی‌های آماری و غیر آماری سؤال‌ها با تعادل مطلوبی بین ویژگی‌های اندازه‌گیری و ساختاری در نظر گرفته شده است. این ویژگی‌ها به‌وسیله وزن‌هایی که توسط متخصصان موضوعی آزمون تولیمو انتخاب شد، در مدل وارد شدند. به عبارت دیگر، ویژگی‌های محتوایی به‌عنوان هدف‌ها فرمول‌بندی شدند. انحراف از هدف‌های محتوایی وزن داده شد و در تابع هدف به همراه فاصله آگاهی سؤال از مقدار هدف قرار داده شد. انتخاب سؤال‌ها در CAT، بر اساس رویکرد WDM به شکلی تنظیم شد تا سؤال‌هایی انتخاب شوند که به‌طور متوالی کوچک‌ترین مجموع انحرافات وزن‌دار را داشته باشند. برای انتخاب یک سؤال از سه گام پیروی شد: ۱- اگر سؤالی که قبلاً در تست نبوده به تست اضافه شود، انحراف برای هر یک از قیود محاسبه شود؛ ۲- انحرافات وزن‌دار در میان همه قیود جمع شود؛ ۳- در پایان، سؤالی با کوچک‌ترین مجموع وزن‌دار انحرافات انتخاب شود. در این روش مدل‌یابی، سؤال‌ها به صورت $i=1,...,N$ نشان داده می‌شود، x_i متغیر تصمیم‌گیری را نشان می‌دهد. اگر سؤال در تست وارد شود، $x_i=1$ و اگر سؤال از تست خارج شود $x_i=0$. در این مدل $j=1,...,J$ صفات تست همراه قیود غیر روان‌سنجی را نشان می‌دهد. حدود پایین و بالای تعداد سؤال‌هایی که در آزمون دارای چنین ویژگی‌هایی هستند را به ترتیب با L_j و U_j نشان می‌دهد، البته ممکن است گاهی با یکدیگر برابر باشد. همچنین، اگر سؤال i دارای ویژگی j باشد، $a_{ij}=1$. و اگر سؤال i دارای ویژگی j نباشد، $a_{ij}=0$. N تعداد سؤال‌ها در خزانه را نشان می‌دهد، W_j وزن اختصاص داده شده به هر قید را نشان می‌دهد، d_{L_j} و d_{U_j} به ترتیب کسری حد پایین و مازاد حد بالا را نشان می‌دهند. e_{L_j} و e_{U_j} ، به ترتیب، اضافی حد پایین و کسری حد بالا را نشان می‌دهد. d_θ انحراف از آگاهی هدف را برای یک آزمودنی نشان می‌دهد. دو جدول (۷) و (۸) به صورت خلاصه اطلاعات مربوط به توابع هدف و قیود مربوط به آن را نشان داده است. قیود تست به‌عنوان

ویژگی‌های غیر آماری یا غیر روان‌سنجی، به همراه ویژگی‌های آماری وارد شبیه‌سازی‌های روش اکتشافی مرحله قبل شد. سپس، انحرافات از این قیدها برای هر یک از ۱۲۰۰۰ تعداد CAT که از کل خزانه بهینه سرهم شده است، محاسبه شد. به‌طور کلی، در این مرحله برای طراحی و ساخت خزانه سؤال بهینه، تلفیقی از دو رویکرد برنامه‌نویسی ریاضی و رویکرد اکتشافی به چشم می‌خورد (مقدسین، ۱۳۹۵).

جدول (۷) اطلاعات مربوط به قیود و وزن‌های آزمون‌ها در مورد به حداقل رساندن انحرافات از قیود

$\text{minimize} \sum_{j=1}^{J} W_j d_{L_j} + \sum_{j=1}^{J} W_j d_{U_j} + W_\theta d_\theta \rightarrow (objective)$	**تابع هدف: به حداقل رساندن میزان انحرافات وزن دار**
در ارتباط با قید زیر	
$\sum_{i=1}^{60} a_{ij} x_i + d_{L_j} - e_{L_j} = L_j \Rightarrow j = 1,...,J$	
$\sum_{i=1}^{60} a_{ij} x_i - d_{U_j} + e_{U_j} = U_j \Rightarrow j = 1,...,J$	
$\sum_{i=1}^{60} I(\theta) x_i + d_\theta - e_\theta = \infty$	

جدول (۸) اطلاعات مربوط به قیود و وزن‌های آزمون‌های CAT تولیمو در مورد بیشینه کردن آگاهی تست

تابع هدف: به حداکثر رساندن تابع هدف در ارتباط با قیود زیر			$\max imize \sum_{i=1}^{I} I_i(\hat{\theta}^{(g-1)})x_i \rightarrow (objective)$	
قید	کد قید	وزن	حداقل	حداکثر
طول تست $\sum_{i=1}^{I} x_i = 140 \rightarrow test-length$	Test lenght	N1	۲۱/٤	۲۱/٤
		N2	۲۹/۸	۲۹/۸
		N3	۱۵/٤	۱۵/٤
تعداد زیرمجموعه‌های تست $\sum_{i=1}^{25} z_1 = 20, \sum_{i=1}^{20} z_2 = 25, \sum_{i=1}^{15} z_3 = 22$	Number of test sets	$z_1 = structure$	۲۰	۲۰
		$z_2 = Reading$	۲۵	۲۵
		$z_3 = lisiting$	۲۲	۲۲
تعداد سؤال‌ها در زیرمجموعه‌های تست $\sum_{i \in V_1} x_1 \le n^{(u)}{}_1 z_1, \sum_{i \in V_1} x_1 \ge n_1^{(l)} z_1, \sum_{i \in V_1} x_2 \le n_2^{(u)} z_2, \sum_{i \in V_1} x_2 \ge n_2^{l} z_2, \sum_{i \in V_1} x_3 \le n_3^{(u)} z_3, \sum_{i \in V_1} x_3 \ge n_3^{(l)} z_3$	Number of item in test sets	For example : $z_{1-1} = 1$ $z_{1-2} = 2$ $\vdots$ $z_{3-11} = 1$	۱	۳
تعداد سؤال‌ها در هر سطح شناختی $\sum_{i \in C_h} x_i \ge n^{(l)}{}_h$ $\sum_{i \in C_h} x_i \le n^{(u)}{}_h$	Number of item per cognitive level	سه حوزه‌ی شناختی: به کار بستن: h_1 تجزیه و تحلیل: h_2 ترکیب: h_3	۱	۷

الف) طراحی و ساخت CAT شبیه‌سازی شده به همراه خزانه‌های سؤال بهینه، با در نظر گرفتن تعادل محتوایی و بدون کنترل مواجهه بیش از حد سؤال با b-bin=0.2

در این مرحله، ویژگی‌های آزمون‌های CAT شبیه‌سازی شده به صورت زیر است: آزمون‌های CAT با خزانه‌های سؤال بهینه‌ای که از طریق سه روش R، MRP، MTI با پهنای b-bin = 0.2، میزان a-bin: $\Delta a2=2\Delta IMaximum$ = 0.4 و بدون هیچ روش کنترل مواجهه‌ای، در هر سه محتوا در نظر گرفته شده‌اند، طراحی شدند. همچنین، به‌منظور مقایسه‌پذیری نتایج CAT عملیاتی و شبیه‌سازی شده، روش برآورد توانایی در کل CATهای شبیه‌سازی شده این مرحله، MI و ملاک توقف آزمون، طول ثابت ۱٤۰ سؤال درنظر گرفته شد.

به دلیل اینکه، توزیع خزانه‌های سؤال در هر یک از محتواها از یکدیگر تفکیک شود، پس از ایجاد خزانه‌های سؤال که با درنظر گرفتن تعادل محتوایی ساخته شدند، توزیع‌های هر کدام از سه محتوا در جداول جداگانه گزارش شده است. جدول (۹)، اندازه‌ها و خلاصه آماره‌های مربوط به پارامترهای سؤال در خزانه‌های سؤال بهینه و عملیاتی را در سه محتوا ارائه کرده است. نتایج نشان داد که در این مرحله، خزانه‌های سؤال بهینه دارای حداقل تعداد سؤال هستند. البته یکی از دلایل آن می‌تواند این قضیه باشد که در ساخت آنها هیچ نوع کنترل مواجهه‌ای صورت نگرفته است. همه خزانه‌های سؤال بهینه، نسبت به خزانه سؤال عملیاتی در پایین توزیع دارای سؤال‌های با دامنه بزرگ‌تری در سطوح دشواری هستند. ولی در بالای توزیع دامنه سؤال‌ها محدودتر است؛ به این دلیل که وقتی قیود محتوایی در تعامل با ویژگی‌های روان‌سنجی سؤال‌ها قرار می‌گیرند، خزانه‌های سؤال بهینه شبیه‌سازی شده دارای ویژگی‌های روان‌سنجی دقیق‌تری می‌شوند، به‌طوری که، دامنه دشواری سؤال‌ها دقیق‌تر می‌شود. خزانه‌های سؤال بهینه دارای میانگین دشواری بالاتری نسبت به خزانه‌های عملیاتی هستند ولی پراکندگی کمتری دارند.

