فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی **Journal of Educational Measurement & Evaluation Studies**
سال پنجم، شمارهٔ ۱۰، تابستان ۱۳۹٤، صص ۱۷۷-۱۳۳ **Vol. 5, No. 10, Summer 2015**

طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری با در نظر گرفتن امنیت آزمون

مریم مقدسین*
محمدرضا فلسفی‌نژاد**
علی دلاور***
احسان جمالی****
نورعلی فرخی*****

تاریخ دریافت مقاله: ۹۳/۱۰/۰۱
تاریخ پذیرش مقاله: ۹٤/۰٦/۱٦

چکیده

سنجش انطباقی کامپیوتری به خزانه سؤالی نیاز دارد که به خوبی طراحی شده و برای ساخت آزمون‌های مجزا، تعداد مناسبی سؤال داشته باشد. همچنین شامل سؤال‌هایی باشد که از لحاظ محتوایی متعادل باشد و هزینه ساخت آزمون را کاهش دهد. یکی از روش‌های طراحی خزانه سؤال، روش رِکیس است، که در آن از روش مونت‌کارلو برای تعیین ویژگی‌های یک خزانه سؤال بهینه استفاده می‌شود. در این پژوهش، از این روش برای طراحی خزانه‌هایی که با مدل سه پارامتری لُجستیک مدرج شده‌اند، استفاده شده است. برای کنترل نرخ مواجهه سؤال از روش سیمپسون-هتر و برای شبیه‌سازی سؤال‌های آزمون از سه روش تصادفی (R)، تصادفی آمیخته و پیش‌بینی (MRP) و حداقل آگاهی آزمون (MTI) استفاده شده است. عملکرد خزانه‌های سؤال شبیه‌سازی شده و عملیاتی با در نظر گرفتن مجموعه‌ای از ملاک‌های ارزیابی، با یکدیگر مقایسه شده‌اند. نتایج نشان می‌دهد که خزانه MRP نسبت به سایر خزانه‌های بهینه از امنیت و دقت اندازه‌گیری بالاتری برخوردار است و شامل سؤال‌های زیادی با ضریب تشخیص بالا است. نرخ همپوشی آزمون در آن کمتر و درصد کوچکی از سؤال‌های بیش مواجهه و کم مواجهه دارد. در مقابل اندازه خزانه MTI نسبت به سایر خزانه‌های بهینه کوچک‌تر و سؤال‌هایی با ضریب تشخیص پایین‌تر دارد. به‌طور کلی با در نظر گرفتن عامل کنترل مواجهه، خزانه‌های بهینه، بهتر از خزانهٔ عملیاتی عمل می‌کند.

واژه‌های کلیدی: سنجش انطباقی کامپیوتری، خزانه سؤال بهینه، روش مونت‌کارلو-رِکیس، کنترل نرخ مواجهه سؤال، روش سیمپسون-هتر.

---

*دانشجوی دوره دکتری سنجش و اندازه‌گیری دانشگاه علامه طباطبایی (نویسنده مسئول (mmoghadasin@yahoo.com)
** دانشیار دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبایی
*** استاد دانشکده روانشناسی و علوم تربیتی دانشگاه علامه‌طباطبایی
**** استادیار سازمان سنجش آموزش کشور
***** دانشیار دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبایی

مقدمه

سنجش انطباقی کامپیوتری[1] (CAT) شیوه‌ای از اجرای آزمون است که در آن سؤال‌ها به صورتی انتخاب می‌شوند که با سطح توانایی آزمودنی‌ها مطابقت داشته باشد و بدون از دست دادن دقت آزمون، طول آن را کاهش داد (رکیس[2]، ۲۰۱۰). در این شیوه توانایی آزمودنی بعد از پاسخ به هر سؤال، دوباره برآورد می‌شود. همچنین، سؤال‌های بعدی به شکلی که دارای ویژگی‌های بهینه‌ای باشند، به صورت سلسله مراتبی[3] (با توجه به تطابق درجهٔ دشواری سؤال با سطح توانایی آزمودنی) از خزانه سؤال انتخاب می‌شود (وندرلیندن و گلاس[4]، ۲۰۱۰). مؤلفه‌های اساسی سنجش انطباقی کامپیوتری (CAT) عبارتند از: مدل نظریهٔ سؤال پاسخ[5] که سؤال‌ها بر اساس آن مدرج می‌شوند، خزانه سؤال مدرج شده[6]، الگوریتم انتخاب سؤال، روش آماری برآورد توانایی آزمودنی و قاعده اتمام آزمون (واینر، دورانس، ایگنور، فلاگر، گرین، میسلوی، استنبرگ و تیسن[7]، ۲۰۰۰). اخیراً روش‌های سنجش انطباقی کامپیوتری (CAT)، برای سنجش‌های سرنوشت‌ساز[8] به مقدار فراوانی به کار می‌روند. از این‌رو، تعادل محتوایی[9] (چنگ و چانگ[10]، ۲۰۰۹) و قواعد امنیتی، مانند کنترل مواجهه سؤال[11] (سیمپسون و هتر[12]، ۱۹۸۵) اهمیت به‌سزایی دارد. یکی از مؤلفه‌های مهم سنجش انطباقی کامپیوتری (CAT) که پژوهش‌های محدودی به آن اختصاص یافته، خزانه سؤال است. ویژگی‌های جذاب سنجش انطباقی کامپیوتری (CAT) در صورتی تحقق می‌یابند که سؤال‌های موجود در خزانه‌ای که برای اجرا به کار می‌روند مناسب باشند (رکیس، ۲۰۱۰). به طور مسلم افزایش کیفیت خزانه سؤال، نحوهٔ 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

1.Computerized Adaptive Test
2.Reckase
3.Hierarchically
4.Glas
5.Item response theory model
6.Calibrated Item pool
7.Wainer, Dorans, Eignor, Flaugher, Green , Mislevy, Steinberg& Thissen
8.High.Stake
9.Content balancing
10.Cheng&Chang
11.Exposure Control
12.Sympson&Hetter

(فلاگر، به نقل از واینر و همکاران، ۲۰۰۰). بنابراین، سنجش انطباقی کامپیوتری، به خزانه سؤالی نیاز دارد که خوب طراحی شده و شامل تعداد مناسبی سؤال باشد، به طوری که آزمون‌های مجزایی برای مطابقت با سطوح توانایی آزمودنی‌ها فراهم سازد. هر خزانه سؤال بهینه همچنین باید شامل سؤال‌هایی با تعادل محتوایی مناسب باشد، به گونه‌ای که به استفاده بهینه از سؤال منجر شود و هزینه ایجاد سؤال را کاهش دهد (گو[1] و رکیس، ۲۰۰۷). هنگامی خزانه سؤال بهینه‌ای خواهیم داشت که هر زمان الگوریتم انتخاب سؤال سنجش انطباقی کامپیوتری (CAT) سؤالی را برای اجرا جستجو کند، دقیقاً همان سؤالی که مطلوب و مورد نظر ما است، در خزانه سؤال موجود باشد (رکیس، ۲۰۱۰). برخلاف آزمون‌های سنتی مداد-کاغذی، که انتخاب سؤال‌ها به نوعی است که بهترین سنجش را برای آزمودنی‌هایی با توانایی متوسط فراهم می‌کند، سنجش انطباقی می‌تواند دامنه گسترده‌ای از توانایی را پوشش دهد. از این‌رو، به سؤال‌هایی با کیفیت بالا برای دامنه گسترده‌ای از توانایی نیاز است. به همین دلیل خزانه‌های سؤال، در سنجش انطباقی، باید به پیش‌فرض‌های مدل روان‌سنجی که زیربنای مدرج‌سازی، اجرا و نمره‌گذاری است، توجه کنند. بنابر این، تلاشی که برای نوشتن خزانه سؤال‌های سنجش انطباقی لازم است، بسیار بیشتر از آزمون‌های مداد-کاغذی است (میلمن و آرتر[2]، ۱۹۸٤). خزانه سؤال بهینه باید بر اساس مؤلفه‌های دیگر سنجش انطباقی کامپیوتری (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 سنجش انطباقی کامپیوتری (CAT) که توسط رکیس (۱۹۸۹) تعیین شده است، به طور هم‌زمان الزامی است (برگستروم و لانز[3]، ۱۹۹۹). در انتخاب سؤال برای سنجش انطباقی کامپیوتری (CAT) چهار هدف متضاد، وجود دارد. اول اینکه، انتخاب سؤال باید دقت اندازه‌گیری را با انتخاب سؤالی که آگاهی و یا دقت پسین[4] سطح توانایی برآورد شده آزمودنی را بیشینه می‌کند، به حداکثر برساند. دوم، انتخاب سؤال باید با محدود کردن میزان مواجهه سؤال‌ها، از امنیت خزانه سؤال محافظت کند. سوم، انتخاب سؤال باید این اطمینان را ایجاد کند که

1.Gu
2.Millman &Arter
3.Bergstrom& Lunz
4.Posterior precision

آزمودنی‌ها، آزمونی با تعادل محتوایی مناسبی را دریافت کرده‌اند (پارشال، دیوی و نرینگ[1]، ۱۹۹۸). هدف چهارمی که به بهینه‌شدن بیشتر خزانه سؤال کمک می‌کند، این است که انتخاب سؤال باید استفاده از سؤال را به گونه‌ای بیشینه کند، که همه سؤال‌های خزانه استفاده شوند (استوکینگ و سوآن‌سون[2]، ۱۹۹۸). مسائل مربوط به انتخاب سؤال شبیه بادکنکی است که وقتی یک طرف آن را فشار می‌دهیم طرف دیگر آن متورم می‌شود، یعنی، زمانی که به یک جنبه آن توجه می‌شود، از سایر جنبه‌های آن غافل می‌شویم (استوکینگ و لوئیس[3]، ۲۰۰۰).

طراحی خزانه سؤال بر دو رویکرد عمده استوار است. رویکرد اول، توسط ولدکمپ[4] و وندرلیندن (۲۰۰۰) ایجاد شده است. این رویکرد از روش برنامه‌نویسی ریاضی[5] برای طراحی خزانه سؤال استفاده می‌کند. در این رویکرد فرض می‌شود که مجموعه بزرگی از سؤال‌ها که "خزانه اصلی[6]" نامیده می‌شود، از قبل وجود دارد و تنها باید خزانه‌های قابل استفاده، از آن انتخاب شود (بلو و آرمسترونگ[7]، ۲۰۰۹؛ وندرلیندن، آریل[8] و ولدکامپ، ۲۰۰۶). در این رویکرد از "آزمون سایه[9]" برای طراحی خزانه سؤال استفاده می‌شود، و از ویژگی‌های خزانه سؤال موجود به عنوان نقطه شروع استفاده می‌شود (ولدکامپ و وندرلیندن، ۲۰۰۰). به عبارت دیگر، سنجش انطباقی کامپیوتری (CAT) با رویکرد آزمون سایه اجرا می‌شود و آزمون با برنامه‌نویسی عدد صحیح خطی دو ارزشی[10] یا برنامه‌نویسی صفر و یکی سرهم می‌شود (وندرلیندن، ریس[11]، ۱۹۹۸). برخی مطالعات مبتنی بر این رویکرد به هدف طراحی خزانه سؤال با برنامه‌نویسی اعداد صحیح

1.Parshall, Davey&Nering
2.Stocking &Swanson
3.Lewis
4.Veldkamp
5.Mathematical programming
6.Master pool
7.Belov &Armstrong
8.Ariel
9.Shadow a test (SAT)

روش آزمون سایه، روش پشتیبانی مفیدی برای سرهم کردن آزمون‌های انطباقی است. ایده زیربنایی این روش، حل یک مسئله بزرگ به صورت یک توالی از مسائل هم‌زمان کوچک‌تر است. این رویکرد بر اساس این پیش‌فرض شکل گرفته است که، اگر بخواهیم از یک خزانه بزرگ مجموعه‌ای آزمون سرهم کنیم، ابتدا تعداد آزمون‌هایی که باید سرهم شود مشخص می‌شود.

10.Binary linear integer programming
11.Reese

رسیده‌اند (آریل، ولدکمپ و وندرلیندن، ۲۰۰٤). ولدکمپ و وندرلیندن (۱۹۹۹)، پنج گام برای طراحی الگوی بهینه خزانه سؤال سنجش انطباقی کامپیوتری (CAT) با روش برنامه‌ریزی ریاضی، توصیف کردند[۱]. همچنین، وندرلیندن (۲۰۰۵b) توانست با استفاده از این روش، ویژگی‌های بهینه خزانه سؤال را شبیه‌سازی کند. مزیت این روش این است که طراح را قادر می‌سازد تا ویژگی‌های پیچیده آزمون را مدل‌یابی کند. یعنی، ابتدا ویژگی‌های سؤال‌ها را تعریف و آنها را به عدد تبدیل کند، سپس نرم‌افزار ویژه‌ای برای شبیه‌سازی خزانه سؤال بهینه تعبیه کند. با این وجود، خزانه سؤال طراحی شده با روش برنامه‌ریزی ریاضی، به طور گسترده‌ای در انتخاب سؤال، به روش "آزمون سایه" وابسته است و به دانش زیادی در مورد نرم‌افزار بهینه‌سازی ویژه نیاز دارد. همچنین، بسته به روشی که ویژگی[۲] سؤال بخش‌بندی می‌شود، فضای طراحی می‌تواند بسیار بزرگ شود و فرآیند شبیه‌سازی از لحاظ محاسباتی دشوار شود (گو و رکیس، ۲۰۰۷). یکی از محدودیت‌های بالقوهٔ این رویکرد آن است که برای به‌دست آوردن راه‌حل بهینه، به نرم افزارهای جبر خطی از قبیل "CPLEX" و"LINDO" نیاز دارد، که کاربرد این روش را کمی دشوار می‌کند و ممکن است، کدها و معادله‌های آن برای اکثریت کاربران در دسترس نباشد. در این صورت اگر برنامه نیازمند اصلاح و یا تغییر باشد، کنترلی بر آن نخواهد داشت و چه بسا این احتمال وجود دارد که همیشه راه حل قابل اجرا و عملی[۳] در دسترس نباشد (چانگ[٤]، ۲۰۰۷؛ روبین[٥] و همکاران، ۲۰۰۵). محدودیت دیگر این رویکرد این است که سؤال‌ها از قبل در خزانه موجود است و از روی آنها خزانه‌ای کوچک‌تر سرهم می‌شود (گو و رکیس، ۲۰۰۷). در این رویکرد از ویژگی‌های خزانهٔ سؤال موجود به عنوان نقطه شروع استفاده می‌شود (رکیس، ۲۰۱۰).

رویکرد دوم طراحی خزانه سؤال روش اکتشافی رکیس (۲۰۰۳) است، که برای برطرف کردن محدودیت‌های رویکرد برنامه‌نویسی ریاضی، ایجاد شده است. این روش که بر اساس روش مونت‌کارلو[٦]، ویژگی‌های یک خزانه سؤال بهینه را تعیین می‌کند (گو و رکیس، ۲۰۰۷)، برخلاف روش برنامه‌نویسی ریاضی، بسیار سر راست است و در

۱. برای جزئیات بیشتر در مورد این رویکرد به وندرلیندن (۲۰۰۵a) مراجعه شود.