در این مرحله، خزانه بهینه MTI در هر سه محتوا دارای حداقل تعداد سؤال است، ولی تفاوت زیادی با خزانه‌های MRP ندارد. بر اساس ادبیات پژوهشی خزانه‌های MTI، نسبت به خزانه‌های دیگر دارای حداقل ضریب تشخیص بهینه هستند، امّا در این شبیه‌سازی، خزانه‌های MTI دارای حداقل میانگین پارامتر a نیستند و دلیل این امر، تعامل ویژگی‌های محتوایی و پارامترهای روان‌سنجی است. با

این وجود، خزانه‌های بهینه MRP دارای بیشترین مقدار پارامتر a هستند. ولی، خزانه‌های بهینه R و MTI دارای میانگین پارامتر a مشابهی هستند. خزانه‌های MTI به دلیل ماهیت ایجاد سؤال‌هاشان بیشترین مقدار پارامتر a آن نسبت به خزانه‌های دیگر حداقل است و کمترین مقدار پارامتر a آن نیز نسبت به خزانه‌های دیگر بیشتر است. به عبارت دیگر، دارای حداقل میزان پراکندگی در پارامتر a هستند. توزیع خزانه‌های سؤال R نسبت به دو خزانه دیگر ،دارای یک توزیع یکنواخت‌تری در سراسر ماتریس پارامترها است و این نتیجه به دلیل ماهیت روشی است که پارامترهای سؤال را ایجاد کرده است. در این روش، پارامترهای a و b در سراسر منحنی توانایی پراکنده شده‌اند. توزیع پارامتر دشواری و تشخیص سؤال‌ها در این روش بسیار مشابه خزانه سؤال عملیاتی است و در تمام محتواها دارای مقادیر پارامتر متنوع‌تری است. امّا سؤال‌های دشوار در خزانه‌های بهینه MRP پارامتر ضریب تشخیص بالاتری دارند و برعکس، سؤال‌های آسان دارای پارامترهای ضریب تشخیص متوسط یا پایین‌تری هستند. این نتایج باعث می‌شود که تعداد آزمون‌هایی که در خزانه‌های R از قیود محتوایی تخطی می‌کنند، در سرتاسر پارامتر توانایی یکنواخت باشد. خزانه‌های MRP در پارامترهای توانایی بالاتر از متوسط، دارای تخطی از قیود محتوایی کمتری هستند.

جدول (۱۰)، عملکرد CATهایی که بر اساس خزانه‌های سؤال بهینه طراحی شدند را با CAT عملیاتی مقایسه کرده است. با این وجود، برآورد توانایی در هر سه خزانه بهینه و عملیاتی، دارای سطح معینی از اریب منفی است، یعنی اجرای آزمون به شکل CAT باعث کم برآورد شدن پارامتر توانایی شده است ولی مقدار این اریب‌ها بسیار ناچیز است. میانگین مجذور خطا (MSE) برآورد پارامتر توانایی در خزانه‌های سؤال بهینه کوچک‌تر از خزانه سؤال عملیاتی است. در میان خزانه‌های سؤال بهینه MTI عملکرد بهتری در این شاخص برای برآورد توانایی نشان می‌دهد، زیرا با برنامه‌ریزی دقیق‌تری با توجه به تعامل میزان حداقل آگاهی با ویژگی‌های محتوایی ایجاد شده است؛ بنابراین، برآورد توانایی را با دقت بیشتری برآورد می‌کند. در مجموع، خزانه‌های سؤال بهینه دارای نرخ همپوشی پایین‌تری هستند، با وجود اینکه سؤال‌های کمتری دارند.

جدول (۹) اندازه خزانه سؤال و آماره‌های پارامتر سؤال، بدون S-H (b-bin=0.2)، با تعادل محتوا

خزانه سؤال	اندازه خزانه	a				b				c			
		میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل
Content 1(structure)													
OP	٤۰۰	۰/٦۳	۰/۲۵	۲/۱۵	-۰/۰۵	-۰/۰۲	۱/۱۳	۳/۹۸۱	-۲/۵۹٦	۰/۰۷	۰/۰۲۲	۰/۲۱۷۹	۰/۰۰۰۵
ROP_1	۲۸۲	۱/۷۹	۰/۲۷۷	۲/۷۸	۰/٤٦	۰/۰۱٦	۱/۱۵	۳/٦۵	-۳/۸۲	۰/۱٤٦	۰/۰۷	۰/٤۱۰۲	۰/۰۰۱
ROP_2	۲۳۰	۲/۰۱۲	۰/۲۸۱	۳/۱۲	۰/۹۵	۰/۰۲۰۱	۱/۰٦۳	۳/۸۲۵	-۳/۱۱۸	۰/۱٤۸	۰/۰۷۲	۰/٤۳	۰/۰۲٤
ROP_3	۲۲٦	۱/۵٦۲	۰/۲۵۳	۲/۳۲٤	۰/۹۸۷	۰/۰۱٤٦	۱/۰۰۸	۳/۸٤۱	-۳/۵۹۲	۰/۱٤۲	۰/۰٦٤	۰/٤٦	۰/۰۰۱
Content 2 (reading)													
OP	۵۰۰	۱/۱۹	۰/٦٤	۳/۹۳	۰/۲٤۵	۰/۹۷	۰/۷۷	٤/۸٤	-۱/۷۵۸	۰/۲۱	۰/۰۹۹	۰/٤۷٦	۰/۰۹۱
ROP_1	۳۳۱	۱/۹٦۳	۰/۲٦۳	۲/٦۵٤	۰/۵۲٦	۰/٦۸۷۵	۰/۹۷٦	۳/۵۵	-۳/۸۵٤	۰/۱۷۱	۰/۰۷٤	۰/٤۲	۰/۰۱
ROP_2	۲۸۰	۲/۸۸۲	۰/۲۷۸	۲/۹۳۵	۰/۹۲۱	۰/۲۱۲۸	۱/۰٤۷	۳/۷٤	-۳/٤۹	۰/۱۸۳	۰/۰٦۸	۰/٤٦	۰/۰۰۱۵
ROP_3	۲۷۷	۱/۸۱٤	۰/۲۷۲	۲/۵۱	۱/۱۵	۰/۰۱۳۵	۰/۹۹۳	۳/۱۰٤	-۳/۱۰٤	۰/۱۷۸	۰/۰۸۱	۰/٤۸	۰/۰۰۲

Content 3 (listening)													
OP	۵۰۰	۰/۵٤	۰/٤۱	۳/۸۸۹	۰/۱٦٦	۱/۹۷	۱/۸۲	۵/٦۸۹	-۲/٤۰۹	۰/۰۸	۰/۰۱۲	۰/۱۹۲	۰/۰۰٤
ROP_1	۳۲۵	۱/۷۸۳	۰/۲٤۱	۲/۷۸۹	۰/۲۳۵	۰/۰۲۳۸	۰/۹۹۸	۳/۵۳۲	-۳/٤۵٦	۰/۱۰۱	۰/۰۲۲	۰/٤۹۷	۰/۰۰۳
ROP_2	۲۷۹	۲/۱٤۸	۰/۳۰۲	۳/۰٤٦	۰/۹۵	۰/۰۱۸٤	۰/۹۹٤	۳/۵۸۹	-۳/۱٤	۰/۱۲۹	۰/۰۳۵	۰/٤٦۵	۰/۰۰۲
ROP_3	۲۷۲	۱/۷۹٤	۰/۳۱۲	۲/٤۸۹	۰/۹۰۲	-۰/۰۱۰۹	۱/۰۰۷	۳/۵۹۸	-۳/۰۸۹	۰/۰۹٤	۰/۰٤٤۸	۰/٤۵	۰/۰۰۱

ارزیابی عملکرد CAT های شبیه‌سازی شده براساس خزانه سؤال بهینه، با در نظر گرفتن تعادل محتوایی و بدون کنترل مواجهه سیمپسون-هتر (S-H): (b-bin=0.2)

جدول (۱۰) خلاصه ملاک‌های ارزیابی عملکرد CATهای شبیه‌سازی شده بدون S-H ،(b-bin=0.2) و تعادل محتوا با CAT عملیاتی

آماره‌ها	OP	R	MRP	MTI
Bias	-۰/۰۱۱٤	-۰/۰۰۲۰	-۰/۰۰۰٦۱	-۰/۰۰۱۸۹
MSE	۰/۰۱۷٤۵	۰/۰۱۲٦۷	۰/۰۱۱۳۸	۰/۰۰۷٦
کجی نرخ مواجهه	۹۸/۸۵٦	۵۱/٤۰۲	٤۸/٤۳٦۵	٤۳/٤۹٤
نرخ همپوشی سؤال	۰/۵۱۷۳	۰/۳۷٦۵	۰/۳۸٤٤	۰/۳۹٤۱
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۸/٪٦٦۹	۹/٪٦٤٤	۱۱/٪۵۲٤	۱۰/٪۰۰۳
درصد سؤال‌های با نرخ مواجهه کوچک‌تر از 0.02	۵۱/٪۱۲٦	۲۵/٪۱۳۱	۲۱/٪۱۵	۱۹/٪۲۸۹
درصد تست‌هایی که از قیود تست تخطی دارند	۵٤/٪۸	۳/٪۰۰۲	۲/٪۰۰۱۳	۲/٪۰۰۱۵
اندازه خزانه سؤال	۱٤۰۰	۹۳۸	۷۸۹	۷۷۵

جدول‌های (۱۱و۱۲) نتایج مربوط به میزان تخطی از قیود را در تست‌های سرهم‌بندی شده برای ۱۲۰۰۰ آزمودنی در CATهای شبیه‌سازی شده را برحسب پارامترهای توانایی نشان می‌دهد. در CAT عملیاتی، ۱۱۹۰۳ تست سرهم شدند و در CATهای شبیه‌سازی شده، ۱۲۰۰۰ تست سرهم شدند. برای قابل مقایسه کردن میزان تخطی‌ها، درصد فراوانی نسبی آنها گرفته شد. نتایج نشان داد که در CAT عملیاتی در تمام سطوح میزان تخطی بیشتر از CATهای شبیه‌سازی شده است. در خزانه‌هایی که به‌روش R طراحی شدند، میزان تخطی‌ها در دو دامنه توانایی بیشتر است و تقریباً الگویی مشابه با خزانه عملیاتی دارد. در خزانه‌هایی که به‌روش MRP و MTI طراحی شدند، در سطوح پایین توانایی میزان تخطی از قیود بیشتر است. در خزانه‌های MRP در سطوح بالای توانایی به دلیل وجود سؤال‌های بیشتر، تخطی از قیود در تست‌هایی که سرهم می‌شود، به حداقل خود رسیده است.