2. attributes
3.Feasible
4.Chang
5.Robin
6.Mont Carlo

مطالعات گوناگون طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری (CAT) استفاده شده است (رکیس، ۲۰۰۳ و ۲۰۰۹؛ رکیس و هی، ۲۰۰٤، ۲۰۰۵، ۲۰۰۸، ۲۰۰۹a، ۲۰۰۹b؛ گو، ۲۰۰۷). در این رویکرد، استفاده از برنامه‌ریزی اعداد صحیح کنار گذاشته شده است و در آن فرض نمی‌شود که سؤال‌ها از قبل وجود دارد. در عوض، در این رویکرد سؤال‌ها برحسب پارامترهای "IRT" به گونه‌ای شبیه‌سازی می‌شوند، که با برآوردهای اخیر توانایی مطابقت داشته باشند و میزان آگاهی سؤال بهینه‌ای را ایجاد کنند. در این روش ، ابتدا خزانه سؤال هدف بر اساس ویژگی‌های غیرآماری از قبیل محتوا، به خزانه‌های کوچک‌تری تقسیم می‌شود، سپس فرآیند سنجش انطباقی کامپیوتری (CAT) به طوری شبیه‌سازی می‌شود، که خزانه‌های سؤال کوچک‌تر به طور هم‌زمان ساخته شوند. شبیه‌سازی با یک آزمودنی که به طور تصادفی از توزیع مورد انتظار جامعه فرضی، در دامنه مشخص شده انتخاب می‌شود، آغاز و سنجش انطباقی کامپیوتری (CAT) برای او اجرا می‌شود. هر سؤال به نحوی شبیه‌سازی می‌شود که براساس برآورد جدید توانایی آزمودنی، سؤال بهینه‌ای باشد. فرایند مشابهی برای آزمودنی بعدی نیز تکرار می‌شود، سپس، به همین ترتیب، این فرایند برای کل نمونه مورد نظر ادامه می‌یابد و سؤال‌ها برای نمونه بزرگی از آزمودنی‌ها شبیه‌سازی و به خزانه سؤال اضافه می شود. بدین ترتیب براساس این روش که به آن روش "bin-and-union" نیز گفته می‌شود، خزانه سؤال بهینه ساخته می‌شود (رکیس، ۲۰۰۳، ۲۰۰۹؛ رکیس و هی، ۲۰۰٤، ۲۰۰۹a). برخلاف مسئله سرهم کردن[1] خزانه سؤال در رویکرد اول که در آن یک خزانه سؤال از یک خزانه بزرگ[2] در دسترس بر طبق ویژگی‌های مطلوب سرهم می‌شود (وندرلیندن، آریل و ولدکمپ، ۲۰۰٦؛ وندرلیندن، ۲۰۰۰a، ۲۰۰۰b، ۲۰۰۵b)، در مسئله طراحی خزانه سؤال در رویکرد دوم، فرض بر این است که هیچ سؤال واقعی در دسترس نیست. رویکرد دوم از این‌رو می‌تواند مفید باشد که در زمان شروع به طراحی یک خزانه سؤال هیچ سؤالی در دسترس نیست (هی، رکیس، ۲۰۱۰). کاربرد این روش در مدل یک پارامتری در تحقیقات گسترده‌ای توسط رکیس (۲۰۰۳)؛ رکیس و هی (۲۰۰٤ و۲۰۰۵) با موفقیت‌های زیادی روبه‌رو بوده است. امّا، تعمیم روش رکیس به مدل‌های دو و سه پارامتری با پیچیدگی‌هایی همراه است. از جمله اینکه برای کاربرد این رویکرد در مدل دو و سه پارامتری لازم است علاوه بر

1.Assemble
2.Master

پارامتر دشواری سؤال، هم‌زمان به دو پارامتر ضریب تشخیص و ضریب حدس نیز توجه شود (گو، رکیس، ۲۰۰۷؛ رکیس، ۲۰۱۰). تعمیم این رویکرد به مدل‌های دو و سه پارامتری در دو پژوهش گو (۲۰۰۷) و هی و رکیس (۲۰۱۱) مشاهده شده است. در رویکرد اکتشافی، امکان کنترل مواجهه بیش از حد سؤال وجود دارد، امّا تا به حال روشی جامع در مورد کنترل مواجهه سؤال در این رویکرد صورت نگرفته است. یکی از پیشنهادهای رکیس در آخرین پژوهشی که از وی منتشر شده است این بود که رویکرد اکتشافی می‌تواند زمانی که کنترل مواجهه سؤال و سیستم امنیتی آزمون اهمیت زیادی دارد نیز در شبیه‌سازی وارد شود (رکیس، ۲۰۱۰؛ هی و رکیس، ۲۰۱۱). طراحی خزانه‌های سؤال بهینه بر اساس رویکرد اکتشافی رکیس، زمانی که عامل کنترل مواجهه، وارد می‌شود کار بسیار دشواری است، زیرا زمانی‌که سؤال‌ها شبیه‌سازی می‌شوند، فرضی هستند و از خزانه‌ای نامتناهی ایجاد می‌شوند. یکی از روش‌های مشهوری که برای کنترل مواجهه وجود دارد، روش سیمپسون–هتر است که وارد کردن آن در طراحی خزانه‌های سؤال بهینه، اهمیت زیادی دارد. استفاده از این روش کنترل مواجهه در رویکرد اکتشافی با چالش‌هایی رو به رو بوده است. از این‌رو، در پژوهش‌های مرتبط با این رویکرد از آن استفاده نشده است (رکیس، ۲۰۱۰؛ هی و رکیس، ۲۰۱۱) و در موارد کاربرد کنترل مواجهه، جدول آگاهی ایجاد شده است (گو و رکیس، ۲۰۰۷). در این پژوهش روش سیمپسون–هتر با استفاده از جداول آگاهی و با ذخیره‌سازی سؤال‌ها شبیه‌سازی شد و اجرای مجدد این خزانهٔ بهینه برای نمونه‌ای از همان آزمودنی‌های جمعیت هدف، در رویکرد اکتشافی رکیس وارد شده است.

هدف اصلی پژوهش حاضر، طراحی خزانه‌های سؤال بهینه‌ای است که بر پایه رویکرد اکتشافی رکیس (۲۰۰۳) طراحی و بر اساس مدل سه پارامتری لُجستیک مدرج شده‌اند و در آنها از روش سیمپسون– هتر برای کنترل مواجهه بیش از حد سؤال، استفاده شده است. هدف فرعی این پژوهش، ساخت خزانهٔ سؤال عملیاتی برای سنجش مهارت ریاضی، به عنوان ملاکی برای ارزیابی خزانه‌های سؤال بهینه شبیه‌سازی شده بود. سؤال‌های پژوهش عبارتند از اینکه:

۱)آیا روش شبیه‌سازی رکیس (۲۰۰۳) قابل تعمیم به مدل سه پارامتری لُجستیک است؟

۲)عملکرد خزانه سؤال بهینه برای سنجش انطباقی کامپیوتری (CAT) زمانی که در الگوریتم انتخاب سؤال، مواجههٔ بیش از حد سؤال کنترل نمی‌شود در مقابل زمانی که مواجههٔ بیش از حد کنترل می‌شود، چگونه است؟

۳)آیا برحسب ملاک‌های ارزیابی[1] خزانه‌های سؤال بهینه‌ای که بر اساس شبیه‌سازی مونت‌کارلو طراحی می‌شوند بهتر از خزانه سؤال عملیاتی واقعی است؟

شرح مفهوم "بین" (bin)

در مدل سه پارامتری لُجستیک، مقدار بیشینه آگاهی یک سؤال از طریق سه پارامتر تعیین می‌شود. یک سؤال با ضریب تشخیص (a) بالا، نسبت به سؤالی با ضریب تشخیص پایین آگاهی بیشتری دارد. با این وجود، چانگ و ییِنگ[2] (۱۹۹۹)، نشان دادند که، زمانی‌که برآورد توانایی (θ) از توانایی واقعی آزمودنی (θ) فاصله دارد، میزان آگاهی کمتری در سطح (θ) ایجاد می‌شود. یک سؤال با پارامتر (c) کوچک‌تر، در سطح (θ) بیشینه‌اش آگاهی بیشتری ایجاد می‌کند. امّا معمولاً در خزانه سؤالی که به خوبی طراحی شده باشد، تغییرپذیری پارامترهای (c) سؤال‌ها به اندازه‌ای کم است. که، عامل (c) تأثیر کمی بر میزان آگاهی به‌دست آمده از سؤال‌ها دارد. بنابراین پارامترهای a و b دو عامل اولیهٔ تأثیرگذار بر میزان آگاهی هر سؤال هستند. سؤال‌هایی با توابع آگاهی مشابه، پارامترهای a و b مشابهی دارند. این قاعده به ایجاد مفهوم "بین" (bin) در شبیه‌سازی خزانه سؤال منتهی می‌شود. مفهوم "بین" (bin) در مدل یک پارامتری، به عنوان فواصلی از مقادیر پارامتر b که در آن فواصل، سؤال‌ها در طول دامنه سطوح (θ) مقادیر آگاهی مشابهی ایجاد می‌کنند. ولی در مدل سه پارامتری، محدوده "بین" (bin) بر اساس پارامتر a و b تعیین می‌شود. در این مدل "بین"ها از یک طرح بخش‌بندی شده مشبک بر اساس مقادیر a و b تشکیل می‌شوند (نمودار ۱). هر خانه مشبک که دامنه‌ای از مقادیر پارامترهای a و b را در برمی‌گیرد، با "ab-bin" نشان داده می‌شود. کل کناری هر ردیف با "a-bin" و کل کناری هر ستون با "b-bin" مشخص می‌شود. سؤال‌هایی که درون هر خانه قرار می‌گیرند، با توجه به پارامترهای a و b مربوط به سؤال در کل دامنه توانایی، آگاهی مشابهی دارند و بیشینه آگاهی آنها در اطراف مرکز توانایی داخل خانه‌ای "بین" (bin) است که سؤال‌ها در آن قرار می‌گیرند.

1.evaluation criteria
2.Chang&Ying

در مدل دو و سه پارامتری، محدوده‌های "b-bin" با تقسیم کردن محور θ (یا محور پارامتر b)، به فواصل برابر به‌دست می‌آید، در صورتی‌که پهنای محدوده‌های "a-bin" به‌گونه متفاوتی تقسیم‌بندی می‌شود. دلیل این قضیه این است که مقدار بیشینه آگاهی هر سؤال با فرض اینکه پارامتر (c) ثابت است، متناسب با تابع درجه دوم پارامترهای a است (لرد، ۱۹۸۰). بنابراین، رابطهٔ بین بیشینه آگاهی که هر سؤال می‌تواند فراهم کند و پارامتر a، بر اساس معادله (۱) محاسبه می‌شود (لرد، ۱۹۸۰؛ گو و رکیس، ۲۰۰۷):

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1 - 20c_i - 8c_i^2 + +(1 + 8c_i)^{3/2}] \qquad (۱)$$

معادله ۲ تغییرات بین تابع بیشینه آگاهی (ΔM) برای سؤال‌هایی با پارامترهای a متفاوت را بهتر نشان می‌دهد:

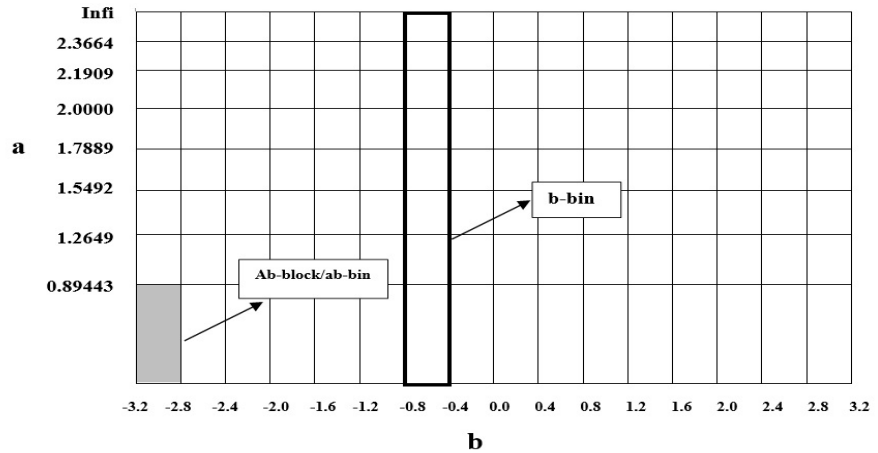
$$\Delta M = \frac{D^2[1-20c-8c^2+(1+8c)^{3/2}]}{8(1-c)^2}\Delta a^2 \qquad (۲)$$

از آنجا که میانگین پارامتر c سؤال‌های موجود در خزانه عملیاتی، برابر با ۰/۲۵ است، قسمت ثابت اول فرمول برابر با ۰/٤٤۷ (یعنی، $\Delta M = 0.447\Delta a^2$) خواهد بود. به این ترتیب، می‌توان حدود "a-bin" را که در آن تغییرات پارامترهای (a) باعث تغییر اندکی در میزان آگاهی می‌شود محاسبه کرد. از آنجا که پارامتر (a) به‌طور قراردادی از صفر بیشتر است، می‌توان با توجه به مقدار مورد انتظار تغییر آگاهی، حدود پارامتر (a) را گام به گام محاسبه کرد. نمودار (۱) یک طرح "بین" (bin) که براساس یک تغییر ۰/٤ در میزان آگاهی سؤال، زمانی‌که دامنه (b) برابر با ۰/٤ و (c) برابر با ۰/۲۵ است را نشان می‌دهد. خزانه سؤالی که به‌طور بهینه طراحی شده است، باید در هر "b-bin" دارای تعداد کافی سؤال باشد و این اطمینان را ایجاد کند که سؤال‌هایی با پارامتر شیب (a) به اندازه کافی بالا، در دسترس هستند (برای توصیف بیشتر در خصوص این مفهوم به گو و رِکیس، ۲۰۰۷ مراجعه شود).

روش تحقیق

در این بخش ابتدا خزانه سؤال عملیاتی و سپس روش‌های ایجاد خزانه‌های سؤال بهینه شبیه‌سازی شده، توصیف شده‌اند. آنگاه روش کنترل نرخ مواجهه سؤال، مراحل شبیه‌سازی خزانه‌های سؤال بهینه و روش‌های ارزیابی آنها مورد بحث قرار گرفته است

و در نهایت متغیرهای مستقل پژوهش ذکر شده‌اند.



نمودار (۱) نمونه‌ای از ab-bin یا ab-block

۱) خزانه سؤال عملیاتی

خزانه سؤال عملیاتی مربوط به محتوای ریاضی حسابان–دیفرانسیل است. سؤال‌های موجود در این خزانه به منظور اندازه‌گیری توانایی حل مسائل حساب دیفرانسیل (این محتوا به‌صورت خلاصه شامل: معادلات گویا، اصم، نامعادله، رابطه، تابع، ماتریس‌ها، دنباله و سری‌ها، روابط مثلثاتی، توابع، حدّ و پیوستگی، مشتق، انتگرال است و برای دانش‌آموزانی که در مقطع پیش‌دانشگاهی تحصیل می‌کردند و برای آزمون سراسری گروه ریاضی و فنی در سال ۱۳۹۳ آماده می‌شدند، طراحی شده است و به عنوان مبنایی برای محک و ارزیابی خزانه‌های سؤال شبیه‌سازی شده این مطالعه، به کار رفته است. خزانه سؤال عملیاتی شامل ۸٤۰ سؤال حساب دیفرانسیل بود که توسط هشت طراح سؤال ساخته شد و توسط شش طراح دیگر از نظر موضوعی و تخصصی ارزیابی شد. سپس سه روان‌سنج که در درس ریاضی نیز تخصص داشتند بررسی شد و مشکلات محتوایی و ایرادهای احتمالی گزینه‌ها رفع شد. در پایان، سؤال‌ها در ۱٤ مرحله آزمون متوالی– از ابتدای نیم سال اول سال تحصیلی ۹۳–۹۲ تا اوسط فروردین ماه سال ۱۳۹۳، به صورت مداد– کاغذی برای پانصد نفر از آزمودنی‌هایی که معرف جامعه اصلی بودند، اجرا شد[1]. داده‌های به‌دست آمده از این نمونه با استفاده از مدل سه پارامتری لُجستیک

۱. البته قابل ذکر است که سؤال‌های خزانه عملیاتی تنها برای نمونه‌ای (پانصد نفری) که برای مدرج کردن سؤال‌ها انتخاب شد، اجرا شد و دیگر افراد جامعه در ۱٤ مرحله کنکور آزمایشی، سؤال‌هایی متفاوت با این پانصد نفر دریافت می‌کردند. این

توسط نرم‌افزار "BILOG" مدرج و سؤال‌هایی که با مدل سه پارامتری لُجستیک برازش داشتند در خزانه ذخیره شدند (دی‌آیالا، ۲۰۰۹). همچنین، مفروضات "IRT" شامل تک بعدی بودن و استقلال موضعی در مورد هریک از سؤال‌ها بررسی شد (همبلتون و همکاران، ۱۹۹۱؛ دی‌آیالا، ۲۰۰۹). این خزانه عملیاتی به شکلی طراحی شد تا ملاک‌های توصیف شدهٔ وندرلیندن (۲۰۰۰a) را داشته باشد[1]. برنامه سنجش انطباقی کامپیوتری (CAT) با زبان "PHP" نوشته شد و از پایگاه داده "MYSQL" برای ذخیره‌سازی سؤال‌ها در خزانه استفاده شد. روش انتخاب سؤال در فرایند اجرای سنجش انطباقی کامپیوتری (CAT) روش بیشینه آگاهی (MI) بوده است. همچنین، برای صرفه‌جویی در زمان محاسبه، از یک جدول آگاهی نیز استفاده شده است. برای به‌دست آوردن برآورد اخیر توانایی هر آزمودنی، قبل از اینکه دو پاسخ صحیح و غلط در الگوی پاسخ او مشاهده شود، از روش میانگین پسین (MAP) (اوون[2]، ۱۹۷۵) استفاده شد، به طوری‌که پیشینهٔ مورد انتظار از توزیع نرمال پیروی می‌کرد. پس از مشاهدهٔ دو پاسخ صحیح و غلط در الگوی پاسخ فرد، برای برآورد توانایی از شیوه بیشینه درست‌نمایی (MLE) استفاده شد. برآورد (θ) بعد از سؤال آخر، به عنوان نمره آزمودنی به حساب آمد. همچنین روش سیمپسون–هتر برای کاهش بیش مواجههٔ سؤال‌هایی که میزان آگاهی بالایی داشتند استفاده شد و نرخ مواجهه هدف برابر با ۱÷۳=۰/۳۳ قرار داده شد. هر آزمون بعد از اجرای بیست سؤال، به اتمام می‌رسید. از آنجا که در این مطالعه تعادل محتوایی به عنوان عاملی امنیتی در ساخت خزانه‌های سؤال بهینه در نظر گرفته نشده بود، به منظور کنترل محتوای سؤال‌های ارائه شده، تنها از یک محتوا (حساب دیفرانسیل) در سنجش انطباقی کامپیوتری (CAT) عملیاتی استفاده شد. توانایی اولیه برای هر فرد روی صفر تنظیم شد و برنامه سنجش انطباقی کامپیوتری (CAT) به نحوی برنامه‌ریزی

کار به دو دلیل انجام شد، اول برای اینکه اطمینان حاصل شود که تمام پانصد نفری که برای مدرج کردن خزانه به کار می‌روند تا پایان ۱٤ مرحله آزمون ثابت باقی می‌مانند تا توسط محقق مورد پیگیری قرار گیرند و مدرج کردن سؤال‌ها با دقت انجام شود. دوم برای جلوگیری از افشا شدن آزمون برای کل افراد جامعه که نمونه نهایی (۳۵۰ نفری) نیز از بین آنها انتخاب می‌شد و حفظ امنیت سؤال‌های خزانه عملیاتی.

۱. این ملاک‌ها عبارتنداز: الف) خزانه باید به اندازه کافی بزرگ باشد تا این اجازه را به ما بدهد که چندین هزار خرده آزمون همپوش از سؤال‌ها بدست آید ب) سؤال‌ها دامنه کاملی از سطوح دشواری نسبت به جامعه مورد نظر را پوشش دهند ج) خزانه باید شامل ترکیب مناسبی از سؤال‌ها با ضرایب تشخیص بالا و پایین باشد تا هزینه ساخت سؤال را با در نظر گرفتن دقت آزمون کاهش دهد. برای اطلاعات بیشتر نیز می‌توان به وندرلیندن، پاشلی، ۲۰۰۰ مراجعه کرد.

2.Owen

شد که برای همه افراد، سؤال یکسانی که پارامتر دشواری آن صفر (b=0) باشد، اجرا کند. تمام دانش‌آموزان مقطع پیش‌دانشگاهی که خود را برای آزمون سراسری گروه ریاضی سال ۱۳۹۳ آماده می‌کردند و در کنکورهای آزمایشی ١٤ مرحله‌ای نیز شرکت می‌کردند، جامعهٔ این پژوهش محسوب شدند. همچنین، نمونه‌ای ۳۵۰ نفری از دانش‌آموزان این جامعه به صورت تصادفی منظم (از آنجا که فهرست کاملی از اسامی شرکت کنندگان در آزمون‌های آزمایشی مرحله‌ای در دسترس بود، سعی شد اسامی به صورت تصادفی فهرست شوند، از این روش استفاده شد) انتخاب و آزمون در فواصل فروردین ماه تا خرداد ماه ۱۳۹۳ (قبل از آزمون سراسری ۹۳) به صورت برخط و کامپیوتری برای آنها اجرا شد. این نمونه از یک توزیع نرمال توانایی، با میانگین ۰/۱۷- و انحراف استاندارد ۰/۹۵ پیروی می‌کرد.

۲) روش‌های ایجاد سؤال‌های بهینهٔ شبیه‌سازی شده

برای ایجاد سؤال‌های بهینه[1] در رویکرد رکیس باید پارامترهای سؤال شبیه‌سازی شوند. برای اینکه ویژگی‌های سؤال‌های شبیه‌سازی شده از لحاظ کاربردی مفید باشد، باید از اطلاعات پیشین، که از خزانه عملیاتی حاصل می‌شود، استفاده کرد. این اطلاعات شامل، توزیع پارامترهای سؤال‌های عملیاتی، رابطه بین پارامترهای سؤال، اعتبار[2] آزمون و برآوردهای توانایی آزمودنی‌ها است. نتایج تحلیل ٨٤٠ سؤال موجود در خزانه نشان داد که بین پارامترهای (a) و (b) سؤال‌های عملیاتی هیچ نوع همبستگی معنی‌دار آماری (p= ۰/٦٥٨) وجود نداشت. توزیع پارامتر (a) در کل سؤال‌های خزانه عملیاتی نرمال با میانگین ۱/۰۸۹ و انحراف استاندارد ۰/۲۷٤ بود. نرمال بودن این توزیع براساس آزمون کُلموگروف-اسمیرنف[3] تأیید شد (p<۰/۲۸). همچنین، پارامتر (c) با میانگین ۲/٦٥٤ و انحراف استاندارد ١٥/٦٤٩ از توزیع بتا[4] پیروی می‌کرد. برای این‌که الگوی رابطه پارامترهای (a) و (b) در سؤال‌های خزانه عملیاتی به صورت دقیق‌تری مشخص شود، ابتدا همهٔ سؤال‌ها بر اساس مقادیر (b) به سه گروه تقسیم شد (٤- تا ۱/۷۲۰-؛ ۱/۷۲۰- تا ۱/۷۲۰ و ۱/۷۲۰ تا ٤)، سپس همبستگی بین پارامترهای (a) و (b) برای هر سه گروه محاسبه شد. نتایج نشان داد که، تنها در گروه سوم سؤال‌ها (۱/۷۲۰ تا ٤) بین پارامترهای

1.Generation optimal items
2.Reliability
3.Klmogorov–Smirnov
4.Beta

(a) و (b) همبستگی معنی‌داری (r_{ab}=۰/٦٧٨ ،p= ۰/۰۰۱)[1] وجود داشت. بنابراین، در این گروه، یک رگرسیون ساده خطی برابر با $a_i = 1.243 + 0.275b_i + e_i$ ، برای پیش‌بینی a توسط b محاسبه شد[2] که در آن e_i یک عنصر تصادفی با توزیع $N(0, \sigma_e^2)$ است. انحراف استاندارد این توزیع (σ_e) براساس معادله (۳)، که بر ایده مک‌برد و وایس (۱۹۷٦) استوار است، محاسبه شد (r_{ab}=۰/٦٧٨ و S_a=۰/۲۷٤).

$$\sigma_e = S_a\sqrt{(1 - r_{ab}^2)} = 0.274\sqrt{(1 - 0.678^2)} = 0.202 \qquad (۳)$$

براساس ایده‌های گو (۲۰۰۷)، مک‌برید و وایس[3] (۱۹۷٦)، سه روش برای ایجاد خصوصیات بهینه سؤال در مدل سه پارامتری وجود دارد که عبارتند از: روش تصادفی[4] (R)، روش تصادفی آمیخته و پیش‌بینی[5] (MRP)، و روش کمینه آگاهی آزمون[6] (MTI).

الف) روش تصادفی(R)

این روش هنگامی به کار می‌رود که پارامترهای (a) و (b) در آزمون سنجش انطباقی کامپیوتری (CAT) عملیاتی از لحاظ آماری با یکدیگر همبستگی نداشته باشند. به منظور ایجاد یک سؤال بهینه با استفاده از روش "R"، گام‌های زیر دنبال شد:

۱- بر اساس خزانه سؤال عملیاتی پارامترهای a_i و c_i به ترتیب دارای توزیع نرمال$\cong$ a $N(1.089, 0.274)$ و بتا c $\cong Beta(2.654, 15.649)$ در نظر گرفته شدند.

۲- پس از مشخص شدن a_i و c_i، b_i با استفاده از معادله (٤) محاسبه شد. در اینجا $\theta_{max} = (\hat{\theta}_{ij})$ برآورد جدید توانایی است:

$$\theta_{max} = b_i + \frac{1}{Da_i}\ln\left[0.5\left(1 + \sqrt{1 + 8c_i}\right)\right] \Longrightarrow \qquad (٤)$$

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

---

۱. همبستگی در گروه (٤- تا ۱/۷۲۰-) برابر با (۰/۸۱ =p، ۰/۰۵۸=rab)، در گروه (۱/۷۲۰- تا ۱/۷۲۰) برابر با (۰/۵٤ =p، ۰/۰۹۹=rab) است.

۲.در روش پیش‌بینی (P) برای شبیه‌سازی پارامتر a در خزانهٔ بهینه، از رگرسیون ساده پیش‌بینی (a) توسط (b) در خزانه عملیاتی استفاده می‌شود که در قسمت مربوط شرح داده شده است.

3.McBride & Weiss
4.Random Procedure (R)
5.Mixed Random and Prediction Procedure (MRP)
6.Minimum Test Information Procedure (MTI)

ب) روش تصادفی آمیخته و پیش‌بینی (MRP)

روش "MRP"، یک روش آمیخته است. قسمت تصادفی (R) آن در بالا توصیف شد، در قسمت پیش‌بینی (P)، که بر ایده مک‌برید و وایس (۱۹۷٦) استوار است، یک خزانه سؤال "مناسب[1]" با پارامترهای بهینه، از طریق معادله رگرسیون پارامترهای (a_i) براساسپارامترهای (b_i) شبیه‌سازی می‌شود. این روش بر این واقعیت استوار است که پارامترهای (a) و (b) به طور معنی‌داری با یکدیگر همبستگی دارند (چانگ و وندرلیندن، ۲۰۰۳؛ وندرلیندن، اسکرامز و اسچنیپکا[2]، ۱۹۹۹). به عبارت دیگر، واریانس پارامتر (a) با افزایش پارامتر (b)، بیشتر می‌شود. این مطلب بیانگر این است که با استفاده از تبدیلات لگاریتمی، پارامترهای (a) به طور خطی با پارامترهای (b) مرتبط می‌شود (گو و رکیس، ۲۰۰۷). برای مدل‌یابی این روابط، پارامتر (a) برای یک سؤال شبیه‌سازی شده، برابر با تابع رگرسیونی تبدیل لگاریتمی پارامتر $a^{(a')}$ بر پارامتر (b) است (رکیس و هی، ۲۰۰٤). همچنین، $e \cong N(0,\ \sigma^2)$ توزیع نرمال دارد. با اضافه کردن یک عبارت خطا در تابع رگرسیونی، پراکندگی نرمال در پارامترهای a به وجود می‌آید.

$$a'_i = \log(a_i) = B_0 + B_1 b_i + e_i \Longrightarrow \ ai = \exp(a'_i) \qquad (۵)$$

در این پژوهش به منظور ایجاد یک سؤال بهینه با استفاده از رویکرد "MRP"، اگر مقدار(b_i) سؤال، که براساس ($\hat{\theta}$) در هر مرحله از اجرای آزمون تقریب زده شد، پایین‌تر از ۱/۷۲۰، یعنی، جزو گروه‌هایی بود که در آن همبستگی معنی‌داری بین پارامتر (a) و (b) وجود نداشت، از روش "R" استفاده شد. امّا، اگر مقدار(b_i) برابر یا بالاتر از ۱/۷۲۰ بود یعنی جزو گروهی بود که در آن همبستگی معنی‌داری بین پارامتر (a) و (b) وجود داشت، روش‌های زیر به کار رفت:

۱- پارامتر (c_i) براساس توزیع بتا $c \cong Beta(2.654, 15.649)$ ایجاد شد.

۲- پارامتر (a_i) براساس $a_i = 1.243 + 0.275 b_i + e_i$ ایجاد شد که b_i براساس ($\hat{\theta}$) به دست آمده در هر گام انتخاب سؤال تقریب زده شد. همچنین، e_i از توزیع نرمال $N(0,\ 0.202^2)$ پیروی می‌کرد.

۳- دوباره پارامتر (b_i) از طریق معادله (٤) محاسبه شد.

1.perfect
2.Scrams and Schnipka

ج) روش حداقل آگاهی آزمون (MTI)

برای ساخت آزمونی که بتواند میزان آگاهی بیشتری ایجاد کند، به سؤال‌هایی با ضرایب تشخیص بالا نیاز است. چون معمولاً ساخت این نوع سؤال‌ها گران و دشوار است، (مخصوصاً اگر سؤال‌ها آسان باشد)، این دشواری دو چندان می‌شود. روش "MTI" این اطمینان را ایجاد می‌کند که آزمون‌ها، برای برآورد توانایی، دقت کافی دارند، ولی شامل سؤال‌هایی با ضرایب تشخیص بسیار بالا نیستند. در این روش، مقداری آگاهی هدف بر دامنه‌ای از مقیاس (θ) قرار می‌گیرد. هر سؤالی که برای آزمودنی اجرا می‌شود، در مقدار آگاهی هدف آزمون سهیم است (گو و رِکیس، ۲۰۰۷). برای اجرای این رویکرد، گام اول، تعیین آگاهی هدف آزمون است. براساس اطلاعات پیشین در مورد خزانه عملیاتی و توزیع برآوردهای توانایی، حداقل آگاهی هدف آزمون می‌تواند براساس معادله (٦) تعیین شود:

$$S_e = S_o\sqrt{1 - r_{xx'}} \rightarrow I_{\hat{\theta}} = \frac{1}{S_e^2} \qquad (٦)$$

(S_o)انحراف استاندارد برآوردهای توانایی($\hat{\theta}$)؛ (S_e) خطای استاندارد برآورد، $r_{xx'}$ اعتبار آزمون و $I_{\hat{\theta}}$ آگاهی آزمون را نشان می‌دهد. زمانی‌که $I_{\hat{\theta}}$ معلوم شد، می‌توان آگاهی مورد انتظاری که هر سؤال باید فراهم کند را با تقسیم $I_{\hat{\theta}}$ بر طول آزمون به دست آورد. با توجه به این واقعیت که آگاهی واقعی که هر سؤال می‌تواند ایجاد کند، به برآورد جدید توانایی مشروط است، ممکن است آگاهی واقعی کاملاً مطابق با آگاهی مورد انتظار نباشد. بنابراین، آگاهی هدف سؤال باید هر بار پس از اینکه یک سؤال اجرا می‌شود به‌روز شود. معادله (۷) برای به‌روز کردن آگاهی هدف سؤال به کار می‌رود. در این رابطه T، آگاهی آزمون و L، طول آزمون را نشان می‌دهد (گو و رِکیس، ۲۰۰۷؛ هی و رِکیس، ۲۰۱۱):

$$I_i = \frac{T_{target} - T_{admin}}{L_{target} - L_{admin}} \qquad (٧)$$

در این پژوهش، با استفاده از تحلیل داده‌های سنجش انطباقی کامپیوتری (CAT) عملیاتی، حداکثر آگاهی هدف آزمون در دامنه توانایی (۱/٦۲٤- تا ۱/۰۸۸) به دست آمد، که تقریباً برابر با (۲۱/٤) بود. به عبارت دیگر، آگاهی هدف آزمون با توجه به سطوح متفاوت توانایی محاسبه شد. برای آزمودنی‌هایی با توانایی واقعی (۱/٦۲٤- تا ۱/۰۸۸) آگاهی هدف آزمون برابر با ۲۱/٤ (چون $S_e = 0.216$، پس، $I_{\hat{\theta}} = \frac{1}{0.047} = 21.4$)، برای آزمودنی‌هایی با توانایی‌های واقعی (۱/۰۸۸ تا ۲/۵) و (۱/٦۲٤- تا ۲/۵) آگاهی

هدف آزمون برابر با ۱۸/٤ (چون $S_e = 0.233$؛ پس، $I_{\hat{\theta}} = \frac{1}{0.054} = 18.4$) و برای بقیه آزمودنی‌ها، آگاهی هدف آزمون برابر با ۱۵/٤ (چون $S_e = 0.255$؛ پس،$= I_{\hat{\theta}} = \frac{1}{0.065}$ 15.4) به‌دست آمد. پس از آنکه آگاهی آزمون معلوم شد، آگاهی مورد انتظاری که هر سؤال باید ایجاد کند، براساس معادله (۷) به‌دست آمد. در روش "MTI" سؤال‌ها در سه مرحله ایجاد شدند (رکیس و هی، ۲۰۰٤، گو و رکیس، ۲۰۰۷؛ هی و رکیس، ۲۰۱۱):

۱- پارامتر c_i براساس توزیع بتا $c \sim Beta(2.654, 15.649)$ ایجاد شد.

۲- پارامتر a_i براساس معادله (۸) ایجاد شد.

$$a_i = \sqrt{\frac{8(1-c_i)^2 I_i}{D^2[1-20c_i-8c_i^2+(1+8c_i)^{\frac{3}{2}}]}} \quad (۸)$$

معادله (۸) در واقع، از سازماندهی دوباره معادله (۹) به‌دست آمده است. البته M_i می‌تواند به جای I_i در معادله (۸) قرار گیرد:

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{\frac{3}{2}}] \quad (۹)$$

۳- با توجه به اینکه هم پارامتر a_i و هم پارامتر c_i معلوم بودند، پارامتر b_i با استفاده از معادله (٤) محاسبه شد.