جدول (۱۱) تعداد تست‌ها بر اساس تخطی از قیود محتوایی در CATهای شبیه‌سازی شده، بدون S-H و CAT عملیاتی

سطوح توانایی / خزانه سؤال	total CAT	OP	R (ROP_1)	-٤	-۳/۵	-۳	-۲/۵	-۲	-۱/۵	-۱	-۰/۵	۰	۰/۵	۱	۱/۵	۲	۲/۵	۳	۳/۵	٤	کل
	۱۱۹۰۳	۳۵۷	٤۸	۳۵۷	٤۷٦	۵۹۵	۳۵۷	۳۵۷	۳۵۷	۹۵	۱۱	۱۱	۲۲	۱۱۹	۲۳۸	٤۷	۳۵۷	۸۳۳	۸۳۳	۹۵۲	٦۰۱۷
	۱۲۰۰۰	٤۸	۳٦	۳٦	۳۰	۹	۱۲	۹	۰	۰	۰	۰	۰	۲٤	۹	۱۲	۱۲	۲٤	۳٦	٦۰	۳۱۲

MRP (ROP_2)	MTI (ROP_3)
۱۲۰۰۰	۱۲۰۰۰
٤٨	٤٨
٣٦	٤٨
٢٤	٣٦
٣٦	٢٤
٣٦	١٢
١٢	٠
٠	٠
٠	٠
٠	٠
٠	٠
١٢	١٢
٦	٦
٨	٦
١٢	١٢
٠	٠
٠	٥
١	٤
٢	١٢
٤	١٢
٢٢٥	٢٢٥

جدول (۱۲) درصد تست‌ها بر اساس تخطی از قیود محتوایی در CAT های شبیه‌سازی شده، بدون S-H و CAT عملیاتی

خزانه سؤال / سطوح توانایی	total CAT	OP	R (ROP_1)	MRP (ROP_2)	کل
-٤	٠/٠٣	٠/٠٤	٠/٠٠٤	٠/٠٠٤	٤
-٣/٥	٠/٠٤	٠/٠٤	٠/٠٠٣	٠/٠٠٣	٣/٥
-٣	٠/٠٥	٠/٠٤	٠/٠٠٢٥	٠/٠٠٢	٣
-٢/٥	٠/٠٤	٠/٠٤	٠/٠٠١	٠/٠٠٣	٢/٥
-٢	٠/٠٤	٠/٠٤	٠/٠٠٠٨	٠/٠٠٣	٢
-١/٥	٠/٠٤	٠/٠٤	٠	٠/٠٠١	١/٥
-١	٠/٠٠٨	٠/٠٠١	٠	٠	١
-٠/٥	٠/٠٠١	٠/٠٠١	٠	٠	٠/٥
٠	٠/٠٠١	٠/٠٠١	٠	٠	٠
٠/٥	٠/٠٠٢	٠/٠١	٠	٠	٠/٥
١	٠/٠١	٠/٠٢	/٠٠٢	٠/٠٠١	١
١/٥	٠/٠٢	٠/٠٠٤	٠/٠٠٠٨	٠/٠٠٠٧	١/٥
٢	٠/٠٠٤	٠/٠٣	٠/٠٠١	٠/٠٠٠٦	٢
٢/٥	٠/٠٣	٠/٠٧	٠/٠٠١	٠	٢/٥
٣	٠/٠٧	٠/٠٧	٠/٠٠٢	٠	٣
٣/٥	٠/٠٧	٠/٠٧	٠/٠٠٣	٠/٠٠٠٢	٣/٥
٤	٠/٠٨	٠/٠٨	٠/٠٠٥	٠/٠٠٠٤	٤
کل	٠/٥٤٨	٠/٠٣	٠/٠٣	٠/٠٢	

MTI (ROP_3)	۱۲۰۰۰	۰/۰۰۴	۰/۰۰۴	۰/۰۰۳	۰/۰۰۲	۰/۰۰۱	۰	۰	۰	۰	۰	۰/۰۰۱	۰/۰۰۲	۰/۰۰۳	۰/۰۰۴	۰/۰۰۴	۰/۰۰۵	۰/۰۰۴	۰/۰۰۱	۰/۰۰۱	۰/۰۰۲

ب) ساخت و طراحی CATهای شبیه‌سازی‌شده آزمون تولیمو به همراه خزانه‌های سؤال بهینه با در نظر گرفتن تعادل محتوایی و کنترل مواجهه بیش از حد سؤال با b-bin=0.2

در این مرحله، ویژگی‌های آزمون‌های CAT شبیه‌سازی شده به صورت زیر است: آزمون‌های CAT با خزانه‌های سؤال بهینه، با سه روش R، MRP، MTI با پهنای b-bin = 0.2، میزان a-bin: $\Delta a2=2\Delta IMaximum$ = 0.4 و با روش کنترل مواجهه S-H، طراحی شدند. توزیع‌های هر کدام از سه محتوا در جدول‌های جداگانه گزارش شده است. جدول (۱۳) اندازه‌ها و خلاصه آماره‌های مربوط به پارامترهای سؤال در خزانه‌های سؤال بهینه و عملیاتی را در سه محتوا ارائه کرده است. نتایج نشان داد که با وجود کنترل مواجهه سؤال در این مرحله، خزانه‌های سؤال بهینه باز هم شامل حداقل تعداد سؤال هستند. در این مرحله هم همانند خزانه‌هایی که بدون کنترل مواجهه ایجاد شدند، هر سه خزانه بهینه دارای سؤال‌های با دامنه وسیع‌تری در سطوح دشواری هستند و میانگین پارامتر دشواری بالاتری دارند. در این مرحله نیز خزانه‌های بهینه MTI در هر سه محتوا دارای حداقل تعداد سؤال است. ولی خزانه‌های بهینه MRP در مقایسه با خزانه‌های MRP که کنترل مواجهه سؤال نداشتند، به تعداد سؤال‌های بسیار بیشتری نیاز دارند. همچنین، نسبت به دو خزانه دیگر با شرایط مساوی نیز به سؤال‌های بیشتری نیاز دارند. در این مرحله نیز خزانه‌های MTI دارای حداقل میانگین پارامتر a و خزانه‌های بهینه MRP دارای بیشترین مقدار پارامتر a هستند. خزانه‌های MTI به دلیل ماهیت ایجاد سؤال‌هاشان دارای حداقل میزان پراکندگی در پارامتر a هستند. شکل توزیع پارامترهای سؤال در هر سه خزانه‌های سؤال بهینه مشابه خزانه‌های سؤال بهینه در مرحله قبل است، با این تفاوت که میانگین ضریب تشخیص در هر سه خزانه نسبت به خزانه‌های مرحله قبل بالاتر است. بررسی نتایج عملکرد این CATها در جدول (۱٤) آورده شده است. برآورد توانایی در هر سه خزانه بهینه و عملیاتی، دارای سطح معینی از اریب منفی

است، ولی مقدار این اریب‌ها بسیار ناچیز است. میانگین مجذور خطا (MSE) برای برآورد پارامتر توانایی در خزانه‌های سؤال بهینه کوچک‌تر از خزانه سؤال عملیاتی است. در میان خزانه‌های سؤال بهینه، MRP عملکرد بهتری در این شاخص برای برآورد توانایی نشان می‌دهد، زیرا سؤال‌های با ضریب تشخیص بالاتری دارند؛ بنابراین، برآورد توانایی را با دقت بیشتری برآورد می‌کند. همچنین، خزانه‌های سؤال بهینه دارای نرخ همپوشی تست پایین‌تری هستند، با وجود اینکه سؤال‌های کمتری دارند. نتایج جدول (۱۵) نشان می‌دهد که در بخش اول، دوم و سوم خزانه‌های بهینه‌ای که با روش R دارای بیشترین نرخ مواجهه هستند و با اینکه بیشترین تعداد سؤال را دارد، دارای نرخ بالاتری از مواجهه است. در این مرحله نیز خزانه‌های MTI دارای کمترین نرخ مواجهه هستند. البته نتایج خزانه‌های MTI و MRP در نرخ مواجهه سؤال، بسیار مشابه است. این نتایج نشان می‌دهد که با وجود اینکه، خزانه‌های MTI دارای تعداد سؤال کمتری هستند، نرخ‌های مواجهه کمتری دارند. همان‌طور که جدول (۱٤) نیز نشان داد، نرخ سؤال‌های کم‌مواجهه شده در خزانه‌های سؤال بهینه MTI نیز کمتر از بقیه است، همه این نتایج گویای به‌صرفه بودن این نوع خزانه‌ها است.

جدول (۱۳) اندازه خزانه سؤال و آماره‌های پارامتر سؤال، با S-H (b-bin=0.2)، با تعادل محتوا

خزانه سؤال	اندازه خزانه	a				b				c			
		میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل
Content 1(structure)													
OP	٤۰۰	۰/٦۳	۰/۲۵	۲/۱۵	-۰/۰۵	-۰/۰۲	۱/۱۳	۳/۹۸۱	-۲/۵۹٦	۰/۰۷	۰/۰۲۲	۰/۲۱۷۹	۰/۰۰۰۵
ROP_4	۳۱۳	۱/۹۰۱	۰/۳۰۱	۳/۰۲٤	۰/۹۰۲	-۰/۰۳٤٦	۰/۹۹۸	۳/۹۹۸	-۲/۹۸۹	۰/۰٤۸	۰/۰۷۲۳	۰/٤۳	۰/۰۰۱

ROP_5	۳۰۱	۲/۲۸۹	۰/۳۰۱۲	۳/۱۸۶	۰/۹۳۸	۰/۰۵۴۹	۱/۰۹۲	۳/۹۹٤	-۳/٤۹٥	۰/۰٤۲	۰/۰٦۸	۰/٤۲۱	۰/۰۰٤
ROP_6	۲۷۱	۱/٦۱۳	۰/۲٦۷	۲/٥۲۸	۱/۱۸۹	-۰/۰۲۱۸۰	۱/۱۱۸	۳/۹۹٦	-۳/٦٤۷	۰/۰٤٤	۰/۰۷۱	۰/٤٦۹	۰/۰۰۱
Content 2 (reading)													
OP	٥۰۰	**۱/۱۹**	۰/٦٤	**۳/۹۳**	**۰/۲٤٥**	۰/۹۷	۰/۷۷	٤/۸٤	-۱/۷٥۸	۰/۲۱	**۰/۰۹۹**	**۰/٤۷٦**	**۰/۰۹۱**
ROP_4	۳۱۹	۱/۹۸۷	۰/۲۷٤	۳/۱٥٤	۰/۹۰٥	-۰/۰۹٤۸	۱/۱۸۲	۳/٥۹۳	-۳/٥۸۷	۰/۱۷۹	۰/۰۹۱	۰/٤٦	۰/۰۰٤
ROP_5	۳۰٤	۲/٦۱۰	۰/۳۰٤	۳/۱۸۳	۱/۰۸٦	۰/۲۱٥	۱/۰۰۳	۳/۹۹۷	-۳/٤۸۲	۰/۱۸۲	۰/۰۸۷	۰/٤٥۸	۰/۰۰۳
ROP_6	۲۷٥	۱/٥٤۲	۰/۲۸٦	۲/٥۰۲	۱/۰۸۷	-۰/۰۳٦۸	۱/۱۲۲	۳/۷۸۲	-۳/٥۷۹	۰/۱۷٦	۰/۰۸۹	۰/٤٤٥	۰/۰۰۷
Content 3 (listening)													
OP	٥۰۰	**۰/٥٤**	**۰/٤۱**	**۳/۸۸۹**	**۰/۱٦٦**	**۱/۹۷**	**۱/۸۲**	**٥/٦۸۹**	**-۲/٤۰۹**	**۰/۰۸**	**۰/۰۱۲**	**۰/۱۹۲**	**۰/۰۰٤**
ROP_4	۳۲۸	۱/۷٦۹	۰/۲۹۸	۳/۰۰٦	۰/۹۱۸	۰/۰۰٤۲۸	۱/۰۰۱	۳/٥۸۲	-۳/٥٥۲	۰/۰۸۱	۰/۰۲۸	۰/٤۹۲	۰/۰۰۲
ROP_5	۳۱۰	۲/۱۲٥	۰/۲۹۹	۲/۸۸٦	۰/۹٤۸	-۰/۰۰٤٦۱	۱/۰۱٤	۳/۹۹۸	-۳/۱۸٤	۰/۰۸٤	۰/۰۳۷	۰/٤۸۳	۰/۰۰۱
ROP_6	۲۸۸	۱/۸۲۹	۰/۲۷٤	۲/٥۲۳	۱/۱۰٤	-۰/۰٤۹٦	۱/۱۰۲	۳/۹۹۳	-۳/٤۹٥	۰/۰۸۳	۰/۰۱۹	۰/٤٦	۰/۰۰۱٥

نتایج ارزیابی عملکرد خزانه‌های سؤال بهینه شبیه‌سازی شده با در نظر گرفتن تعادل محتوایی در اجرای CAT آزمون تولیمو و با کنترل مواجهه سیمپسون-هتر (b-bin=0.2)

جدول (۱٤) خلاصه آماره‌های عملکرد CAT با S-H (b-bin=0.2)، با تعادل محتوا

آماره‌ها	OP	R	MRP	MTI
Bias	-۰/۰۱۱٤	-۰/۰۰۱۰	-۰/۰۰۰۱۱	-۰/۰۰۰٦۲
MSE	۰/۰۱۷٤٥	۰/۰۱۲٦٥	۰/۰۰۷۳	۰/۰۱۱۲٤
کجی نرخ مواجهه	۹۸/۸٥٦	۳۰/۲٥۲	۲٦/۷٤۱	۲٦/۹۹٤
نرخ همپوشی سؤال	۰/٥۱۷۳	۰/۲۲٤۳	۰/۲۳٦۷	۰/۲٤۰۱
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۸/۰٦٦۹	٥/۰۱٥۳	٤/۰٦۳٥	۳/۰۸۷۲
درصد سؤال‌های با نرخ مواجهه کوچکتر از 0.02	٥۱/۰۱۲٦	۲۳/۰۰۸٦	۲٤/۰٤۲	۱۷/۰۳۲
درصد تست‌هایی که از قیود تست تخطی دارند	٥٤/۰۸	۳/۰٤	۲/۰٥	۲/۰٦
اندازه خزانه سؤال	۱٤۰۰	۹٦۰	۹۱٥	۸۳٤

نتایج مربوط به جدول‌های (۱٦و۱۷) نشان داد که در CAT عملیاتی در تمام سطوح توانایی درصد فراوانی نسبی تخطی بیشتر از CATهای شبیه‌سازی شده است. در خزانه بهینه R میزان تخطی‌ها در دو دامنه توانایی بیشتر است و تقریباً الگویی مشابه با خزانه عملیاتی دارد. در خزانه‌های بهینه MRP و MTI در سطوح پایین توانایی میزان تخطی از قیود بیشتر است. در خزانه MRP در سطوح بالای توانایی به دلیل وجود سؤال‌های بیشتر، تخطی از قیود در تست‌هایی که سرهم شدند، به حداقل خود رسیده است. در هر سه خزانه سؤال بهینه، میزان تخطی از قیود محتوایی تست‌ها بیشتر از زمانی است که کنترل مواجهه سؤال وارد نشده بود؛ این نتیجه دلیلی بر این امر است که وارد کردن کنترل مواجهه S-H بر انتخاب سؤال‌ها تأثیر می‌گذارد و این امکان وجود دارد که برنامه CAT، سؤالی را برای اجرا انتخاب کند که کاملاً با قیود محتوایی هماهنگ نباشد و در عمل میزان این تخطی‌ها را بیشتر کند. میزان اریب در هر چهار خزانه منفی و بسیار کوچک است. زمانی که مواجهه سؤال کنترل می‌شود،

این میزان اریب کمتر از زمانی است که کنترل نمی‌شود. زمانی که عامل S-H در شبیه‌سازی وارد نشده است، خزانه MTI دارای بهترین عملکرد از نظر دقت اندازه‌گیری توانایی است. در مرحله قبل، خزانه MRP و R دارای خطاهای اندازه‌گیری تقریباً مشابهی بودند. دلیل اینکه خزانه MTI در مرحله قبل دارای خطای اندازه‌گیری توانایی کمتری نسبت به MRP بود، این بود که در طراحی این خزانه در طول پیوستار توانایی میزان آگاهی‌های متفاوتی وارد شد که با کدهای محتوایی هماهنگ بود. تعامل میزان آگاهی با کدهای محتوایی باعث می‌شد که دقت اندازه‌گیری بالا رود. امّا زمانی‌که عامل S-H وارد شد، خزانه MTI دارای دقت اندازه‌گیری کمتری است نسبت به زمانی که S-H وارد نشده بود؛ زیرا سؤال‌هایی که دارای میزان آگاهی متناسب با دامنه توانایی مورد نظر و کد محتوایی مناسب بودند، بیشتر از ۰/۳۳ ارائه شده و برای کنترل این قضیه، برنامه مجبور شد که از سؤال‌های binهای همجوار استفاده کند که باعث شد دقت اندازه‌گیری کاهش یابد. در این مرحله، خزانه MRP دارای بهترین عملکرد از نظر خطای اندازه‌گیری برآورد توانایی است. در مجموع، هر سه خزانه بهینه در هر دو مرحله بهتر از خزانه‌های سؤال عملیاتی از نظر اندازه خزانه، دقت اندازه‌گیری و امنیت آزمون عمل کردند. در کل، بررسی ملاک دقت اندازه‌گیری توانایی در هر یک از سطوح توانایی نشان داد خزانه‌های سؤالی که با کنترل مواجهه سؤال طراحی می‌شوند، دارای دقت بیشتری هستند نسبت به خزانه‌هایی که بدون کنترل مواجهه طراحی می‌شوند. این نتیجه به دلیل این است که سؤال‌های اضافه شده به خزانه‌های بهینه با کنترل مواجهه S-H دارای سؤال‌های با ضرایب تشخیص بالاتری هستند. در کل، به نظر می‌رسد که خزانه MTI از سؤال‌های موجود در خزانه استفاده بیشتری می‌کند و دارای حداقل سؤال‌های کم‌مواجهه شده است. همچنین، از نرخ همپوشی تست کمی با وجود اینکه دارای حداقل تعداد سؤال است، برخوردار است. درمجموع، بدون توجه به عامل S-H، خزانه‌های بهینه MTI نسبت به خزانه‌های R وMRP دارای سؤال‌های کمتری هستند و از امنیت بالایی نیز برخوردارند و از سؤال‌ها استفاده بیشتری می‌کنند. خزانه‌های MRP زمانی که عامل S-H وارد می‌شود، دارای دقت اندازه‌گیری بالاتری است ولی از نظر اقتصادی به صرفه نیست. بنابراین، پیشنهاد می‌شود زمانی‌که به‌صرفه بودن طراحی خزانه‌های سؤال، تعادل محتوایی و امنیت آزمون عامل بسیار مهمی است، برای کاهش تعداد سؤال‌های مورد نیاز در خزانه CAT از روش MTI استفاده

شود. همچنین، نکته دیگری که باید به آن توجه کرد این است که وقتی عامل S-H وارد می‌شود، در هر سه خزانه سؤال بهینه، میزان تخطی از قیود محتوایی تست‌ها بیشتر از زمانی است که S-H وارد نشده بود؛ این نتیجه دلیلی بر این امر است که وارد کردن کنترل مواجهه S-H بر انتخاب سؤال‌ها تأثیر می‌گذارد و این امکان وجود دارد که برنامه CAT، سؤالی را برای اجرا انتخاب کند که از قیود محتوایی تخطی دارد.