۳) کنترل نرخ مواجهه سؤال

روش کنترل مواجهه سیمپسون–هتر[1] (S-H) یکی از رایج‌ترین شیوه‌های انتخاب مشروط سؤال است. در این روش به هر سؤال یک عدد در دامنهٔ "صفر و یک" به عنوان پارامتر کنترل مواجهه اختصاص داده می‌شود. این عدد بر اساس فراوانی انتخاب سؤال، که مبتنی بر شبیه‌سازی چرخشی[2] سنجش انطباقی کامپیوتری (CAT) است، تعیین می‌شود. به سؤال‌های دارای فراوانی‌های اجرایی زیاد، پارامترهای کنترل مواجهه کوچک‌تری اختصاص داده می‌شود. در طول اجرای آزمون، پارامتر کنترل مواجههٔ سؤال انتخاب شده با عدد یکنواخت تصادفی که دامنه آن نیز بین صفر تا یک است، مقایسه می‌شود. اگر پارامتر کنترل مواجهه بزرگ‌تر از عدد تصادفی باشد، سؤال اجرا می‌شود و اگر کوچک‌تر باشد، سؤال به خزانه سؤال بازگردانده می‌شود. به همین صورت، فرآیند یکسانی برای بهترین سؤال بعدی صورت می‌گیرد. پارامتر کنترل مواجهه مشابه آستانه[3]

1.Sympson-Hetter procedure
2.Iteration
3.Threshold

است. با کنترل آستانه در روش سیمپسون–هتر اجرای سؤال‌هایی که به طور فراوانی در سنجش انطباقی کامپیوتری (CAT) استفاده می‌شوند محدود می‌شود و نرخ بیشینه مواجهه سؤال برای سؤال‌هایی که کمتر مورد استفاده قرار می‌گیرند را تضمین می‌کند. معمولاً پارامترهای کنترل مواجهه در روش سیمپسون–هتر با مجموعه‌ای از شبیه‌سازی‌های چرخشی اجراهای واقعی سنجش انطباقی کامپیوتری (CAT) تعیین می‌شود. به عبارت دیگر، این پارامتر نسبت نرخ مواجهه هدف برای احتمال انتخاب سؤال در آزمون است. در این شیوه فرض می‌شود که $P(S_i)$ احتمال انتخاب سؤال (i) برای یک آزمودنی است که به طور تصادفی انتخاب شده و (A_i) اجرای آن سؤال را نشان می‌دهد. نرخ مواجهه سؤال (i) می‌تواند به صورت $P(A_i)$ تفسیر شود: یعنی احتمال اجرای سؤال (i) برای آزمودنی که به‌طور تصادفی نمونه‌گیری شده است. روش سیمپسون–هتر سؤال‌های اجرا شده را از سؤال‌هایی که انتخاب می‌شوند، براساس رابطهٔ $(P(A_i) = P(A_i|S_i)P(S_i))$ جدا می‌کند و $P(A_i)$ را براساس $P(A_i|S_i)$، یعنی نسبت انتخاب‌هایی که به اجرا منجر می‌شود، کنترل می‌کند. برای هر نرخ مواجهه معین $r_i > 0$، می‌توان $P(A_i) \leq r_i$ را براساس $P(A_i|S_i) \leq \frac{r_i}{P(S_i)}$ به‌دست آورد. اگر $P(S_i)$ معلوم باشد، یا بتوان آن را تقریب زد، این روش می‌تواند به آسانی با ایجاد یک متغیر تصادفی با توزیع U=(0, 1) اجرا شود (سیمپسون و هتر، ۱۹۸۵).

٤) مراحل شبیه‌سازی خزانه‌های سؤال بهینه

برنامه‌های شبیه‌سازی شده با استفاده از نسخهٔ اصلی برنامه MATLAB (MathWorks, 2014)، به منظور شبیه‌سازی الگوی خزانه سؤال و ارزیابی خزانه‌های سؤال شبیه‌سازی شده و عملیاتی ایجاد شد. شبیه‌سازی الگوی خزانه سؤال بهینه در گام‌های زیر انجام شد:

گام اول- مدل‌یابی شیوه‌های سنجش انطباقی کامپیوتری (CAT): از آنجا که یکی از اهداف این پژوهش ساخت خزانه سؤال بهینه برای سنجش مهارت ریاضی بود، در شبیه‌سازی خزانه‌های بهینه نیز تمام ویژگی‌های روان‌سنجی آزمون عملیاتی، به دقت وارد شد. طول آزمون‌ها مانند آزمون‌های عملیاتی، ثابت (۲۰ سؤال) قرار داده شد. آزمون‌ها به صورت تک محتوایی و روش انتخاب سؤال‌ها روش "MI" به همراه جدول آگاهی در نظر گرفته شد. به منظور برآورد توانایی آزمودنی در طول اجرای آزمون، پیش از اینکه آزمودنی در الگوی پاسخ خود حداقل دو پاسخ صحیح و غلط ایجاد کند، از

روش برآورد اوون (۱۹۷۵) برای میانگین پسین[۱] استفاده شد. پس از ایجاد حداقل دو پاسخ صحیح و غلط در الگوی پاسخ، از روش بیشینه درست‌نمایی استفاده شد. لازم به ذکر است که پهنای متعدد "b–bin" در تمام خزانه‌ها برابر با ۰/۲ و میزان تغییر آگاهی در پارامتر a برابر با ۰/٤ (یعنی، $\Delta a2=2\Delta IMaximum = 0.4$:a-bin) قرار داده شد.

گام دوم–ایجاد جامعه آزمون دهندگان: از آنجا که، خزانه سؤال عملیاتی برای آزمودنی‌های دارای توزیع توانایی نرمال طراحی شد، در شبیه‌سازی خزانه سؤال بهینه نیز از توزیع نرمال پیروی شد. دو توزیع حجم نمونه در شبیه‌سازی خزانه‌های سؤال بهینه به کار رفت. به این معنی که، خزانه‌های سؤال بهینه با یک نمونه شبیه‌سازی و با نمونه‌ای دیگر ارزیابی شدند. الف) تعداد شش هزار توانایی (θ) از توزیع $N(0,1)$ به طور تصادفی انتخاب شد و به عنوان توانایی واقعی آزمودنی‌ها وارد تحلیل شد. این نمونه به منظور تعیین ویژگی‌های خزانه سؤال بهینه به کار رفت ب) پیوستار توانایی در دامنه (٤– تا ٤+) با فواصل (۰/۱۲۵ به ٦٥) طبقه تقسیم و در هر یک از این سطوح پانصد آزمودنی قرار گرفت (در کل ۳۲۵۰۰ آزمودنی). این نمونه به منظور ارزیابی عملکرد کلی شبیه‌سازی‌ها و محاسبه آماره‌های مشروط توانایی مورد استفاده قرار گرفت.

گام سوم–ایجاد پارامترهای سؤال: برای هر آزمون سنجش انطباقی کامپیوتری (CAT)، سؤال اول طوری طراحی شد که برای سطح توانایی صفر بهینه باشد. بعد از هر پاسخ، سؤال‌های بهینه‌ای برای برآورد اخیر توانایی تولید شد. البته فرض بر این بود که سؤال‌ها بر پایه مدل سه پارامتری لُجستیک مدرج شده‌اند. بنابراین، پارامترهای (a، b و c) با سه روش (R، MRP و MTI) تولید شدند. در هر سه روش، پارامتر (c) بر اساس توزیع بتا[۲] تولید شد. پارامترهای (a) بسته به برآورد اخیر توانایی و روش ایجاد پارامتر (R، MRP و MTI) ایجاد شدند و پارامترهای (b) طوری تولید شدند که سؤال حداکثر میزان آگاهی در برآورد توانایی جدید را ایجاد کند.

---

۱. توزیع پیشین توانایی با میانگین صفر و انحراف استاندارد یک در نظر گرفته شد.

۲. میانگین و واریانس پارامتر (c) برابر با میانگین و واریانس پارامتر (c) در خزانه عملیاتی بود. بهترین توزیعی که با این پارامتر برازش داشت توزیع بتا بود.

گام چهارم-ایجاد داده‌های پاسخ: پاسخ آزمودنی‌ها به دنبال هر سؤالی که بر طبق مدل سه پارامتری لُجستیک ایجاد شد، تولید شد. در مدل سه پارامتری لُجستیک، احتمال اینکه آزمودنی (j) به سؤال (i) پاسخ صحیح دهد، به صورت زیر محاسبه می‌شود:

$$P_i(\theta_j) \equiv c_i + (1 - c_i)(1 + \exp[-1.7a_i(\theta_j - b_i)])^{-1} \quad (۱۰)$$

$P_i(\theta_j)$، احتمال اینکه فرد j (J،...، ۲، ۱=j) با پارامتر θ، به سؤال i (I،...، ۲، ۱=i) پاسخ صحیح دهد را نشان می‌دهد و در آن a_i پارامتر شیب، b_i پارامتر دشواری و c_i پارامتر حدس سؤال i است. از آنجا که توانایی واقعی آزمودنی در شبیه‌سازی معلوم بود، بعد از اجرای هر سؤال برای آزمودنی، $P_i(\theta_j)$ محاسبه شد. پس از آن، عدد تصادفی m_{ij} از توزیع یکنواخت U(0, 1) استخراج و با $P_i(\theta_j)$ مقایسه شد. اگر m_{ij} برابر یا کمتر از $P_i(\theta_j)$ بود، پاسخ برابر با یک، در غیر این‌صورت برابر با صفر در نظر گرفته می‌شد.

گام پنجم-تعدیل پس از شبیه‌سازی: برای هر ترکیبی از روش‌ها و متغیرهای مستقل پژوهش (که در ادامه توصیف شده‌اند) ده تکرار صورت گرفت، تا جایی که برآورد نسبتاً ثابتی از خزانه سؤال بهینه به‌دست آمد. از ده تکرار الگوها و تعداد مواجهه سؤال، قبل انجام تعدیل پس از شبیه‌سازی، میانگین گرفته شد.

۵) روش‌های ارزیابی خزانه‌های سؤال شبیه‌سازی شده

در این پژوهش عملکرد خزانه‌های سؤال بهینه شبیه‌سازی شده با خزانه‌های سؤال عملیاتی، بر اساس مجموعه‌ای از ملاک‌های تجربی ارزیابی و مقایسه شد. توزیع نمونه‌گیری دوم مطالعه (یعنی، توزیع ۳۲۵۰۰ نفری) در ارزیابی خزانه سؤال استفاده شد. ملاک‌های ارزیابی خزانه سؤال به شرح زیر است (چانگ و ییِنگ[1]، ۱۹۹۹؛ رِکیس و هی، ۲۰۰۵):

۱- **آگاهی شرطی آزمون[2]:** آگاهی آزمون در هر یک از سطوح توانایی، برابر است با مجموع کل آگاهی هر یک از سؤال‌های آزمون در آن سطح توانایی. چون آزمون‌های این مطالعه دارای طول ثابت بیست (I،...، ۲، ۱=i) بودند، آگاهی آزمون، به عنوان شاخص کارایی[3] آزمون در نقاط مختلف توانایی در نظر گرفته شد. هرچه میزان آگاهی آزمون در یک سطح توانایی بیشتر باشد، کارایی آزمون در آن سطح نسبت به سایر سطوح توانایی نیز بیشتر است. برای سطح توانایی چندمین j (۶۵،

1.Chang&Ying
2.Conditional test information
3.Efficacy index

... ،۲ ،۱=j) آگاهی آزمون در زیر ارائه شده است. نمادهای استفاده شده در این رابطه قبلا تعریف شده‌اند:

$$I(\theta_j) = \sum_{i=1}^{I} a_{ij}^2 \frac{(P_{ij}-c_i)^2.q_{ij}}{(1-c_i)^2.p_{ij}} \quad (۱۱)$$

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

$$SEM(\theta_j) = \sqrt{\frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \bar{\theta}_{ij})^2} \quad (۱۲)$$

اگر θ_j، توانایی j ام (٦٥، ... ،۲ ،۱=j) 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 ٥۰۰=N_i، تعداد کل تکرارهای CAT اجرا شده در θ_j است. $\hat{\theta}_{ij}$ ($\hat{\theta}_{ij}$=۱، ۲،...،٥۰۰) برآورد θ_{ij} و $\bar{\theta}_{ij} = \frac{1}{N_i}\sum_{i=1}^{500}\hat{\theta}_{ij}$، میانگین ٥۰۰ برآورد θ_{ij} ($\hat{\theta}_{ij}$) در θ_j است.

۳- **اریب و میانگین مجذور خطا[2] (MSE)**: در معادله‌های ۱۳ و ١٤، N تعداد کل آزمودنی‌ها در تمام نقاط ثابت توانایی (٦٥ نقطه ثابت توانایی از ٤- تا ٤)، یعنی برابر با ($\sum_{j=1}^{65}\sum_{i=1}^{500} N_{ij} = 32500$) آزمودنی است و $\hat{\theta}_i$ برآوردکننده چندمین آزمودنی (i) با سطح توانایی واقعی θ_i است:

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i) \quad (۱۳)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2 \quad (۱٤)$$

٤- **اریب شرطی[3] و میانگین مجذور خطای شرطی[4] (CMSE)**: در معادله‌های ۱٥ ١٦ مقدار θ_j همان مقادیر ثابت (یعنی؛ ٤+، ۳/۸۷٥+، ... ۳/۸۷٥-، ٤-=θ_j) است، که در اینجا برآوردهای $\hat{\theta}_{ij}$ در هر سطح توانایی، از مقدار ثابت θ_j کم می‌شود:

$$Conditional\ Bias(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j) \quad (۱٥)$$

$$CMSE(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j)^2 \quad (١٦)$$

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

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:

$$X^2 = \sum_{i=1}^{n} \frac{(r_i - \frac{L}{n})^2}{\frac{L}{n}} \qquad (۱۷)$$

در این معادله r_i، نسبت نرخ مشاهده شده i‌امین سؤال، L=۲۰ طول آزمون و n تعداد سؤال‌های خزانه است. معادله ۱۷ اختلاف بین نرخ مواجهه سؤال مشاهده شده و ایده‌آل را محاسبه می‌کند و مقدار اثربخشی استفاده از خزانه سؤال را نیز تعیین می‌کند. مقدار کای دو کوچک نشان می‌دهد که بیشتر سؤال‌ها استفاده شده‌اند.

٦- **درصد سؤال‌های بیش مواجهه شده**[2]: نرخ مواجهه هر سؤال را می‌توان به عنوان نسبت تعداد دفعات اجرای سؤال به تعداد کل آزمودنی‌ها در نظر گرفت. در مجموع، سطح متوسط نرخ مواجههٔ سؤال مناسب است. نرخ بالای مواجههٔ هر سؤال بدین معنی است که خطر فاش شدن سؤال برای آزمودنی‌های بعدی افزایش می‌یابد. اگر چنین باشد، هم امنیت و هم روایی[3] آزمون به دلیل نرخ بالای مواجهه سؤال مورد تهدید قرار می‌گیرد. بنابراین، درصد سؤال‌های بیش مواجهه شده، به‌عنوان ملاک مهمی برای ارزیابی موفقیت برنامه سنجش انطباقی کامپیوتری (CAT) در نظرگرفته می‌شود (هَو و چانگ[4]، ۲۰۰۱).

۷- **درصد سؤال‌های کم مواجهه شده**[5]: نرخ کم مواجهه شدن یک سؤال بدین معنی است که یک سؤال به ندرت در برنامه سنجش انطباقی کامپیوتری (CAT) مورد استفاده قرار گیرد. خزانه سؤالی که سؤال‌های بسیار زیادی با نرخ مواجهه خیلی پایینی دارد، فایده کمی دارد. دو موضوع به‌صرفه‌بودن طراحی سؤال‌ها و مناسب بودن شیوه انتخاب آنها، به دلیل نرخ مواجهه کم سؤال به چالش کشیده می‌شوند. نرخ مواجهه پایین‌تر از ۰/۰۲ به عنوان سؤال کم مواجهه شده در نظرگرفته می‌شود (هَو و چانگ، ۲۰۰۱؛ رِکیس، ۲۰۰۹).

1.Skewness of item exposure rate distribution
2.Percentage of overexposed items
3.Validity
4.Hau & Chang
5.Percentage of underexposed items

۸- **نرخ همپوشی آزمون**[1]: 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[2]. به‌طور ایده‌آل، تعداد سؤال‌های مشترک بین دو آزمودنی که به طور تصادفی نمونه‌گیری شده‌اند، باید حداقل باشد (چانگ و ییِنگ، ۱۹۹۹؛ چن، آنکنمان، اسپری[3]، ۱۹۹۹):

$$\bar{T} = \frac{\sum_{i=1}^{n}\binom{m_i}{2}}{L\binom{N}{2}} = \frac{\sum_{i=1}^{n} m_i(m_i-1)}{LN(N-1)} \qquad (۱۸)$$

در این رابطه N، تعداد آزمون‌های سنجش انطباقی کامپیوتری (CAT) (با طول ثابت) اجرا شده، L، تعداد سؤال‌ها در هر یک از سنجش انطباقی کامپیوتری (CAT)، n تعداد سؤال‌های خزانه و m_i تعداد دفعاتی است که سؤال i برای همه N تعداد آزمون سنجش انطباقی کامپیوتری (CAT) اجرا شده است.