جدول (۱۵) درصد سؤال‌های بیش مواجهه شده و کم مواجهه شده در هر محتوا

آماره	بخش ۱				بخش ۲				بخش ۳			
خزانه‌های سؤال	OP	R	MRP	MTI	OP	R	MRP	MTI	OP	R	MRP	MTI
نرخ مواجهه بزرگتر	۹٪/۸۱	۵٪/٦٤	٤٪/۸۷	٤٪/۱۸۹	۷٪/۵۸	۵٪/۵۲	٤٪/۹۲	۳٪/۷۵	۷٪/۰۸	۵٪/۲٦	٤٪/۱٦	۳٪/۹۲
اندازه خزانه	٤۰۰	۳۱۳	۳۰۱	۲۷۱	۵۰۰	۳۱۹	۳۰٤	۲۷۵	۵۰۰	۳۲۸	۳۱۰	۲۸۸

1/3

جدول (١٦) تعداد تست‌ها بر اساس تخطی از قیود محتوایی در CATها، بدون عامل کنترل مواجهه S-H

سطوح توانایی / خزانه سؤال	total CAT	OP	R (ROP_4)	MRP (ROP_5)	MTI (ROP_6)	کل
-٤	٣٥٧	٣٥٧	٦٠	٦٠	٥٤	٢٩٥
-٣/٥	٤٧٦	٤٧٦	٤٣	٤٩	٤٨	٣٠٧
-٣	٥٩٥	٥٩٥	٣٣	٣١	٣٧	٤٠٤
-٢/٥	٣٥٧	٣٥٧	٢٤	٤٨	٣٠	٦٠١٧
-٢	٣٥٧	٣٥٧	١٥	٣١	٢٤	٩٥٢
-١/٥	٣٥٧	٣٥٧	٦	٢١	١٢	٨٣٣
-١	٩٥	٩٥	٦	٧	٢٤	٨٣٣
-٠/٥	١١	١١	١٢	٩	١٢	٣٥٧
٠	١١	١١	٦	٠	٠	٢٣٨
٠/٥	٢٢	٢٢	٤	٠	٩	٤٧
١	١١٩	١١٩	٢٧	١٨	٤	١٨
١/٥	٢٣٨	٢٣٨	١٢	٧	٧	١٨
٢	٤٧	٤٧	١٨	٦	٦	٤٧
٢/٥	٣٥٧	٣٥٧	١٨	٠	٠	٣٥٧
٣	٨٣٣	٨٣٣	٢٤	٤	٩	٨٣٣
٣/٥	٨٣٣	٨٣٣	٣٦	٤	١٣	٨٣٣
٤	٩٥٢	٩٥٢	٦٠	٠	١٨	٩٥٢
کل	٦٠١٧	١١٩٠٣	١٢٠٠٠	١٢٠٠٠	١٢٠٠٠	

جدول (۱۷) درصد تست‌ها بر اساس تخطی از قیود محتوایی در **CAT**ها، بدون عامل کنترل مواجهه **S-H**

سطوح توانایی / خزانه سؤال	OP	R (ROP_4)	MRP (ROP_5)	MTI (ROP_6)	کل
total CAT	۱۱۹.۳	۱۲۰۰۰	۱۲۰۰۰	۱۲۰۰۰	۰/۵۴۸
۴-	۰/۰۳	۰/۰۰۵	۰/۰۰۵	۰/۰۰۴۵	۰/۰۳۴
۳/۵-	۰/۰۴	۰/۰۰۳۶	۰/۰۰۴۱	۰/۰۰۴	۰/۰۲۵
۳-	۰/۰۵	۰/۰۰۲۸	۰/۰۰۲۶	۰/۰۰۳۱	۰/۰۲۶
۲/۵-	۰/۰۴	۰/۰۰۲	۰/۰۰۴	۰/۰۰۲۵	۰/۰۱۵
۲-	۰/۰۴	۰/۰۰۱۳	۰/۰۰۲۶	۰/۰۰۲	۰/۰۱۱
۱/۵-	۰/۰۴	۰/۰۰۰۵	۰/۰۰۱۸	۰/۰۰۱	۰/۰۰۳
۱-	۰/۰۰۸	۰/۰۰۰۵	۰/۰۰۰۶	۰/۰۰۲	۰/۰۰۱
۰/۵-	۰/۰۰۱	۰/۰۰۱	۰/۰۰۰۸	۰/۰۰۱	۰/۰۰۱
۰	۰/۰۰۱	۰/۰۰۰۵	۰	۰	۰/۰۰۰۳
۰/۵	۰/۰۰۰۲	۰/۰۰۰۳	۰	۰/۰۰۰۸	۰/۰۰۰۳
۱	۰/۰۱	/۰۰۲۳	۰/۰۰۱۵	۰/۰۰۰۳	۰/۰۰۰۱
۱/۵	۰/۰۲	۰/۰۰۱	۰/۰۰۰۶	۰/۰۰۰۶	۰/۰۰۰۱
۲	۰/۰۰۴	۰/۰۰۱۵	۰/۰۰۰۵	۰/۰۰۰۵	۰/۰۰۰۱
۲/۵	۰/۰۳	۰/۰۰۱۵	۰	۰	۰/۰۰۰۱
۳	۰/۰۷	۰/۰۰۲	۰/۰۰۰۳	۰/۰۰۰۸	۰/۰۰۰۱
۳/۵	۰/۰۷	۰/۰۰۳	۰/۰۰۰۳	۰/۰۰۱۱	۰/۰۰۰۱
۴	۰/۰۸	۰/۰۰۵	۰	۰/۰۰۱۵	۰/۰۰۰۱

۳- مقایسه عملکرد CATهای طراحی شده بر اساس خزانه سؤال بهینه و CAT طراحی شده بر اساس خزانه سؤال عملیاتی

الف) بررسی نسبت کارایی[1] میانگین مجذور خطاهای برآورد پارامتر توانایی

به منظور بررسی کارایی هر یک از مدل‌های طراحی شده CAT، در این بخش کارایی هر مدل نسبت به مدل دیگر ارزیابی شده است. بر این اساس، مقدار میانگین مجذور خطا در هر مدل بر مدل دیگر تقسیم شده و اگر حاصل این تقسیم کوچک‌تر از یک بود، مدلی که در صورت قرار دارد کاراتر است و بالعکس (چانگ و ینگ، ۱۹۹۹؛ ریکیسی، ۲۰۰۵). به عبارت دیگر، اگر رابطه زیر بین دو مدل برقرار باشد، مدل اول، مدلی با کارایی بیشتر است و میزان میانگین مجذور خطاهای کمتری دارد. در جدول (۱۸)، نتایج مربوط به مقایسه تمام MSEهای ممکن الگوهای طراحی شده برای CATهای شبیه‌سازی شده و عملیاتی را ارائه کرده است. با بررسی خانه‌های این ماتریس می‌توانیم، کارایی هریک از مدل‌ها را نسبت به هم بررسی کنیم. در مجموع، نتایج نشان می‌دهد که هر ٦ الگوی خزانه سؤال بهینه، از خزانه عملیاتی کاراتر است. بنابراین، طراحی سؤال به منظور اجرای آزمون CAT از طراحی سؤال برای آزمون‌های مداد-کاغذی، به‌مراتب به‌صرفه‌تر و بهینه‌تر است. همچنین، در بین خزانه‌های بهینه نیز ROP_5 از همه کاراتر است.

$$\frac{MSE_1}{MSE_2} < 1 \Rightarrow MSE_1 \rightarrow \text{efficiency} \uparrow$$

[1]. efficiency

جدول (۱۸) مقایسه **MSE** ها در الگوهای **CAT** شبیه‌سازی شده و عملیاتی (کارایی الگوها)

خزانه سؤال	MSE		OP	ROP_1	ROP_2	ROP_3	ROP_4	ROP_5	ROP_6
OP	۰/۰۱۷٤٥	OP	–	۱/۳۸	۱/٥۳	۲/۲۹	۱/۳۸	۲/۳۹	۱/٤۳
ROP_1	۰/۰۱۲٦۷	ROP_1	۰/۷۳	–	**۰/۹۲**	۱/٦۷	۱/۰۰۱	۱/۷٤	۱/۰۳
ROP_2	۰/۰۱۱۳۸	ROP_2	۰/٦٥	۰/۹۰	–	۱/٤۹	۰/۹۰	۱/٥٦	۰/۹۳
ROP_3	۰/۰۰۷٦	ROP_3	۰/٤٤	۰/٦۰	۰/٦٤	–	۰/٦۰	۱/۰٤	**۰/٦۲**
ROP_4	۰/۰۱۲٦٥	ROP_4	۰/۷۲	۰/۷۲	۱/۱۱	۱/٦٦	–	۱/۷۳	۱/۰۳
ROP_5	۰/۰۰۷۳	ROP_5	۰/٤۲	۰/٤۲	۰/٦٥	۰/۹٦	۰/٥۸	–	۰/٦۰
ROP_6	۰/۰۱۱۲٤	ROP_6	۰/٦٤	۰/۸۸	۰/۹۸	۱/٤۷	۰/۸۹	۱/٥٤	–

ب) بررسی نتایج مقایسه کجی توزیع نرخ مواجهه سؤال در الگوهای **CAT** شبیه‌سازی شده بر اساس خزانه سؤال بهینه و **CAT** عملیاتی

آماره χ^2 برای اندازه‌گیری میزان کجی توزیع مواجهه سؤال، به کار می‌رود (چانگ و ییِنگ، ۱۹۹۹). این آماره اختلاف بین نرخ‌های مواجهه سؤال مشاهده شده و ایدئال را محاسبه و مقدار اثربخشی استفاده از خزانه سؤال به کار رفته برای CAT را نیز تعیین می‌کند. مقدار χ^2 پایین نشان می‌دهد که بیشتر سؤال‌ها استفاده شدند. نسبت اندازه‌های χ^2 از توزیع F پیروی می‌کند. همچنین می‌توان برای مقایسه نرخ‌های مواجهه سؤال در دو روش، فرمول زیر را به کار برد:

$$F_{method_i,method_j} = {\chi^2_{method_i}} \Big/ {\chi^2_{method_j}}$$

اگر $F<1$ باشد، پس روش اول نسبت به روش دوم، از نظر تعادل کلی نرخ‌های مواجهه سؤال بهتر در نظرگرفته می‌شود. نتایج جدول (۱۹) نشان می‌دهد که سؤال‌ها در خزانه‌هایی که با وارد کردن عامل S-H شبیه‌سازی شده‌اند (یعنی، ROP_4, ROP_6,...)، به‌طور متعادل‌تری از سؤال‌ها استفاده می‌کنند. با اینکه دارای تعداد سؤال‌های بیشتری هستند، چون از مواجهه بیش از حد سؤال و همچنین کم مواجهه

شدن سؤال‌ها، حداقل تا نرخ ۰/۰۲، جلوگیری می‌کنند، استفاده از خزانه سؤال را متعادل‌تر می‌کنند. بررسی خانه‌های ماتریس بالا، نشان‌دهنده آن است. این نتایج گویای این مطلب است که تعداد یا اندازه خزانه سؤال، عامل تعیین‌کننده‌ای برای میزان کجی نرخ مواجهه نیست. تنها تفاوت آنها وارد شدن عامل تعادل محتوایی است.

جدول (۱۹) مقایسه نتایج میزان تعادل کلی نرخ‌های مواجهه سؤال در الگوهای CAT شبیه‌سازی شده

خزانه سؤال	χ^2	$\chi^2_{method_i} / \chi^2_{method_j}$		J_1	J_2	J_3	J_4	J_5	J_6	J_7
				OP	ROP_1	ROP_2	ROP_3	ROP_4	ROP_5	ROP_6
OP	۹۸/۸۵۶	I_1	OP	*	۱/۹۲	۲/۰٤	۲/۲۷	۳/۲۷	۳/٦۹	۳/٦٦
ROP_1	۵۱/٤۰	I_2	ROP_1	۰/۵۲	*	۱/۰٦	۱/۱۸	۱/۷۰	۱/۹۲	۱/۹۰
ROP_2	٤۸/٤۳٦	I_3	ROP_2	۰/٤۹	۰/۹٤	*	۱/۱۱	۱/٦۰	۱/۸۱	۱/۷۹
ROP_3	٤۳/٤۹٤	I_4	ROP_3	۰/٤٤	۰/۸۵	۰/۹۰	*	۱/٤٤	۱/٦۲	۱/٦۱
ROP_4	۳۰/۲۵۲	I_5	ROP_4	۰/۳۱	۰/۵۹	۰/٦۲	۰/۷۰	*	۱/۱۳	۱/۱۲
ROP_5	۲٦/۷٤۱	I_6	ROP_5	۰/۲٦	۰/۵۲	۰/۵۵	۰/٦۱	۰/۸۸	*	۰/۹۹
ROP_6	۲٦/۹۹	I_7	ROP_6	۰/۲۷	۰/۵۳	۰/۵٦	۰/٦۲	۰/۸۹	۱/۰۰۹	*

بحث و نتیجه‌گیری

هدف از اجرای این پژوهش، طراحی از طریق استخراج یک مدل بهینه CAT و کاربرد آن برای اجرای آزمون تولیمو به شیوه سنجش انطباقی کامپیوتری در سازمان سنجش آموزش کشور بود. برای محقق کردن این هدف، براساس تلفیق دو رویکرد برنامه‌نویسی خطی و روش مونت‌کارلو، به شش روش، خزانه‌های سؤال بهینه برای آزمون CAT شبیه‌سازی شده، طراحی شد. همچنین، با یک بررسی تجربی، بهترین روش برآورد پارامتر توانایی و مناسب‌ترین ملاک توقف آزمون برای اجرای آزمون تولیمو به شکل CAT استخراج شد. سپس یک برنامه عملیاتی CAT برای آزمون

تولیمو به زبان PHP نیز نوشته شد. در آزمون CAT عملیاتی، ویژگی سؤال‌ها بر اساس سؤال‌های اجرا شده واقعی در ده دوره آزمون مداد-کاغذی و پارامترهای توانایی آزمودنی‌ها نیز بر اساس اجراهای واقعی آن ده دوره به دو شیوه بیشینه درست‌نمایی و پسین مورد انتظار مدرج‌سازی و در پایگاه داده MYSQL ذخیره‌سازی شدند، تا بر این اساس ویژگی کامل یک CAT عملیاتی بازسازی شود. بنابراین، در برنامه CAT، خزانه سؤال و ویژگی‌های محتوایی آن همانند آزمون مداد-کاغذی بود و ارائه بیش از حد سؤال‌ها با روش کنترل مواجهه سیمپسون-هتر کنترل شد. برای استخراج بهترین برنامه CAT عملیاتی، هشت طرح مختلف بر اساس شیوه برآورد پارامتر توانایی و ملاک توقف آزمون برنامه‌ریزی شد. نتایج این مطالعه نشان داد که برای آزمون تولیمو، بهترین روش برآورد پارامتر توانایی، روش بیشینه درست‌نمایی است و بهترین ملاک طول ثابت ۱۴۰ سؤالی. در پایان نتایج CATهای شبیه‌سازی شده و عملیاتی با یکدیگر و همچنین با آزمون مداد-کاغذی مقایسه شد.

یکی از حوزه‌هایی که در آن استفاده گسترده‌ای از CAT شده است، سنجش مهارت و توانایی زبان انگلیسی است. این نوع سنجش طی دو دهه اخیر رشد فزاینده‌ای داشته است. آزمون‌هایی از قبیل TOFEL، GRE و GMAT امروزه به شکل آزمون CAT اجرا می‌شود (رادنر[1] ۲۰۱۰). مطالعات تجربی سیستماتیکی روی کاربرد CAT در مورد مهارت زبان صورت گرفته است، شامل مهارت شنیداری (دانکل[2] ۱۹۹۹؛ مادسن[3] ۱۹۹۱)، مهارت خواندن (چالهوب-دیلول[4] ۱۹۹۹؛ کایا-کارتون، کارتون و داندونولی[5] ۱۹۹۱) و مهارت واژگان (لافر و گلدشتاین[6] ۲۰۰۴). مطالعات پیشین در مورد آزمون CAT واژگان نشان داد که CAT می‌تواند به سطوح برابر یا بالاتری از اعتبار و روایی نسبت به آزمون‌های مداد-کاغذی حتی با تعداد سؤال‌های کمتر برسند (برای مثال ویزپل[7] ۱۹۹۳؛ ۱۹۹۸؛ ویزپل و همکاران، ۱۹۹۴). اغلب مطالعات انجام گرفته در مورد آزمون CAT مهارت زبان نشان می‌دهد که

---

1. Rudner
2. Dunkel
3. Madsen
4. Chalhoub-Deville
5. Kaya-Carton, Carton & Dandonoli
6. Laufer & Goldstein
7. Vispoel

CAT نسبت به آزمون‌های مداد-کاغذی به تعداد سؤال کمتری نیاز دارد (ون-تاو-تسنگ، ۲۰۱۶)، امّا اینکه آیا می‌تواند پارامتر توانایی آزمودنی‌ها را با دقت بیشتری برآورد کند، سؤالی است که ذهن مؤسسات اجرایی آزمون‌ها را به خود جلب کرده است. اجرای آزمون انطباقی به‌خصوص اگر به شیوه کامپیوتری اجرا شود، هزینه‌های اولیه سنگینی را متحمل شرکت‌های برگزاری آزمون خواهد کرد، از این‌رو اگر مزیت CAT که اجرایی دقیق و بهینه برای برآورد توانایی است، برقرار نشود، فایده و مزیتی برای این مراکز نخواهد داشت. از این‌رو، برای بررسی این امر مهم، پژوهش حاضر به بررسی و استخراج یک مدل بهینه برای اجرایی شدن CAT در سازمان سنجش آموزش کشور پرداخته است. به منظور عملیاتی کردن این اجرا، مورد مطالعه در پژوهش حاضر، آزمون تولیمو بوده است. نتایج پژوهش نشان داد که به دلیل آنکه سازمان سنجش آموزش کشور دارای خزانه سؤال بسیار بزرگ و مناسبی برای آزمون تولیمو است (۱۳۵ دوره آزمون تولیمو اجرا شده است که تقریباً باید حدود ۱۸۹۰۰ سؤال در خزانه سؤال وجود داشته باشد. البته ما سؤال‌های تکراری خزانه سؤال و نیز سؤال‌های طراحی شده موجود که هیچ‌گاه اجرا نشده‌اند را در نظر نگرفته‌ایم)، امکان اجرایی شدن این آزمون به شیوه CAT ممکن است. در این پژوهش تنها از ۱۰ دوره از ۱۳۵ دوره برگزار شده آزمون نمونه‌گیری شد، و تنها از یک خزانه سؤال ۱۴۰۰ تایی برای مدرج‌سازی و طراحی آزمون CAT استفاده شد، امّا نتایج نشان داد که CAT عملیاتی شکاف زیادی در شاخص‌های ارزیابی و برآورد توانایی نسبت به CAT بهینه شبیه‌سازی‌شده ندارد. البته این قابل ذکر است که مدل‌های CAT که بر اساس خزانه سؤال بهینه، شبیه‌سازی شدند از خزانه عملیاتی کاراتر و بهینه‌تر هستند، این نکته کاملا طبیعی است؛ زیرا طراحی سؤال به‌منظور اجرای آزمون CAT از طراحی سؤال برای آزمون‌های مداد-کاغذی، به‌مراتب به‌صرفه‌تر (یعنی به سؤال‌ها کمتری نیاز دارد) و بهینه‌تر (یعنی پارامتر توانایی را با دقت بالاتری برآورد می‌کند) است. این نتایج با مطالعه ریکیسی (۲۰۱۰) نیز همخوان است. بنابراین، خزانه‌های سؤال در سنجش انطباقی باید به پیش‌فرض‌های مدل روان‌سنجی که زیربنای مدرج‌سازی، اجرا و نمره‌گذاری است، توجه کنند. بنابراین، تلاشی که برای نوشتن

خزانه سؤال‌های سنجش انطباقی لازم است، بسیار بیشتر از آزمون‌های مداد-کاغذی است (میلمن و آرتر[1] ۱۹۸٤). خزانه سؤال بهینه باید بر اساس مؤلفه‌های دیگر CAT، یعنی طول آزمون، توزیع مورد انتظار توانایی در جامعه آزمودنی‌ها، برآورد توانایی، شیوه‌های انتخاب سؤال و نسبت‌های مواجهه و نرخ همپوشی سؤال (آزمون) نیز تعیین شود.