٦) متغیرهای مستقل

در همهٔ طرح‌های خزانه سؤال دو متغیر مستقل دستکاری شدند. روش طراحی پارامترهای سؤال و روش کنترل مواجهه. نرخ مواجههٔ هدف بر اساس روش سیمپسون-هتر برابر با (۰/۳۳) قرار داده شد. این نرخ در خزانه سؤال عملیاتی هم برابر با ۰/۳۳ قرار گرفت (برای جزئیات بیشتر جدول (۱) را ببینید).

جدول (۱) طرح شبیه‌سازی خزانه‌های سؤال بهینه

طول آزمون	۲۰
توزیع توانایی	$N(0,1)$
کنترل مواجهه	عدم کنترل مواجهه
	روش سیمپسون-هتر(با نرخ مواجهه۰/۳۳)
روش طراحی خزانه سؤال	مدل تصادفی (R)
	مدل تصادفی آمیخته و پیش‌بینی (MRP)
	مدل کمینه آگاهی آزمون (MTI)
پهنای bin	b-bin:۰/۲ ۰/٤I_{max}=۲=Δ۲a-bin:Δα
تعادل محتوایی	آزمون تک‌محتوایی

1.Test overlap rate
2.Expected test length
3.Chen & Ankenmann & Spray

نتایج

در این قسمت نتایج مربوط به شبیه‌سازی خزانه‌های بهینه و مقایسه عملکرد آنها با خزانه سؤال عملیاتی شرح داده می‌شود. با توجه به متغیرهای مستقل ذکر شده در بالا (سه روش ایجاد سؤال بهینه و اعمال یا عدم اعمال شیوه کنترل مواجهه) شش الگوی خزانه سؤال بهینه وجود دارد. بر این اساس، خزانه سؤال عملیاتی به اختصار با OP[1] و خزانه‌های سؤال بهینه، شبیه‌سازی شده به اختصار با ROP[2] نشان داده شده‌اند (برای حالت عدم کنترل مواجهه، ROP_1 برای R، ROP_2 برای MRP، ROP_3 برای MTI و برای حالت کنترل مواجهه ROP_4 برای R، ROP_5 برای MRP، ROP_6 برای MTI).

الف) خزانه‌های سؤال بهینه، بدون کنترل مواجهه بیش از حد سؤال

بر اساس نتایج جدول (۲)، خزانه‌های سؤال بهینه شامل حداقل تعداد سؤال هستند. البته این نتیجه تعجب برانگیز نیست، زیرا هر سه خزانه بهینه با فرض اینکه هیچ روش کنترل مواجهه‌ای بر اجرای سؤال‌ها اعمال نشده، ساخته شده‌اند، در حالی که خزانه سؤال عملیاتی بر اساس روش کنترل مواجهه سیمپسون-هتر ساخته شده است. نتایج حاکی از آن است که همه خزانه‌ها سؤال‌هایی با دامنه دشواری متنوع (تقریباً از ٤- تا ٤+) دارند، هر چند که سؤال‌های موجود در خزانه‌های بهینه، نسبت به خزانه‌های عملیاتی، دامنه دشواری تقریباً کوچک‌تری دارند. خزانهٔ عملیاتی دارای تعداد زیادی سؤال با پارامتر b بین ۰/۵- تا ۱ است. در حالی که، خزانه‌های بهینه توزیع بزرگ‌تری در میانهٔ اندازه‌های متعدد b-bin دارند. خزانه بهینه MTI(ROP_3) شامل حداقل تعداد سؤال است و میانگین پارامتر (a) سؤال‌های آن کوچک‌تر است، به طوری که از ۱/۰۵ تا ۲/٤۲ پراکنده شده‌اند. خزانه سؤال MRP(ROP_2) و R(ROP_1) شباهت بیشتری به یکدیگر دارند. در خزانه MRP(ROP_2) سؤال‌های دشوار شیب بالاتر و سؤال‌های آسان شیب متوسط یا پایین‌تری دارند.

1.Operational item pool
2.Range-optimal item pool

نتایج مربوط به ارزیابی عملکرد این خزانه‌ها در جدول (۳) اراوئه شده است. برآورد توانایی در هر سه خزانه بهینه و عملیاتی، سطح معینی از اریب مثبت دارد، هر چند که مقدار این اریب‌ها در خزانه‌های بهینه ناچیز است. میانگین مجذور خطا (MSE) در خزانه‌های سؤال بهینه کوچک‌تر از خزانه سؤال عملیاتی است. در میان خزانه‌های سؤال بهینه، خزانه MRP(ROP_2) عملکرد بهتری در این شاخص نشان می‌دهد. همچنین مطابق با نتایج جدول (۳)، خزانه‌های سؤال بهینه با اینکه وجود سؤال‌های کمتر، نرخ همپوشی پایین‌تری دارد. این نتیجه حاکی از آن است که نرخ همپوشی آزمون با اندازه خزانه سؤال[1] رابطه ندارد بلکه به ترکیب بهینه سؤال‌ها بستگی دارد.

جدول (۲) آماره‌های سؤال در چهار نوع خزانه، بدون کنترل مواجهه (b-bin=۰/۲) و a-bin: *Δa2=2ΔIMaximum* = 0.4

نوع خزانه	اندازه /خزانه	A				b				c			
		میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل
OP	۸۴۰	۱/۰۸۹	۰/۲۸۴۴	۳/۰۴۵	۰/۱۶۶	-۰/۰۳۹	۰/۷۷۹	۳/۹۸۱	-۳/۵۹۶	۰/۱۴۵	۰/۸۰۱	۰/۵۱۷۹	۰/۰۰۰۵
R(ROP_1)	۲۸۴	۱/۶۷	۰/۲۷	۲/۵۵	۰/۷۸	-۰/۲۴	۱/۰۳	۳/۳۱	-۳/۰۶	۰/۱۴۶	۰/۷۷	۰/۵۵	۰/۰۰۷
MRP(ROP_2)	۱۹۷	۱/۷۸	۰/۳۵	۲/۶۹	۰/۰۴	۰/۰۱۸	۱/۰۵	۳/۹۴	-۳/۱۲	۰/۱۴۲	۰/۰۴۸	۰/۵۱۲۴	۰/۰۰۶
MTI(ROP_3)	۱۸۴	۱/۳۹	۰/۱۳	۲/۴۲	۱/۰۵	-۰/۰۴	۱/۰۷۸	۳/۸۲	-۳/۲۴	۰/۱۴۶	۰/۰۸	۰/۵۳	۰/۰۰۴

جدول(۳) شاخص‌های ارزیابی عملکرد خزانه‌های سؤال بدون کنترل مواجهه، (b-=۰/۲ bin) و a-bin: *Δa2=2ΔIMaximum* = 0.4

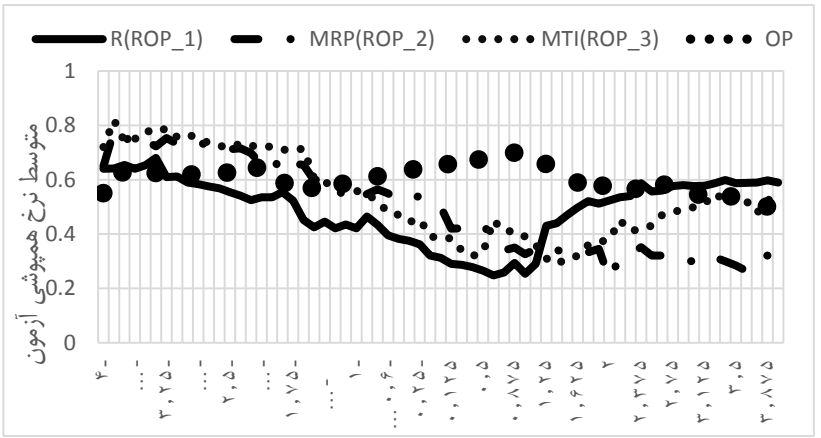
آماره‌ها	OP	R	MRP	MTI
Bias	۰/۰۳۳	/۰۰۸۳	۰/۰۰۷۹	۰/۰۰۸۱
MSE	۰/۱۱	۰/۰۰۹۱	۰/۰۰۲۵	۰/۰۰۸۵
کجی نرخ مواجهه	۳۳/۸۵	۱۸/۰۲	۱۶/۸۲	۱۴/۹۱

۱. همان‌طور که قبلاً ذکر شد، اندازه خزانه سؤال بر اساس ملاک bin-and-union تعیین شده است.

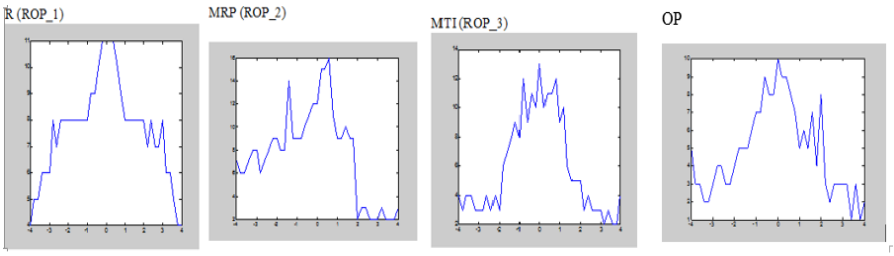
آماره‌ها	OP	R	MRP	MTI
نرخ همپوشی سؤال	۰/٤٦۹۳	۰/٤۰۸۵	۰/٤۰۹٤	۰/٤۱۲٦
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از ۰/۳۳	۱۱/۲۱٪	۱۱/۸۹٪	۱۵/۲۱٪	۱۳/۰۱٪
درصد سؤال‌های با نرخ مواجهه کوچک‌تر از ۰/۰۲	٤۰٪	۱۷/۷۷٪	۱۳/۵۳٪	۷/۲۵٪
اندازه خزانه سؤال	۸٤۰	۲۸٤	۱۹۷	۱۸٤

مطابق نمودار (۲)، خزانه‌های سؤال بهینه، نرخ همپوشی آزمون بیشتری در سطوح توانایی زیر (۲-) نشان می‌دهند. البته، در عمل آزمودنی‌های اندکی در این سطوح توانایی وجود دارد. امّا خزانه سؤال بهینه MRP(ROP_2) نسبت به بقیه خزانه‌ها، در سطوح توانایی بالای (۲)، نرخ همپوشی آزمون کمتری دارد. دلیل این امر می‌تواند این باشد که این خزانه برای سطوح توانایی بالای ۲ سؤال‌های بیشتری با ضریب تشخیص بالا ایجاد کرده است. بنابراین، این قضیه مانع از همپوشی بالا در این سطح توانایی می‌شود. همچنین، خزانه‌های بهینه با درصد پایینی از سؤال‌های کم مواجهه شده دارند. البته خزانه‌های MRP(ROP_2) و MTI(ROP_3) نرخ بالاتری از درصد سؤال‌های بیش مواجهه شده دارند، که به دلیل عدم کنترل مواجهه سؤال‌ها و تعداد بسیار کم سؤال در این دو خزانه بهینه است. طبیعی است که با افزایش تعداد سؤال‌ها در خزانه، نرخ همپوشی و بیش مواجهه شدن کاهش می‌یابد. نمودار (۳)، نتایج مربوط به درصدهای مواجهه سؤال در هر یک از سطوح توانایی را نشان می‌دهد. در هر سه خزانه بهینه، سؤال‌های خیلی آسان و خیلی دشوار نرخ‌های مواجهه کوچک‌تری دارند. بخصوص در خزانه MRP(ROP_2) سؤال‌های خیلی دشوار نرخ مواجهه کمتری دارند، که البته یک دلیل آن می‌تواند این باشد که تعداد بیشتری سؤال با ضریب تشخیص بالا در این سطوح ساخته شده است. نتایج حاکی از آن است که در هر سه خزانه بهینه، سؤال‌هایی با سطوح دشواری متوسط بیشترین قابلیت استفاده را داشته‌اند. همچنان که در نمودار (٤) ملاحظه می‌شود، میانگین آگاهی خزانه‌های سؤال در مقایسه با یکدیگر، در سطوح ثابت توانایی مقادیر متفاوتی دارند. امّا هم در خزانه عملیاتی و هم در سه خزانه بهینه، با وجود مقادیر متفاوت آگاهی، این آگاهی در وسط پیوستار توانایی، به اوج خود می‌رسد. خزانه سؤال R(ROP_1) و خزانه عملیاتی مشابه هم عمل کرده‌اند. خزانه سؤال MTI(ROP_3) در کل دامنهٔ سطوح توانایی، به طور معنی‌داری آگاهی کوچک‌تری را

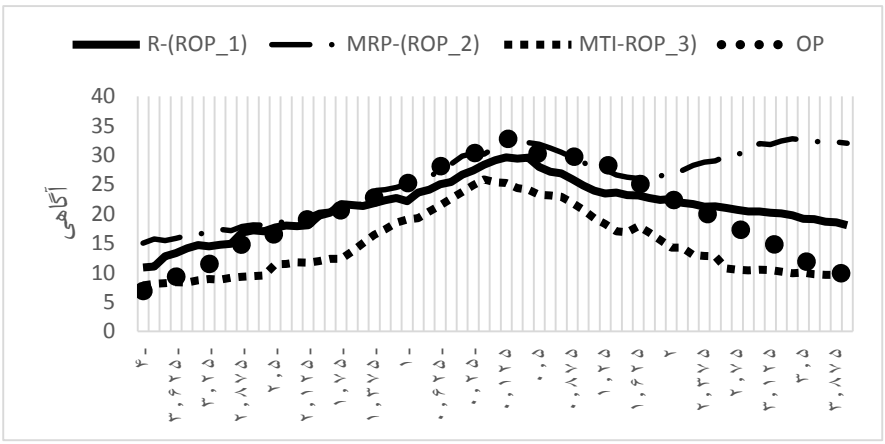
ایجاد کرده‌است، که البته این نتیجه، به دلیل ماهیت روشی است که سؤال‌ها ایجاد می‌شوند. امّا میزان آگاهی که در طول سطوح توانایی ایجاد کرده، فراتر از آگاهی هدف است. نمودارهای (۵) تا (۷) خطای استاندارد شرطی اندازه‌گیری (CSEM)، اریب شرطی و میانگین مجذور خطا (CMSE) را در هر چهار خزانه سؤال نمایش می‌دهند. خطای استاندارد اندازه‌گیری در هر چهار خزانه سؤال در سطوح توانایی زیر (۲-) حداکثر مقادیر است. در سطوح متوسط، توانایی این مقدار به حداقل خود می‌رسد، ولی در سطوح بالای توانایی در هر یک از خزانه‌ها به صورت متفاوتی عمل می‌کند. در هر سه خزانهٔ بهینه میزان خطای استاندارد اندازه‌گیری در سطوح توانایی بالای (۲) کمتر از خزانه عملیاتی است، بخصوص در خزانه MRP(ROP_2) که مقدار خطای اندازه‌گیری به حداقل مقدار خود می‌رسد. دلیل این امر این است که این خزانه برای توانایی‌های بالای ۱/۷۲ سؤال‌هایی با ضریب تشخیص بالاتر ایجاد کرده است. مطابق نمودار (٦)، میزان اریب در سطوح توانایی پایین و بالا خزانه سؤال MTI(ROP_3) از سایر خزانه‌ها بالاتر است. دلیل این نتیجه آن است که در این سطوح، توانایی حداقل مقدار آگاهی مورد نیاز برابر با ۱۵/٤ است. این قضیه باعث می‌شود سؤال‌هایی با ضریب تشخیص پایین‌تر در این سطوح ساخته شود که با توجه به کوتاه بودن طول آزمون در این خزانه اریب مثبت به‌وجود می‌آید. نتایج در نمودار ۷ نشان می‌دهد که مقدار میانگین مجذور خطا (MSE) درهر سه خزانه بهینه، کوچک‌تر از خزانه سؤال عملیاتی است.