همچنین، بررسی نتایج مربوط به شیوه برآورد پارامتر توانایی، نشان داد که روش MI نسبت به روش EAP از دقت بالاتری در برآورد پارامتر توانایی برخوردار است. این نتیجه با پژوهش کلندر (۲۰۱۱) هم‌خوانی دارد. کلندر (۲۰۱۱) هر دو روش را برای برآورد پارامتر توانایی کارا نشان داد، البته تنها بر اساس ملاک همبستگی بین برآورد توانایی CAT و مداد-کاغذی، که ملاک کافی برای ارزیابی نیست ولی با پژوهش شریفی و همکاران (۱۳۹۵) هم‌خوانی ندارد. زیرا آنها نشان دادند که روش EAP در برآورد توانایی دقت بیشتری دارد، همچنین، همبستگی بین برآورد توانایی CAT و اجرای مداد-کاغذی بالاتری دارد. یکی از دلایل متفاوت بودن نتیجه پژوهش حاضر با این پژوهش، می‌تواند این باشد که در آن مطالعه، مواجهه بیش از حد سؤال کنترل نشده بود و همچنین، تعادل محتوایی در آزمون رعایت نشده بود. دلیل دیگر این است که نوع این آزمون ملاک‌مرجع است در صورتی‌که آزمونی که در آن پژوهش بررسی شد، نرم‌مرجع بود. دلیل آخر احتمالی، می‌تواند این موضوع باشد که در آن مطالعه، آزمودنی‌هایی که در الگوی پاسخ خود سؤال غلط یا درست نداشتند، دارای برآوردی نامتناهی می‌شدند و از آزمون حذف می‌شدند، ولی در پژوهش حاضر برای رفع نقص این روش، تا زمانی که آزمودنی پاسخ درست یا غلط در الگوی پاسخ خود نداشت، از روش برآورد اوون (۱۹۷۵) استفاده می‌شد.

همچنین در خصوص شیوه انتخاب سؤال، برخی از پژوهش‌ها نشان دادند (چانگ، ۲۰۱٤، ژنگ و چانگ[2] ۲۰۱۵)، به‌دلیل آنکه انتخاب سؤال در CAT بر اساس روش بیشینه آگاهی مبتنی است، زمانی که CAT برای آزمون‌های خطیر یا مقیاس وسیع به‌کار می‌رود، قضیه برآورد توانایی را پیچیده و بغرنج می‌کند و گاهی باعث کم‌برآورد شدن یا بیش‌برآورد شدن توانایی می‌شود. بنابراین، پژوهشگران اندازه‌گیری

---

[1].Millman &Arter
[2]. Zheng & Chang

چندمرحله‌ای[1](MST) را توسعه دادند، تا به بیش‌برآورد و کم‌برآورد شدن پارامتر توانایی کمک کند. در روش آزمون‌گیری یا اندازه‌گیری چندمرحله‌ای (MST) بر اساس یک سؤال پارامتر توانایی آزمودنی برآورد نمی‌شود، بلکه بر اساس یک مرحله برآورد می‌شود. هر مرحله شامل تعدادی ماژول[2]برنامه است، درحالی‌که هر ماژول سطح دشواری متفاوتی دارد. این راهکار کمک می‌کند تا بیش‌برآورد یا کم‌برآورد توانایی آزمودنی‌ها در شروع آزمون کاهش یابد (چایمانکول، پازایفول و کانجانوواسی[3] ۲۰۱٦). در پژوهش حاضر، سؤال‌ها به‌صورت چندمرحله‌ای اجرا نشده است، ولی بهتر است برای آزمون‌هایی شبیه آزمون تولیمو که یک آزمون خطیر است، از این روش استفاده کرد.

همچنین، نتایج مربوط به بررسی ملاک توقف مناسب آزمون، نشان داد که برای اجرای آزمون تولیمو به شکل CAT، طول ثابت، بیشترین میزان دقت برآورد پارامتر و همچنین همبستگی بالاتر بین برآوردهای توانایی نسخه CAT و مداد-کاغذی را نشان می‌دهد. این نتیجه با پژوهش گاردنر و همکاران[4](۲۰۰٤) همخوان است. به بیان آنها زمانی که ملاک توقف آزمون طول متغیر بر اساس خطای استاندارد ثابت است، کارایی سنجش انطباقی را پایین می‌آورد، زیرا سؤال‌های اضافی و غیر ضروری که با محتواهای معین شده برای آزمون هماهنگ نیست، برای آزمودنی‌هایی که به ملاک توقف آزمون نمی‌رسند، ارائه می‌کند. برای آزمودنی‌هایی که زودتر به ملاک توقف آزمون می‌رسند، گاهی تعادل محتوایی را کامل نمی‌کند و عادلانه بودن آزمون را زیر سؤال می‌برد. بنابراین، می‌توان اذعان داشت که برای آزمون‌های ملاک‌مرجع که تعادل محتوایی از اهمیت بسزایی برخوردار است، طول ثابت آزمون به‌عنوان ملاک توقف آزمون بیشترین دقت را دارد و ملاک خطای استاندارد ثابت، تخطی از قیود را بالامی‌برد.

در خصوص تهیه خزانه سؤال بهینه برای اجرایی شدن CAT برای آزمون تولیمو، از بین سه مدل R، MRP و MTI، الگوی ROP_5 (یعنی، MRP) از همه کاراتر

---

1. multistage testing
2. module
3. Chaimongkol, Pasiphol & Kanjanawasee
4. Gardner

است. زیرا علاوه بر اینکه مواجهه سؤال را کنترل می‌کند، سؤال‌های متناسب با نوع آزمون طراحی می‌کند. در طراحی سؤال‌ها برای آزمون‌های مداد-کاغذی اغلب سؤال‌ها به‌شیوه تصادفی از نظر ویژگی‌های سؤال طراحی می‌شوند، مانند مدل R در طراحی خزانه سؤال. همان‌طور که نتایج این پژوهش نیز نشان داد، این نوع طراحی سؤال، کمترین میزان دقت اندازه‌گیری و بیشترین تعداد سؤال استفاده نشده در خزانه سؤال را ایجاد می‌کند. ولی از نظر اقتصادی به صرفه نیست. امّا پیشنهاد می‌شود زمانی که به صرفه بودن طراحی خزانه‌های سؤال، تعادل محتوایی و امنیت آزمون عامل بسیار مهمی هستند، برای کاهش تعداد سؤال‌های مورد نیاز در خزانه CAT از روش MTI استفاده شود. همچنین، در عمل باید ذکر کرد که، خزانه‌های سؤال در آزمون CAT ایستا نیستند و باید پویا باشند. در اغلب برنامه‌های سنجش CAT، تست‌ها را از خزانه‌های سؤال انتخاب می‌کنند، بنابراین، باید همواره سؤال‌های جدید به صورت متوالی پیش‌آزمون شوند و سپس به خزانه اضافه شوند. سؤال‌های بسیار استفاده شده یا قدیمی[1] طی زمان‌های متوالی از خزانه حذف شوند. بنابراین، نظارت بر استفاده مناسب از سؤال‌ها و دوباره جایگزین کردن خزانه از سؤال‌های جدید، دو وظیفه مهم مدیریت و محافظت از خزانه سؤال است (ون‌درلیندن و ولدکمپ، ۲۰۰۰). بنابراین با یک مدیریت دقیق بر خزانه سؤال آزمون تولیمو، اجرای CAT برای این آزمون کاملاً امکان‌پذیر است.

برای اجرایی شدن ساخت یک خزانه سؤال نسبتاً بهینه با امکانات فعلی سازمان سنجش آموزش کشور، نخست باید سؤال‌های موجود در خزانه سؤال تولیمو توسط روان‌سنجان، مدرج‌سازی و سؤال‌های مناسب بر اساس وزن‌های محتوایی‌شان در یک سیستم بانک سؤال هوشمند ذخیره‌سازی شوند. بهتر است سؤال‌ها براساس مدل دو و سه پارامتر لوجستیک مدرج‌سازی شوند. در مرحله دوم، اطلاعات کلی موجود در خصوص پارامتر توانایی آزمودنی‌ها در نرم‌افزار CAT ذخیره‌سازی شود و اجراهای CAT برای هریک از آزمودنی‌های مفروض انجام گیرد، در این مرحله پس از اجراهای متوالی، سؤال‌های نامناسب حذف شوند. در مرحله سوم، سؤال‌هایی تهیه و طراحی شوند که برای اجراهای CAT مناسب باشند تا کارایی خزانه سؤال را بالاتر ببرند. طراحی خزانه سؤال بر اساس نتایجی که از شبیه‌سازی مونت‌کارلو (ریکیسی،

---

[1]. obsolete

۲۰۰۳) و برنامه‌نویسی خطی WDM (ون‌درلیندن، ۲۰۱۰) استخراج شد، الگوی بسیار مناسبی برای طراحان سؤال بر اساس ویژگی‌های آماری و غیر آماری است. نتایج پژوهش حاضر نشان داد که، در این نوع آزمون سرنوشت‌ساز باید مواجهه سؤال کنترل شود تا دقت اندازه‌گیری پارامتر توانایی بالاتر رود.

همچنین، در این پژوهش، به دلیل اینکه امنیت آزمون تولیمو از اهمیت بسزایی برخوردار است، به‌منظور جلوگیری از افشای سؤال‌ها از روش کنترل مواجهه سؤال استفاده شد. که روش بسیار پیچیده‌ای در اجرای عملیاتی برنامه‌های CAT در آزمون‌های خطیر است. همچنین، بهتر است نحوه ارائه سؤال در نمایشگر کامپیوتر به‌شکلی تعبیه شود تا احتمال به خاطر سپردن متن سؤال و افشای آن در خارج از محیط آزمون کاهش یابد (برای مثال، فرصت دو بار خواندن برای هر سؤال و بین هر سؤال ۵ ثانیه زمان گذاشتن). اما مهم‌ترین عامل برای کنترل افشای سؤال‌ها، اعمال روشی پر قدرت در مواجهه بیش از حد سؤال است. در پایان، هر برنامه CAT به‌ویژه اگر به آزمون‌های خطیر مربوط باشد، باید به‌روزرسانی شود؛ به همین دلیل، باید در خصوص این آزمون، همان‌طور که در قسمت یافته‌ها نیز ذکر شد، خزانه‌های سؤال برنامه CAT برای دوره یک‌ساله تهیه شود و برای جلوگیری از افشای سؤال‌های آن باید سالانه خزانه سؤال، بررسی و به‌روزرسانی شود. همچنین، نظارت و نگهداری خزانه سؤال نیز مدنظر اجرا کنندگان قرار گیرد.