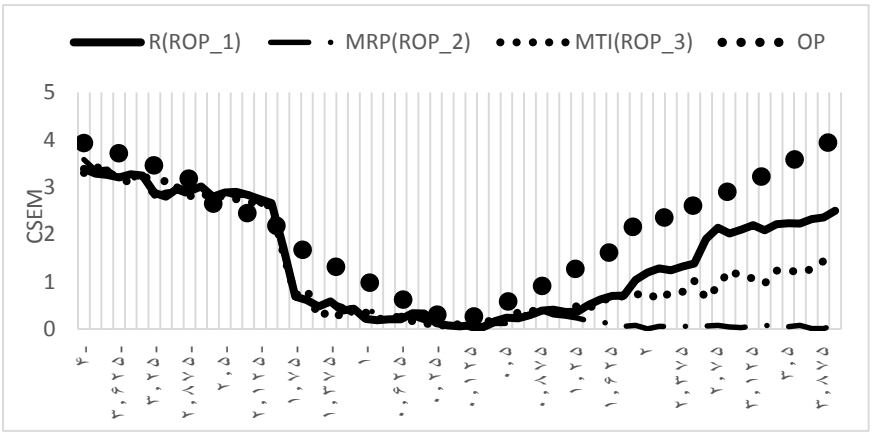
نمودار (۲) نرخ همپوشی آزمون به شرط (θ)، بدون کنترل مواجهه سؤال



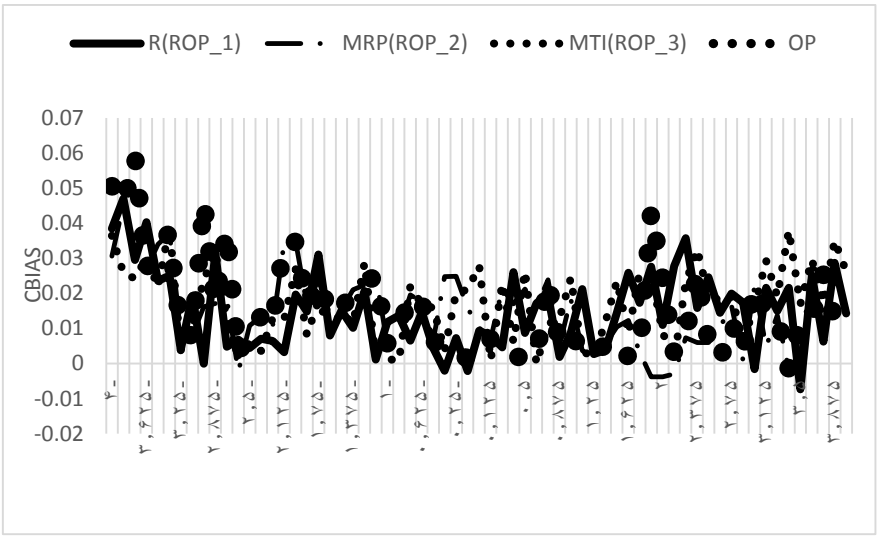
نمودار (۳) درصد سؤال‌های بیش مواجهه شده آزمون به شرط (θ)، بدون کنترل مواجهه سؤال



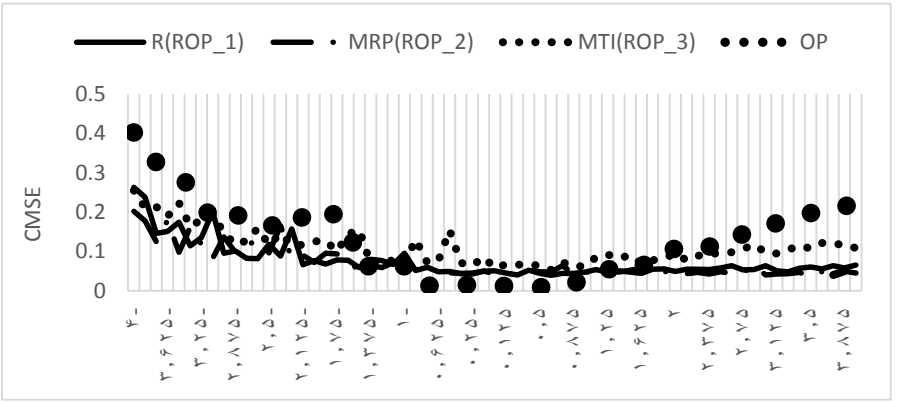
نمودار (٤) متوسط آگاهی آزمون به شرط (θ)، بدون کنترل مواجهه سؤال



نمودار(۵) خطای استاندارد اندازه‌گیری شرطی آزمون به شرط (θ)، بدون کنترل مواجهه سؤال



نمودار (٦) اریب شرطی آزمون به شرط (θ)، بدون کنترل مواجهه سؤال



نمودار (۷) میانگین مجذور خطای آزمون به شرط (θ)، بدون کنترل مواجهه سؤال

ب) خزانه‌های سؤال بهینه با کنترل میزان مواجهه بیش از حد سؤال

در این مرحله، خزانه‌های سؤال بهینه با درنظر گرفتن عامل کنترل مواجههٔ سیمپسون-هتر ایجاد شدند. این عامل به دلیل کنترل امنیت سؤال‌های آزمون در طراحی خزانه‌های سؤال وارد شد. در این روش به هر یک از سؤال‌های موجود در خزانه با روشی تکراری،

یک پارامتر کنترل مواجهه برابر با ۰/۳۳ اختصاص داده شد. در هنگام اجرا سنجش انطباقی کامپیوتری (CAT)، هر سؤالی که برای اجرا انتخاب می‌شد، با نرخ مواجهه ۰/۳۳ مقایسه می‌شد. اگر احتمال انتخاب این سؤال بیشتر از این احتمال بود، سؤال به خزانه بازگردانده می‌شد و سؤال دیگری با همین ویژگی‌ها انتخاب و اجرا می‌شد. این فرآیند با اجراهای متوالی و تکراری برای کل افراد اجرا شد که فرآیندی وقت‌گیر و طولانی بود. در پایان، خزانه‌های سؤال بهینه‌ای ایجاد شد که علاوه بر داشتن ویژگی‌های بهینه برای سنجش انطباقی کامپیوتری (CAT)، امنیت آزمون را نیز تأمین می‌کرد. نتایج نشان داد، خزانه R(ROP_4) در طول ماتریس پارامترهای a و b توزیع یکنواخت‌تری دارد. در صورتی که خزانه MRP(ROP_5) سؤال‌هایی با ضریب تشخیص بیشتری دارد، امّا توزیع آن نسبت به زمانی که کنترل مواجهه اعمال نشده بود یکنواخت‌تر بود. خزانه MTI(ROP_6) سؤال‌های بیشتری با ضرایب تشخیص پایین‌تری داشت. نتایج جدول (٤) حاکی از آن است که خزانه MRP(ROP_5) بیشترین مقدار متوسط ضریب تشخیص و MTI(ROP_6) کمترین مقدار را داشت. البته میزان پراکندگی در پارامتر a در خزانه MTI(ROP_6) کمتر از بقیه خزانه‌ها بود. در حقیقت زمانی‌که در شبیه‌سازی خزانه‌های سؤال بهینه، روش سیمپسون–هتر وارد می‌شود، نسبت به زمانی‌که این عامل مورد توجه قرار نمی‌گیرد، اندازه خزانه سؤال بزرگ‌تر خواهد شد. به طوری‌که نتایج حاصل از این پژوهش نیز نشان داد، خزانه R(ROP_4) نسبت به خزانه R(ROP_1) ۳۰ سؤال بیشتر داشت، خزانه MRP(ROP_5) نسبت به خزانه MRP(ROP_2) ۱۱۵ سؤال بیشتر و خزانه MTI(ROP_6) نسبت به خزانه MTI(ROP_3) ٦۰ سؤال بیشتر داشت. با این وجود، هنوز اندازه خزانه‌های سؤال بهینه، کمتر از خزانه عملیاتی بود. سؤال‌های اضافه شده به خزانه‌های این مرحله ضرایب تشخیص بالاتری داشتند، زیرا آنها به مواجهه بیشتر متمایل بودند. این قضیه باعث می‌شود که متوسط پارامتر a در این نوع خزانه‌ها بالاتر شود. همان‌طور که جدول (۵) نشان می‌دهد، نتایج ارزیابی این خزانه‌ها حاکی از آن است که هر سه خزانه بهینه مقدار اندکی اریب منفی دارند. همچنین، میزان MSE در خزانه‌های بهینه بسیار کوچک‌تر از خزانه عملیاتی است، بخصوص در خزانه MRP(ROP_5) این مقدار به حداقل خود می‌رسد.

جدول (٤) آماره‌های سؤال با کنترل میزان مواجهه سؤال (b-bin=۰/۲) و a-bin: *Δa2=2ΔIMaximum* = 0.4

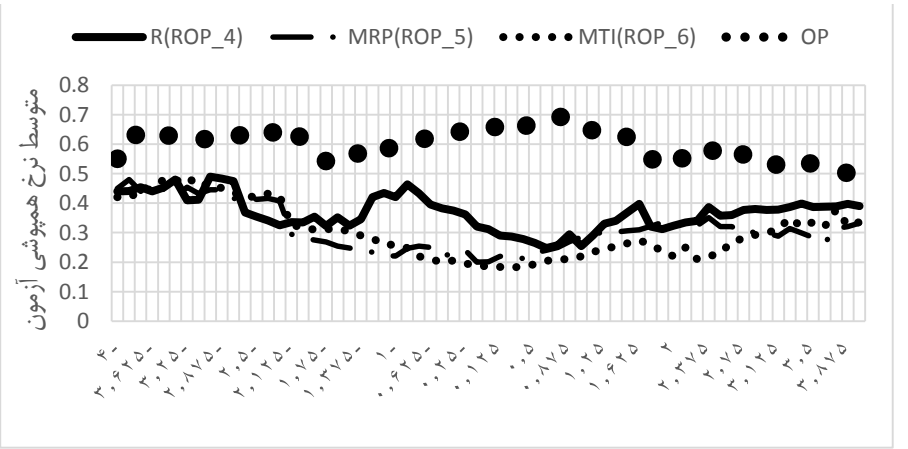
نوع خزانه	اندازه/ خزانه	a				b				C			
		میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل	میانگین	انحراف استاندارد	حداکثر	حداقل
OP	۸٤۰	۱/۰۸۹	۰/۲۸٤٤	۳/۰٤٥	۰/۱٦٦	-۰/۰۳۹	۱/۰۷۹	-۳/۹۸۱	-۳/٥۹٦	۰/۱٤٥	۰/۸۰۱	۰/٥۱۷۹	۰/۰۰۰٥
R(ROP_4)	۳۱٤	۱/٦۸٥	۰/۲۷۹	۲/۷۲	۰/۸۸٤	-۰/۲۷۸	۰/۹٥	۳/۹۲	-۳/۳٤	۰/۱٤۷	۰/۰۷۹	۰/٤۸	۰/۰۰٦
MRP(ROP_5)	۳۱۲	۱/۸۱	۰/٤٤	۲/۹۹	۰/۷٤٥	۰/۰۰۱۳	۰/۹۷	۳/۹۹	-۳/۱۱	۰/۱٤۱	۰/۰۷۸	۰/٤۲۹	۰/۰۰۷
MTI(ROP_6)	۲٤٤	۱/٤۸	۰/۱۱٦	۲/۱۲	۱/۱٥	۰/۰٤۳	۰/۹۱	۳/۸۸	-۳/۱۰	۰/۱٤۷	۰/۰۸	۰/٥۳	۰/۰۰٤

جدول (٥) شاخص‌های ارزیابی عملکرد خزانه سؤال با کنترل میزان مواجهه سؤال (b-bin=۰/۲) و a-bin: *Δa2=2ΔIMaximum* = 0.4

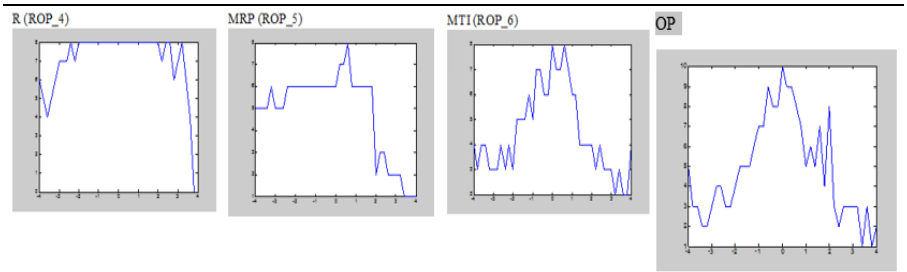
آماره‌ها	OP	R	MRP	MTI
Bias	۰/۰۳۳	-۰/۰۰۰۲۸	-۰/۰۰۰۲۱	-۰/۰۰۰۲٥
MSE	۰/۱۱	۰/۰۰۱۸	۰/۰۰۱٥	۰/۰۰۱٦
کجی نرخ مواجهه	۳۳/۸٥	۷/۲۳۹٦	٦/٥٥٤٦	٦/۰٦۲۱
نرخ همپوشی سؤال	۰/٤٦۹۳	۰/۲٤٥٦	۰/۲٤٤۲	۰/۲٤۳۹
درصد سؤال‌هایی با نرخ مواجهه بزرگ‌تر از ۰/۳۳	۱۱/۲۱٪	٦/۰٥٪	۲/۲٤٪	۲/۰٥٪
درصد سؤال‌هایی با نرخ مواجهه کوچک‌تر از ۰/۰۲	٤۰٪	۲۳/۲۷٪	۲۹/٤۲٪	۳۰/۱٦٪
اندازه خزانه سؤال	۸٤۰	۳۱٤	۳۱۲	۲٤٤

مطابق نمودار (۸) نرخ همپوشی آزمون در خزانه‌های بهینه، کمتر از خزانه عملیاتی است. خزانه MTI(ROP_6) و MRP(ROP_5) کوچک‌ترین نرخ همپوشی آزمون را در

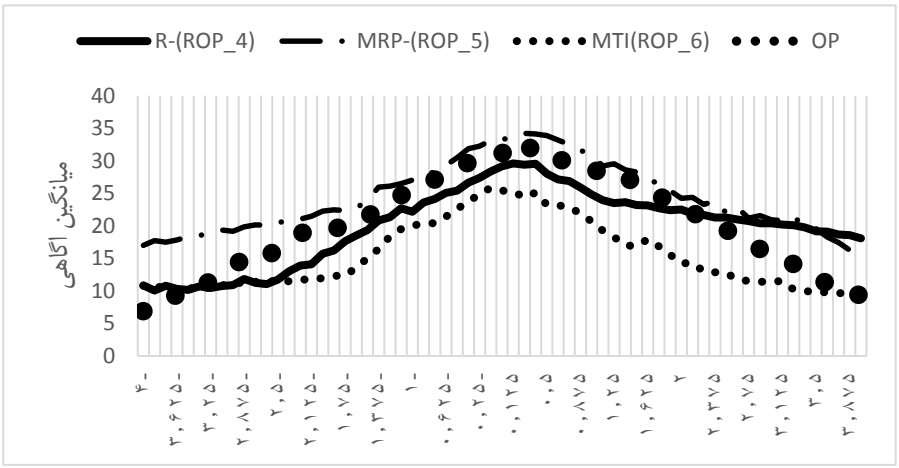
حداکثر سطوح توانایی دارند. مطابق جدول (۵) مقدار کجی نرخ مواجهه سؤال و مقدار درصد سؤال‌هایی با نرخ مواجهه بزرگ‌تر از ۰/۳۳ در خزانه‌های بهینه نسبت به خزانه عملیاتی به حداقل مقدار خود رسیده است. نمودار (۹) نیز این قضیه را نمایش می‌دهد. با وجود اینکه تعداد سؤال‌ها در خزانه‌های بهینه خیلی کمتر از خزانه عملیاتی است، این تفاوت‌ها قابل توجه است. نمودار ۱۰ میانگین آگاهی آزمون را در سطوح متفاوت توانایی در همه خزانه‌های سؤال نشان می‌دهد. خزانه MRP(ROP_5) در تمام سطوح توانایی آگاهی بیشتری فراهم می‌کند. خزانه R(ROP_4) آگاهی مشابه با خزانه عملیاتی ایجاد می‌کند. خزانه MTI(ROP_6) نسبت به سه خزانه دیگر به طور معنی‌داری آگاهی کمتری ایجاد می‌کند، امّا فراتر از میزان آگاهی هدف است. نمودارهای(۱۱، ۱۲ و ۱۳) به ترتیب خطای استاندارد اندازه‌گیری، اریب و میانگین مجذور خطا را در سطوح متفاوت توانایی در هر چهار خزانه نمایش می‌دهد. نتایج این نمودارها نشان می‌دهد که این مقادیر در سه خزانه بهینه مشابه و کمتر از خزانه عملیاتی است. به خصوص این مقادیر در سطوح توانایی بالاتر از ۱/۵ در خزانه MRP(ROP_5) و MTI(ROP_6) به حداقل مقادیر خود می‌رسند.



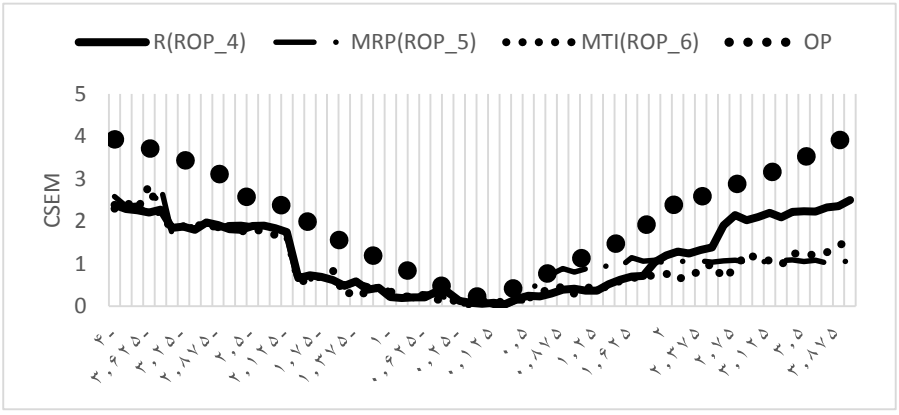
نمودار(۸) نرخ همپوشی آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون-هتر



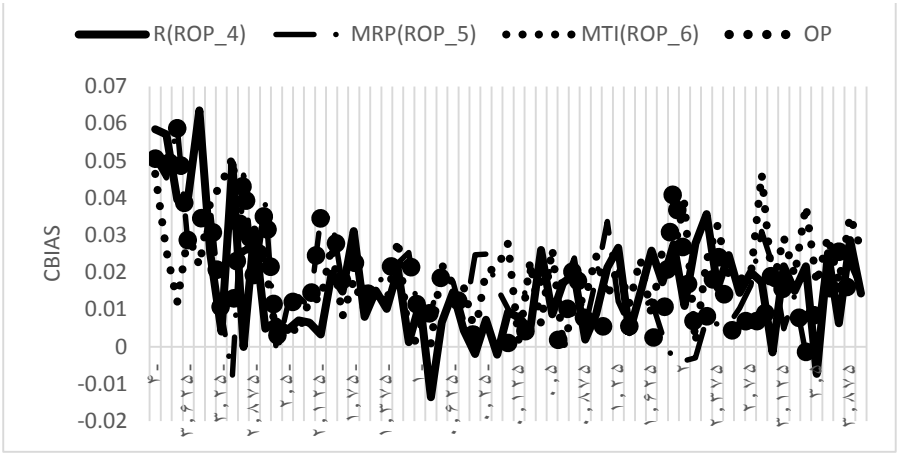
نمودار(۹) درصد سؤال‌ها بیش مواجهه شده آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون–هتر



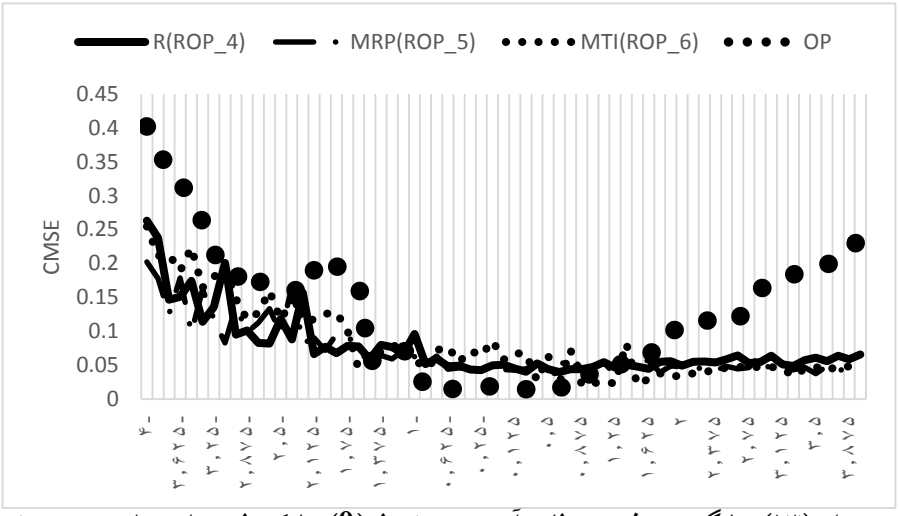
نمودار (۱۰) متوسط آگاهی آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون–هتر



نمودار (۱۱) خطای استاندارد اندازه‌گیری آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون–هتر



نمودار (۱۲) اریب شرطی آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون-هتر



نمودار (۱۳) میانگین مجذور خطای آزمون به شرط (θ)، با کنترل میزان مواجهه به روش سیمپسون-هتر

بحث و نتیجه‌گیری

هدف این پژوهش، ساخت خزانه‌های سؤال بهینه بر اساس رویکرد اکتشافی رکیس، با در نظر گرفتن امنیت آزمون بود. خزانه‌هایی که با توجه به مدل سه پارامتری لجستیک مدرج شده‌اند. نمونه‌های اولیه کاربرد دو روش (R، P) در کارهای مک‌برید و وایس،

(۱۹۷۷) در پیش‌بینی و برآورد پارامترهای سؤال‌ها یافت شد. گو و رکیس (۲۰۰۷) از دو روش پیش‌بینی[1] (PM) و حداقل آگاهی آزمون (MTI) برای طراحی خزانه سؤال بهینه استفاده کردند. همچنین، هی و رکیس (۲۰۱۱) نیز از سه روش (R)، (MRP) و (MTI) در حالت‌های متفاوت b-bin و با تغییرات آگاهی متفاوت سؤال استفاده کردند. در این پژوهش نیز خزانه‌های سؤال بهینه بر مبنای خصوصیات سؤال تعیین شده بر پایهٔ معادله‌های ایجاد شده توسط برن بام[2] (۱۹٦۸) شبیه‌سازی شدند و با دستکاری دو عامل روش ایجاد سؤال بهینه و کنترل مواجهه سؤال، تعداد شش خزانه سؤال بهینه که نتایج آنها در فوق ارایه شد، ایجاد شد. این خزانه‌ها از نظر اندازه خزانه سؤال و متوسط مقادیر پارامتر a، b و c متفاوت بودند. به‌طور خلاصه، اندازه خزانه‌های سؤال از ۱۸٤ (در MTI(ROP_3)) تا ۳۱٤ (در R(ROP_4)) متفاوت بود. همچنین، متوسط مقادیر پارامتر a از ۱/۸۱ (در MRP(ROP_5)) تا ۱/۳۹ (در MTI(ROP_3)) تفاوت داشت و متوسط مقادیر پارامتر b نیز از ۰/۰٤۳ (در MTI(ROP_6)) تا ۰/۲۷۸- (در R(ROP_4)) متفاوت بود. در مقایسه با خزانه عملیاتی، تقریباً همه خزانه‌های بهینه، به استثنای خزانه‌ای که با روش (MTI) ایجاد شد، شامل سؤال‌هایی با مقادیر پارامتر a بالاتر از ۱/۵ بودند. این نتیجه، مشابه با تحقیق یوری[3] (۱۹۷۷) و همچنین هی و رکیس (۲۰۱۱) است. این دو پژوهش مقدار متوسط شیب بالاتر از ۱/۵ را به عنوان ویژگی برتر خزانه‌های سؤال سنجش انطباقی کامپیوتری (CAT) توصیه کرده‌اند. به عبارت دیگر، اگر مقدار متوسط ضریب تشخیص در یک خزانه سؤال سنجش انطباقی کامپیوتری (CAT) بالاتر از ۱/۵ باشد، آن خزانه سؤال یکی از ملاک‌های بهینه‌گی را دارد، که وندرلیندن (a۲۰۰۰) و وندرلیندن، پاشلی[4](۲۰۰۰) نیز این موضوع را خاطر نشان کرده‌اند. در حالی که توزیع پارامترهای b در تمام خزانه‌های بهینه به صورت یکنواختی در سراسر مقیاس توانایی توزیع شده است، این توزیع برخلاف توزیع نرمال زیربنایی توانایی آزمودنی‌هایی بود که آزمون برای آنها شبیه‌سازی شد. این ویژگی نیز در تحقیقات هی و رکیس (۲۰۱۱) و رکیس (۲۰۱۰) مورد تایید قرار گرفت، زیرا ماهیت سنجش انطباقی کامپیوتری (CAT) ایجاب می‌کند که توزیع پارامتر b در طول پیوستار توانایی یکنواخت باشد. همچنان که نتایج نیز نشان می‌دهند روش‌های ایجاد سؤال (R)

1.Prediction model (PM)
2. Birnbaum
3. Urry
4.Pashley

و (MRP) نسبت به روش (MTI)، با وجود تمام شرایط یکسان، به ایجاد خزانه‌هایی با متوسط مقادیر a بالاتر تمایل دارند. دلیل اینکه روش (MTI) متوسط مقادیر a پایین‌تری ایجاد می‌کند، این است که از قاعده حداقل آگاهی آزمون یکسانی برای آزمودنی‌هایی با توانایی واقعی متفاوت استفاده می‌کند. در این مطالعه برای به حداقل رساندن این مشکل، برای شبیه‌سازی سؤال‌ها با توجه به سطوح توانایی آزمودنی‌ها، حداقل مقادیر آگاهی متفاوتی در نظر گرفته شد. به عبارت دیگر، مقدار متوسط آگاهی هدف آزمون در سه نقطه برش متفاوت، که حاصل آن چهار سطح عملکرد بود، محاسبه شد. این نمرات با استفاده از داده‌های آزمون سنجش انطباقی کامپیوتری (CAT) عملیاتی به دست آمد. این سه نمره برش بر اساس پژوهش‌های گو و رکیس (۲۰۰۷) و نیز هی و رکیس (۲۰۱۱) که برای آزمون مهارت ریاضی انجام گرفته بود به‌دست آمده است. در روش (MYI) متوسط مقدار شیب پایین‌تر از دو روش (R) و (MRP) است، زیرا آگاهی هدف آزمون برای آزمودنی‌هایی که توانایی واقعی‌شان در طول دامنه نمره برش از مقدار متوسط فاصله می‌گیرد، متفاوت است. البته با بالا رفتن مقادیر حداقل آگاهی آزمون، روش (MTI) به ایجاد خزانه‌هایی با متوسط مقادیر a بالاتر گرایش پیدا می‌کند. در پرتو ویژگی‌های خزانه سؤال عملیاتی به کار رفته در این مطالعه(یعنی، مقادیر متوسط a برابر با ۱/۰۸۹) به نظر می‌رسد که شیوهٔ به کار رفته در این مطالعه برای قرار دادن آگاهی هدف آزمون، به خوبی عمل می‌کند، زیرا خزانه‌های بهینه ایجاد شده مقادیر معقولی پارامتر سؤال دارند. در کل، ویژگی‌های پارامتر سؤال به‌دست آمده در این مطالعه با نتایج یوری (۱۹۷۷) و جنسما[1] (۱۹۷۷) سازگار است. در این دو مطالعه نیز، میانگین پارامترهای a در خزانه‌های بهینه در حدود ۱/۱ و حداقل مقدار آن نیز بیشتر از ۸/. بوده است. همچنین، دشواری‌های سؤال‌ها نیز به صورت یکنواختی توزیع شده‌اند. این نوع توزیع یکی از مهم‌ترین ضروریات خزانه سؤال برای سنجش انطباقی کامپیوتری (CAT) است. از آنجا که روش انتخاب سؤال، روش بیشینه آگاهی است، این ضرورت، معقول به نظر می‌رسد، زیرا برای تمام آزمودنی‌ها در طول دامنهٔ توانایی، سؤالی برای اجرا انتخاب می‌شود که حداکثر میزان آگاهی را ایجاد کند. حال اگر تعداد کافی از سؤال‌ها در یک سطح دشواری ویژه وجود نداشته باشد، این امکان وجود دارد که الگوریتم انتخاب سؤال، سؤالی را که بهینه و مناسب نیست انتخاب کند. این انتخاب

1. Jensema

باعث می‌شود که مقدار کمتری از آگاهی بهینه سؤال ایجاد شود و در نتیجه، ممکن است کیفیت اندازه‌گیری برای سنجش متناسب[1] به مقدار مورد انتظار نرسد. خلاصه این نتایج حاکی از آن است که، روش اکتشافی رکیس، با وجود دشواری‌هایی در شبیه‌سازی، قابل تعمیم به مدل سه پارامتری است و امکان ساخت خزانه‌های بهینه بر اساس این رویکرد وجود دارد. به عبارت دیگر، ویژگی‌های پارامترهای سؤال‌ها نشان می‌دهد که با استفاده از ایده رکیس و شیوه‌های ایجاد سؤال بهینه، می‌توان خزانه‌های بهینه‌ای بر اساس مدل سه پارامتری شبیه‌سازی کرد. ویژگی‌هایی که با منطق عملیاتی خزانه سؤال سنجش انطباقی کامپیوتری (CAT) منطبق است. با اینکه این نوع شبیه‌سازی‌ها بسیار وقت‌گیر است و نیاز به علم برنامه‌نویسی دقیق دارد، وجود آنها برای توسعه علم سنجش انطباقی لازم است. البته این امر بدیهی است که این تعمیم در پژوهش گو و رکیس (۲۰۰۷) موفقیت‌آمیز بوده است، ولی برنامه‌های این پژوهش در دسترس پژوهشگران داخلی ما قرار نگرفته است. درصورتی‌که برنامه‌ریزی قبل از اجرا و شبیه‌سازی خزانه‌هایی که برای سنجش انطباقی کامپیوتری (CAT) بهینه باشد می‌تواند از هزینه‌های فراوان طراحی سؤال برای سنجش انطباقی کامپیوتری (CAT) بکاهد.

همان‌طور که نتایج نیز نشان می‌دهد، خزانه سؤال عملیاتی اریب[2] و خطای[3] قابل توجهی برای آزمودنی‌هایی که در سطوح بالا و پایین توانایی قرار می‌گیرند ایجاد می‌کند، که دلیل آن کمبود سؤال‌هایی با مقادیر b بالا و پایین است. در حالی که خزانه‌های بهینه ایجاد شده چنین مشکلی ندارند. مطابق با نتایج، تعداد سؤال‌های بیش مواجهه شده در خزانه‌های بهینه‌ای که بدون روش کنترل مواجهه ساخته شدند، با خزانه عملیاتی مشابه است. در حالی‌که، آنها دارای سؤال‌های کمتری نسبت به خزانه عملیاتی هستند. امّا اگر نسبت سؤال‌های بیش مواجهه شده با تقسیم این تعداد بر اندازه خزانه سؤال ارزیابی شود، همه خزانه‌های بهینه‌ای که بدون روش کنترل مواجهه شبیه‌سازی شده‌اند، نسبت به خزانه عملیاتی درصد بالاتری از سؤال‌های بیش مواجهه شده دارند. این نتیجه حاکی از این است که، این نوع خزانه‌های بهینه به طور بالقوه، نگرانی بیشتری در مورد مسائل امنیتی آزمون مطرح می‌کنند. در واقع، این نتیجه می‌تواند به عواملی از قبیل کاربرد روش انتخاب سؤال مبتنی بر بیشینه آگاهی و اجرا نکردن شیوه کنترل مواجهه سؤال در برنامه

1.tailored testing
2.bias
3.error

سنجش انطباقی کامپیوتری (CAT) نسبت داده شود. همچنان که در پیشینه پژوهشی مربوط به این حوزه نیز اشاره شده است (برای مثال، واینر و همکاران، ۲۰۰۰؛ وای[1]، ۱۹۸۸، وای، استفان و اندرسون[2] ۱۹۸۸) قاعده انتخاب سؤال بر اساس روش بیشینه درست آگاهی، حتی به تفاوت‌های خیلی کوچک در آگاهی سؤال بسیار حساس است. بنابراین، اگر بیشینه آگاهی به عنوان ملاکی برای انتخاب سؤال در سنجش انطباقی کامپیوتری (CAT) به کار رود و هیچ نوع روش کنترل مواجهه سؤال نیز روی آن اجرا نشود، همیشه با احتمال زیاد سؤال‌هایی با ضریب تشخیص بالا در معرض بیش مواجهه قرار می‌گیرند. در حالی‌که، بسیاری از سؤال‌ها با ضریب تشخیص پایین یا حتی متوسط انتخاب نمی‌شوند. یک راه حل برای فائق آمدن بر مشکل نرخ مواجهه غیرمتعادل سؤال که به دلیل کاربرد روش بیشینه آگاهی ایجاد می‌شود، کاربرد و اجرای شیوه کنترل مواجهه سؤال است. در این مطالعه، از بین روش‌های کنترل مواجهه سؤال، روش سیمپسون–هتر انتخاب شد. به منظور مطالعه اثرات این روش، یک بار خزانه‌های سؤال، بدون اجرای این روش و بار دیگر با وارد کردن این روش طراحی شد. در حالتی که روش سیمپسون–هتر در برنامه سنجش انطباقی کامپیوتری (CAT) وارد نشد، بیشینه مواجهه سؤال به عنوان نتیجه‌ای طبیعی افزایش یافت. امّا در حالتی که روش سیمپسون–هتر به کار گرفته شد، بیش مواجهه کاهش یافت. با این وجود، در هر دو حالت، خزانه‌های بهینه درصد بسیار پایینی از سؤال‌های کم مواجهه نسبت به خزانه عملیاتی داشتند. همچنین، خزانه‌های بهینه در حالتی که روش کنترل مواجهه در طراحی خزانه وارد شد، میزان نرخ سؤال‌های بیش مواجههٔ کمتری داشت. همان طور که نتایج نشان داد، بدون توجه به عامل کنترل مواجهه سیمپسون– هتر، خزانه‌های سؤال بهینه از نظر اندازهٔ خزانه، دقت و صحت اندازه‌گیری، بهتر از خزانه عملیاتی عمل می‌کنند. البته، خزانه عملیاتی در مقایسه با خزانه‌های بهینه‌ای که روش سیمپسون–هتر برای آنها اعمال نشده است، نرخ مواجهه سؤال پایین‌تری داشت. ولی با در نظر گرفتن عامل کنترل مواجهه، خزانه‌های بهینه دارای امنیت بالاتر و نرخ مواجهه پایین‌تری داشت. با این وجود، خزانه عملیاتی، در هر دو حالت کنترل و عدم کنترل مواجهه سؤال، در سطوح متوسط توانایی نرخ همپوشی آزمون بیشتری نسبت به خزانه‌های بهینه دارد. این موضوع می‌تواند به این دلیل باشد که در خزانه عملیاتی، سؤال‌هایی با درجهٔ دشواری متوسط