منابع

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*، ۶(۱٤)، ۲۰۳-۲۳٤.

مقدسین، مریم (۱۳۹۵). تلفیق رویکرد ریکیسی و رویکرد برنامه‌نویسی ریاضی در طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری. *فصلنامه اندازه‌گیری تربیتی*، ۷(۲٦)، ۱٤۹-۱۹۷.

مقدسین، مریم؛ فلسفی‌نژاد، محمدرضا؛ دلاور، علی؛ جمالی، احسان؛ و فرخی، نورعلی. (۱۳۹٤). طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری با درنظر گرفتن امنیت آزمون. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۵(۱۰)، ۱۳۳-۱۷۸.

مینایی، اصغر؛ و فلسفی‌نژاد، محمدرضا (۱۳۸۹). روش‌های سنجش تک‌بعدی بودن سؤال‌ها در مدل دوارزشی IRT. *فصلنامه اندازه‌گیری تربیتی*، ۱(۳)، ۷۱-۱۰۰.

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).

Barrada, J., Olea, J., Ponsada, V., Abad, F., Ponsoda, V., & Abad, F. J. (2009). Test overlap rate and item exposure rate as indicators of test security in CATs. In Proceedings of *the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [date] from www. psych. umn. Edu/psylabs/CATCentral.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J.Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Boyd, A. M., Dodd, B., & Fitzpatrick, S. (2013). A comparison of exposure control procedures in CAT systems based on different measurement models for testlets. *Applied Measurement in Education*, *26*(2), 113-135.

Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. *Handbook of polytomous item response theory models*, 229-255.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems*.

Brooke, A., Kendrick, D., & Meeraus, A. (1988). *GAMS: A user's guide*. Redwood City CA: The Scientific Press.

Chaimongkol, N., Pasiphol, S., & Kanjanawasee, S. (2016). Computerized Adaptive Testing with Reflective Feedback: A Conceptual Framework. *Procedia-Social & Behavioral Sciences*, 217, 806-812.

Chalhoub-Deville (Ed.). *Issues in computer-adaptive testing of reading proficiency*. Cambridge: University of Cambridge Local Examinations Syndicate.

Chang, H. H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage.

Chang, H. H. (2014). *Psychometrics behind computerized adaptive testing. Psychometrika*. Published online Feb. 2014. DOI: 10.1007/S11336- 014-9401-5.

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.

Chang, S. W., & Twu, B. Y. (1998). *A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing*.

Chang, Y. C. I., & Ying, Z. (2004). Sequential estimation in variable length computerized adaptive testing. *Journal of Statistical Planning and Inference*, *121*(2), 249-264.

Chang, H. H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.

Chang, H. H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a-stratified computerized adaptive testing. *Applied Psychological Measurement*, 27, 262-274.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely con- strained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369-383.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing* (No. ACT-RR-99-5): American College Testing Program, Iowa City, IA.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational & Behavioral Statistics*, *22*(3), 265-289.

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational & Psychological Measurement*, *71*(1), 37-53.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement, 41*(3), 178-194.

CITO (1999). *WISCAT*. Een computergestuurd toetspakket voor rekenen en wiskunde. [Mathcat: A computerized test package for arithmetic and mathematics]. CITO: Arnhem.

CITO (2002). *NT2cat*. Een computergestuurd toetspakket voor Nederlands als tweede taal. [DSLcat. A computerized test package for Dutch as a Second Language]. CITO: Arnhem.

CITO (2008). *TURCAT*. Een computergestuurd toetspakket voor Turks als tweede taal. [TURCAT. A computerized test package for Turkish as a Second Language]. CITO: Arnhem.

Davey, T., & Nering, M. (2002). *Controlling item exposure and maintaining item security*. Computer-based testing: Building the foundation for future assessments, 165-191.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items*. Doctoral dissertation.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational & psychological measurement, 53*(1), 61-77.

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 1-3). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Eggen, T. J. H. M. (2004). *CATs for kids: easy and efficient. Paper presented at the 2004 meeting of Association of Test Publishers*. Palm Springs, CA.

Flaugher, R. (2000). *Item pools. In H. Wainer (Ed.), Computerized adaptive testing: A primer* (pp. 37-59). Mahwah, NJ: Lawrence Erlbaum.

French, B. F., & Thompson, T. D. (2003). The evaluation of exposure control procedures for an operational CAT. In Poster presented at *the annual meeting of the American Educational Research Association (AERA)*, Chicago, IL.

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: a simulation study. *BMC psychiatry*, *4*(1), 13.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment, 5*(8).

Gu, L. (2007). *Designing optimal item pools for computerized adaptive tests with exposure controls*. Unpublished doctoral dissertation. Michigan State University.

Gu, L. & Reckase, M. D. (2007). *Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control*. Paper Presented at the 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park CA: Sage.

Han, K. T. (2012). An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement*, *49*(3), 225-246.

Han, K. T. (2011). *User's manual: SimulCAT*. Retrieved March, 1, 2013.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first. *Journal of Educational Measurement, 38*(3), 249-266.

He, W., & Reckase, M. (2010). *Optimal item pool design for a highly constrained computerized adaptive test*. Unpublished doctoral dissertaion. Michigan State University.

Kanjanawasee, S**.** **(**2012). *Modern Test Theory* (Ed.4). Bangkok: Chulalongkorn University Press.

Kaya-Carton, E., Carton, A. S., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.). *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. Unpublished doctoral dissertation, Middle East Technical University, Ankara, Turkey.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, Strength, and Computer adaptiveness. *Language Learning*, 54, 399-436. doi: 10.1111/j.0023-8333.2004.00260.x.

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376-392.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied measurement in education*, *2*(4), 359-375.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational & Psychological Measurement, 72*(1), 159-175.

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

McBride, J. R., & Weiss, D. J. (1976). *Some properties of a Bayesian adaptive ability testing strategy* (Research Rep No. 76-1). Minneapolis, MN: Psychometric Methods Program, Department of Psychology.

Millman, J., & Arter, J. A. (1984). *Issues in item banking. Journal of Educational Measurement*, 21, 315-330.

National Council of State Boards of Nursing, & National Council of State Boards of Nursing. (2011). *RN practice analysis: Linking the NCLEX-RN examination to practice*. NCSBN Research Brief, 53.

Ozturk, N. B., & Dogan, N. (2015). Investigating Item Exposure Control Methods in Computerized Adaptive Testing. *Educational Sciences: Theory and Practice, 15*(1), 85-98.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical Considerations in Computer-Based Testing*. New York: Springer.

Phankokkruad, M. (2012). *Association Rules for Data Mining in Item Classification Algorithm: Web Service Approach*.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues & Practice, 8*(3), 11-15.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests*. Paper presented at the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D., & He, W. (2004). *The ideal item pool for the NCLEX-RN examination*— Report to NCSBN: Michigan State University.

Reckase, M. D., & He, W. (2005). *Ideal item pool design for the NCLEX-RN exam.* Michigan State University, East Lansing, MI.

Reckase, M. D. (2007). The design of p-optimal item bank for computerized adaptive tests. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

Reckase, M. D. (2009). *Optimal Item Pool Design for the 2009 NCLEX Exam.* A Report SubMTIted to National Council of State Boards of Nursing March 2009.

Reckase, M. D., & He, W. (2009a). *Optimal item pool design for the 2009 NCLEX Exam*--report to the National Council of State Boards of Nursing (NCSBN): Michigan State University.

Reckase, M. D., & He, W. (2009b). *The influence of item pool quality on the functioning of computerized adaptive tests*. Paper presented at the annual meeting of Psychometric Society, Cambridge, U.K.

Reckase, M. D. (2010). Designing Item Pools to Optimize the Functioning of Computerized Adaptive Test. *Psychological Test and Assessment Modeling*,*52*(2), 127-141.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311-327.

Rudner, L. M. (2010). Implementing the Graduate Management Admission Test computerized adaptive test. In W. J. van der

Linden & C. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement*, 17, 167–176.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the *27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, 97, 69-85

Van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23, 195-210.

Van der Linden, W. J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practise*. Boston, MA: Kluwer Academic Publishers.

Van der Linden, W. J., & Pashley, P. J. (2010). *Item selection and ability estimation in adaptive testing*. In W. J. van der Linden & C. W. Glas (Eds.), Elements of adaptive testing. New York: Springer.

Veldkamp, B. P, Vershoor, A. J., Eggen, T. J. (2007). *A Multiple Objective Test Assembly Approach for Exposure Control Problems in Computerized Adaptive Testing*. Measurement and Research Department Reports.

Verschoor, A. J., & Straetmans, G. J. J. N. (2000). MathCAT: A flexible testing system in mathematics education for adults. In W.J. van der Linden, and C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 101-116). Boston, MA: Kluwer Academic Publishers.

Vispoel, W. P. (1993). Computerized adaptive and fixed-item versions of the ITED vocabulary subtest. *Education and Psychological Measurement, 53*, 779-788. doi: 10.1177/0013164493053003022.

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement*, 35, 155-167. doi: 10.1111/j.1745-3984.1998.tb00532.x

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education*, 7, 53-79. doi: 10.1207/s15324818ame0701_5

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000*). Computerized adaptive testing: a primer* (2nd edition). Mahwah, NJ: Lawrence Erlba.

Wang, H. P., Kuo, B. C., Chao, R. C., & Tsai, Y. H. (2012). The Development and Evaluation of a Computerized Adaptive Testing System for Chinese Proficiency-base on CEFR. *Procedia-Social and Behavioral Sciences*, 64, 34-42.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104-1.