1.Way
2.Steffen & Anderson

ضرایب تشخیص بالایی دارند. نرخ همپوشی بالا در خزانهٔ عملیاتی، امنیت آزمون سنجش انطباقی کامپیوتری (CAT) عملیاتی را با خطر رو به رو می‌کند. خزانهٔ بهینه R از نظر نرخ مواجهه و نرخ همپوشی آزمون، تشابه زیادی با خزانه عملیاتی دارد. این قضیه حاکی از آن است که طراحان در ساخت سؤال‌ها برای خزانه‌های سؤال سنجش انطباقی کامپیوتری (CAT) اغلب به صورت تصادفی عمل می‌کنند و این به دلیل غیر قابل استفاده ماندن بسیاری از سؤال‌ها در خزانه سؤال باعث اتلاف هزینه می‌شود. در میان دو عاملی که در طراحی خزانه‌های سؤال بهینه دستکاری شد، روش کنترل مواجهه سؤال، بر اندازهٔ خزانه سؤال، میزان همپوشی آزمون‌ها، درصد سؤال‌های بیش مواجهه و کم مواجهه و کجی نرخ مواجهه تأثیر گذاشته است. این درحالی است که روش‌های ایجاد سؤال بر متوسط مقادیر پارامتر a تاثیرگذار بوده است. بنابر این، روش سیمپسون–هتر با همه دشواری‌هایی که در برنامه‌نویسی دارد، می‌تواند در رویکرد اکتشافی رکیس وارد شود و از مواجهه بیش از حد سؤال جلوگیری کند. این روش امنیت آزمون را نیز تأمین می‌کند.

نتایج ارزیابی عملکرد خزانه‌ها، با استفاده از ملاک‌های تجربی ارزیابی، نشان داد که صرف نظر از عامل کنترل مواجهه سیمپسون–هتر، دقت و صحت اندازه‌گیری در خزانه‌های بهینه بیشتراست و کارایی نسبی بالاتری نسبت به خزانه عملیاتی دارد. همچنین، خزانه‌های بهینه، استفاده متعادل‌تری از سؤال‌های خزانه خود دارند. جالب است که خزانه عملیاتی، شامل سؤال‌های بیشتری نسبت به هر یک از خزانه‌های بهینه است، ولی نرخ همپوشی بالاتری نسبت به خزانه‌های بهینه دارد[1]. همچنین، نتایج نشان داد زمانی که عامل کنترل مواجهه سیمپسون–هتر در شبیه‌سازی خزانه‌های بهینه وارد می‌شود، صرف نظر از روش ایجاد سؤال، دقت و صحت اندازه‌گیری در برآورد توانایی افزایش پیدا می‌کند[2]. زیرا سؤال‌ها به صورت متعادل‌تری استفاده می‌شود و همهٔ سطوح توانایی دقت اندازه‌گیری یکسانی خواهند داشت. به عبارت دیگر، بررسی دقت اندازه‌گیری در هر یک از سطوح توانایی و بررسی نمودارهای مربوط به آگاهی شرطی

۱. در حالتی که برنامه بر اساس روش بدون کنترل مواجهه نوشته شده است، تعداد سؤال‌های خزانه‌های بهینه، تقریباً کمتر از یک سوم سؤال‌های خزانه عملیاتی است.

۲. در حالتی که برنامه بر اساس روش کنترل مواجهه سیمپسون–هتر نوشته شده است، تعداد سؤال‌های خزانه‌های بهینه تقریباً کمتر از یک دوم سؤال‌های خزانه عملیاتی است.

آزمون، نشان می‌دهد که، خزانه‌های سؤال بهینه‌ای که با کنترل مواجههٔ سیمپسون–هتر طراحی می‌شوند، از نظر دقت اندازه‌گیری و امنیت آزمون نسبت به خزانه سؤال عملیاتی و خزانه‌های بهینه‌ای که بدون کنترل مواجهه طراحی می‌شوند، بهتر عمل می‌کنند. همچنین، در بیشتر سطوح توانایی نیز آگاهی بیشتری دارند، زیرا نرخ‌های مواجهه برای همه سؤال‌ها را در حدود و یا پایین‌تر از نرخ مواجهه هدف (۰/۳۳) کنترل می‌کنند. این نتیجه به دلیل این است که سؤال‌های اضافه شده به خزانه‌های بهینه با کنترل مواجهه سیمپسون–هتر، سؤال‌هایی با ضرایب تشخیص بالاتری دارند. بدون تردید، زمانی که، این مولفه در طراحی خزانه سؤال اضافه می‌شود، اندازه خزانه سؤال بزرگ‌تر می‌شود. به طور کلی، بدون توجه به عامل کنترل مواجهه سؤال، خزانه (MRP) در طول سطوح توانایی میزان آگاهی بیشتری ایجاد می‌کند. همچنین، دقت اندازه‌گیری خزانه (MRP) از دو نوع خزانه بهینه دیگر (R و MTI) و خزانه عملیاتی بیشتر است. رویکرد (MTI) خزانه‌های سؤال کوچک‌تری را ایجاد می‌کند که شامل سؤال‌هایی با پارامترهای a کوچک‌تری است. در اصل روش (MTI) خزانه‌هایی با توزیع یکنواخت‌تر ضریب تشخیص در تمام سطوح دشواری سؤال ایجاد می‌کند. در این روش به دلیل اینکه متوسط آگاهی آزمون، به عنوان حداقل آگاهی آزمون در شبیه‌سازی وارد می‌شود، خزانه سؤالی با ضریب تشخیص کمتر نسبت به دو روش (R و MRP) ایجاد می‌کند. امّا روش (MRP) به دلیل اینکه در ایجاد سؤال‌ها به رابطهٔ بین دو پارامتر a و b در سطوح متفاوت توانایی توجه دارد، می‌تواند بسته به سطح دشواری سؤال، ضرایب تشخیص مجزایی را در نظر بگیرد. در این پژوهش این روش توانست در سطوح دشواری بالا، سؤال‌هایی با ضریب تشخیص بیشتر ایجاد کند. در عمل نیز، برای طراحان سؤال، ساخت سؤال‌های دشوار که ضریب تشخیص بالایی داشته باشند، آسان‌تر است. در واقع، خزانه‌های بهینه‌ای با مقادیر متوسط شیب بالاتر ((MRP(ROP_2),(ROP_5)) نسبت به سایر خزانه‌های بهینه و خزانه عملیاتی دقت و صحت اندازه‌گیری بهتری دارند. این نتایج با تحقیقات گو و رِکیس، ۲۰۰۷؛ هی، ۲۰۱۰ ؛ هی و رِکیس ۲۰۱۱ همسو است.

همچنین، نتایج نشان داد که اندازه خزانه سؤال به توزیع جمعیت آزمودنی‌ها و تعداد افرادی که از آنها سنجش انطباقی کامپیوتری (CAT) گرفته می‌شود بستگی دارد. شکل خزانه سؤال طراحی شده، توزیع نرمالی از مقادیر پارامتر دشواری نیست، بلکه توزیعی

مسطح و یکنواخت است و فراوانی‌های کاملاً بالایی در دنباله‌های توزیع دارد. این نتایج نیز با تحقیقات رکیس (۲۰۰۷؛ ۲۰۱۰) همسو است. این امر تأییدی بر این قضیه است که برخلاف آزمون‌های سنتی مداد کاغذی، که انتخاب سؤال‌ها به نوعی است که بهترین سنجش را برای آزمودنی‌هایی با توانایی متوسط فراهم می‌کند، سنجش انطباقی می‌تواند دامنه گسترده‌ای از توانایی را پوشش دهد. از این‌رو به سؤال‌هایی با کیفیت بالا برای دامنه گسترده‌ای از توانایی نیاز دارد (میلمن و آرتر، ۱۹۸٤). البته، طراحی خزانه سؤال برای هر سنجش انطباقی بسیار خاص است و به طراحی شیوه و ویژگی‌های مورد نیاز در سنجش انطباقی کامپیوتری (CAT) و جمعیت آزمودنی بستگی دارد (رکیس، ۲۰۰۱، ۲۰۰۷). در کل، این روش بسیار کلی است و می‌تواند در شکل‌های دیگری از توزیع‌های آزمودنی و شکل‌های متفاوتی از مدل‌های (IRT) نیز به‌کار رود (گو و رکیس، ۲۰۰۷). بنابر این، نتایج حاکی از آن است که گسترش روش رکیس (۲۰۰۳)، بخوبی در طراحی خزانه سؤال بهینه در موقعیت‌های ویژه کار می‌کند و در مقایسه با روش برنامه‌نویسی ریاضی، شیوه سنجش انطباقی کامپیوتری (CAT) را به‌طور سرراست‌تری شبیه‌سازی می‌کند و فرآیند برآورد توانایی در آن انعطاف‌پذیرتر است (گو، ۲۰۰۷). این روش بر تصادفی‌سازی پارامترهای سؤال در شبیه‌سازی سنجش انطباقی کامپیوتری (CAT) تأکید دارد (هی، رکیس، ۲۰۱۱). کاربرد این شیوه، طرح‌های سودمندی را ایجاد می‌کند که مزیت سنجش انطباقی کامپیوتری (CAT) را از بین نمی‌برد (رکیس، ۲۰۱۰).

در مجموع، یافته‌های آماری نشان می‌دهد که روش‌شناسی ایجاد شده در این پژوهش می‌تواند خزانه‌های سؤالی با مشخصات بهینه ایجاد کند. این خزانه‌ها قادرند با استفاده از کارکرد مناسب الگوریتم انتخاب سؤال، آزمونی مناسب برای آزمودنی اجرا کنند و در زمان یکسانی، با دقت و صحت مناسبی توانایی آزمودنی‌ها را برآورد کنند. مشخصات خزانه سؤال بهینه می‌تواند به چندین هدف در ساخت خزانه‌های سؤال عملیاتی کمک کند. ابتدا، این مشخصات می‌توانند به عنوان یک مدل طراحی خزانه سؤال و یک راهنما برای سرهم کردن خزانه سؤال عملیاتی مفید باشند. این مدل این اطمینان را ایجاد می‌کند که در طراحی الگوریتم سنجش انطباقی کامپیوتری (CAT) مورد نظر، نه تنها بهترین کیفیت اندازه‌گیری برقرار شده است (مک‌برید و وایس، ۱۹۷٦)، بلکه این توانایی را ایجاد می‌کند که چندین خزانه سؤال بهینهٔ هم ارز طراحی شود (هی و رکیس، ۲۰۱۱). دوم، ویژگی‌های خزانه‌های بهینه ساخته شده، می‌تواند بینش‌هایی در مدیریت خزانه سؤال عملیاتی ایجاد کنند. سوم، ویژگی‌های خزانه سؤال،

به خصوص، توزیع ویژگی‌های آماری سؤال، می‌تواند به عنوان راهنمایی برای فرآیند نوشتن سؤال به کار روند. این راهنما به نویسندگان سؤال آموزش می‌دهد تا سؤال‌هایی با خصیصه‌های مطلوب مبتنی بر مدل یا طرح مشخص شده‌ای بنویسند (هی و رِکیس، ۲۰۱۱).

در این پژوهش به یکی از مؤلفه‌های دیگر امنیت آزمون که همان تعادل محتوایی آزمون است توجه نشده است. امید است بتوان در پژوهش‌های بعدی، این مؤلفه مهم در طراحی خزانه سؤال بهینه وارد شبیه‌سازی شود. همچنین، تلفیق دو رویکرد برنامه‌نویسی ریاضی و رویکرد اکتشافی در طراحی خزانه‌های بهینه، یکی از پیشنهادهایی است که برای پژوهش‌های آینده توصیه می‌شود.

References

- Ariel, A.; Veldkamp, B. P. & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement*, 41, 345-360.
- 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.
- Belov, D. I. & Armstrong, R. D. (2009). Direct and inverse problems of item pool design forcomputerized adaptive testing. *Educational and Psychological Measurement*, 69 (4), 53-547.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee‘s ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Chang, H. (2007). Book review: Linear models for optimal test design. *Psychometrika*, 72, 279-281.
- 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.
- De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
- Gu, L. (2007). *Designing optimal item pools for computerized adaptive tests with exposure controls*. Unpublished doctoral dissertaion. 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.
- Guide. M. U. S. (2014). The mathwork. Lnc. Natich, MA, 5,333.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park CA: Sage.

- 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.
- He. W., & Reckase, M. (2011). *Optimal item pool design for a highly constrained computerized adaptive test*. Paper presented at the National Council on Measurement in Education, Denver, CO.
- Jensema, C. J. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement*, 1, 111-120.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- 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.
- 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., Davey, T., & Nering, M. (1998). *Test development exposure control for adaptive tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego CA.
- Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice*, 8 (3), 11-15.
- Reckase, M. D. (2001). *Item pool design for computerized adaptive tests*. Invited small group session at the 6th Conference of the European Association of Psychological Assessment, Aachen, Germany.
- 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. (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. (2010). Designing Item Pools to Optimize the Functioning of Computerized Adaptive Test. *Psychological Test and Assessment Modeling*, 52 (2), 127-141.

- Reckase, M. D. & He, W. (2004). *The ideal item pool for the NCLEX-RN examination*--Report to NCSBN: Michigan State University, East Lansing, MI.
- Reckase, M. D. & He, W. (2005). *Ideal item pool design for the NCLEX-RN exam.* Michigan State University, East Lansing, MI.
- Reckase, M. D. & He, W. (2008). *The impact of item disclosure (compromise) on the probability of passing of the NCLEX-RN exam*--report to the National Council of State Boards of Nursing (NCSBN): Michigan State University.
- 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.
- Robin, F.; van der Linden, W. J.; Eignor, D. R.; Steffen, M. & Stocking, M. L. (2005). *A comparison of two procedures for constrained adaptive test construction (ETS Research Rep No. RR-04-39)*. Princeton, NJ: Educational Testing Service.
- Stocking, M. L. & Swanson, L. (1998). Optimal design of item pools for computerized adaptive tests. *Applied Psychological Measurement*, 22, 271-279.
- Stocking, M. L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice (pp.163-182)*. Netherlands: Kluwer Academic Publishers.
- Sympson, J. B. & Hetter, R. D. (1985). *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.
- Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 14, 181-196.
- Van der Linden, W. J. (2000a). Constrained adaptive testing with shadow tests. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer Academic Publishers.
- Van der Linden, W. J. (2000 b). Optimal assembly of tests with item sets. *Applied Psychological Measurement*, 24, 225–240.0.
- Van der Linden, W. J. (2005a). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42, 283-302.

- Van der Linden, W. J. (2005b). *Linear models for optimal test design.* New York: Springer-Verlag.
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259-270.
- 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 PsychologicalMeasurement*, 23, 195-210.
- Van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Boston: Kluwer.
- Van der Linden, W. J., Adelaide, Ariel, & Veldkamp, B. P. (2006). Assembling a computerizedadaptive testing item pool as a set of linear tests. *Journal of Educational and BehavioralStatistics*, 31(1), 81-100.
- Van der Linden, W. J. & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York: Springer.
- Veldkamp, B. P., & van der Linden, W. J. (1999). *Designing item pools for computerizedadaptive testing*. (Research report 99-0). Enschede, the Netherlands: Twente University, Faculty of Educational Science and Technology.
- Veldkamp, B. P. & van der Linden, W. J. (2000). Designing item pools for computerizedadaptive testing. In van der Linden, W. J. & Glas, C. A. W. (Eds.) (2000). *ComputerizedAdaptive Testing: Theory and Practice*. Dordrecht: Kluwer.
- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg,L., & Thissen, D. (2000).Computerized 1998). *Developing, maintaining, and renewing the item inventory adaptive testing: a primer* (2nd edition). Mahwah, NJ: Lawrence Erlbaum.
- Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement, Issues and Practice*, 17, 17-27.
- Way, W.D., Steffen, M., & Anderson, G.S. *to support computer-based testing*. Paper presented at the colloquium on computer-based testing: Building the foundation for future assessments, Philadelphia, PA.