فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
دوره هفتم، شماره ۲۶، زمستان ۹۵، ص ۱۴۹ تا ۱۹۷

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 7, No. 26, Winter 2016

تلفیق رویکرد ریکیسی و رویکرد برنامه‌نویسی ریاضی در طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری

مریم مقدسین[۱]

تاریخ دریافت: ۹۵/۰۱/۱۴
تاریخ پذیرش: ۹۵/۱۰/۱۵

چکیده

سنجش انطباقی کامپیوتری (CAT) شیوه‌ای از سنجش توانایی است که دقت برآورد توانایی را افزایش می‌دهد و بدون از دست دادن دقت اندازه‌گیری آزمون، طول آن را کاهش می‌دهد. بااین‌وجود، سنجش انطباقی در صورتی خوب عمل می‌کند که دارای خزانه سؤالی باشد که در آن تعداد کافی سؤال با کیفیت مناسب وجود داشته باشد. بسیاری از محققان خاطرنشان کردند که برای ساخت خزانه سؤالی برای (CAT)، نه‌تنها اندازه خزانه سؤال مهم است، بلکه توزیع پارامترهای سؤال‌های خزانه نیز از اهمیت به‌سزایی برخوردار است. بااین‌وجود، تحقیقات اندکی در مورد این‌که چگونه این ویژگی‌های مطلوب تعیین می‌شود، وجود دارد. هدف اصلی این پژوهش، تلفیق ایده "bin-and-union" برگرفته از رویکرد ریکیسی (۲۰۰۳) که یک روش شبیه‌سازی مونت‌کارلو برای تعیین ویژگی‌های خزانه سؤال است، با رویکرد برنامه‌نویسی ریاضی بوده است. در پژوهش فعلی، این روش برای ساخت یک خزانه سؤال بهینه برای آزمون سنجش انطباقی ریاضی به‌کار رفته است. خزانه سؤال پژوهش حاضر بر اساس مدل سه پارامتری مدرج شده است و روش سیمپسون-هتر برای کنترل مواجهه سؤال به کار رفته است. این طرح شامل برآوردهایی از اندازه مطلوب خزانه سؤال و توزیع مطلوب پارامترهای سؤال‌ها و ویژگی‌های غیر آماری آن بوده است. فرآیند طراحی این خزانه شامل تعیین مجموعه‌ای از ویژگی‌های مطلوب خزانه سؤال، با در نظر گرفتن چندین عامل مهمی که ممکن بود بر نتایج موردنظر طراحی یک خزانه سؤال اثر گذارد، بوده است. عملکرد خزانه‌های سؤال شبیه‌سازی‌شده و عملیاتی با در نظر گرفتن مجموعه‌ای از ملاک‌های

۱. استادیار گروه روانشناسی بالینی، دانشگاه خوارزمی (نویسنده مسئول) mmoghadasin@yahoo.com

ارزیابی، با یکدیگر مورد مقایسه قرار گرفته‌اند. نتایج ارزیابی نشان داد که مکانیزم به‌کاررفته برای تعیین ویژگی‌های مطلوب خزانه سؤال به‌خوبی عمل می‌کند و برای تعیین ویژگی‌های مطلوب خزانه سؤال مناسب است.

واژگان کلیدی: خزانه سؤال بهینه، سنجش انطباقی کامپیوتری، رویکرد ریکیسی، رویکرد برنامه‌نویسی ریاضی، مدل حداقل انحرافات وزن‌دار (WDM)

مقدمه

کیفیت خزانه سؤال، به‌عنوان یک عامل مهم برای افزایش کیفیت اندازه‌گیری در سنجش انطباقی کامپیوتری (CAT)، در نظر گرفته شده است (فلاگر[1]، ۲۰۰۰؛ جنسما[2]، ۱۹۷۷؛ مک‌برید و وایس[3]، ۱۹۷۶؛ ریکیسی[4]، ۱۹۷۶؛ ۲۰۰۳؛ وندرلیندن، ادلاید و ولدکمپ[5]، ۲۰۰۶؛ ولدکمپ و وندرلیندن، ۲۰۰۰؛ اکسینگ و همبلتون[6]، ۲۰۰۴). حتی در همان اوایل دهه ۱۹۷۰ – آغاز پژوهش‌های مرتبط با CAT- محققان به‌طور ضمنی و یا به‌صراحت اذعان داشتند که ویژگی‌های خزانه سؤال نقش مهمی در دستیابی به بهترین نتایج ممکن در سنجش انطباقی، خواهد داشت (مک‌برید و وایس، ۱۹۷۶). بااین‌وجود، دستورالعمل‌های اندکی در مورد چگونگی ساخت یک خزانه سؤال با کیفیت بالا، برای سنجش انطباقی کامپیوتری قابل‌دستیابی بودند (هی[7]، ریکیسی، ۲۰۱۰؛ هی و ریکیسی، ۲۰۱۱).

مطالعات قبلی در مورد طراحی خزانه سؤال بهینه برای CAT روی دو رویکرد عمده مبتنی هستند. رویکرد اول، توسط ولدکمپ و وندرلیندن (۲۰۰۰) ایجاد شده است، این رویکرد از روش برنامه‌نویسی ریاضی[8] برای طراحی خزانه سؤال استفاده می‌کند. در این

1. Flaugher
2. Jensema
3. McBride & Weiss
4. Reckase
5. Van der linden, Adelaide & Weldkamp
6. Xing & Hambelton
7. Hi
8. mathematical programming

رویکرد، CAT از طریق رویکرد تست سایه[۱]، اجرا می‌شود (وندرلیندن، ریس[۲]، ۱۹۸۸) که تست از طریق برنامه‌نویسی عدد صحیح خطی دو ارزشی[۳] یا برنامه‌نویسی ۱-۰ تشکیل می‌شود. در این رویکرد ویژگی‌های تست بر اساس یک تابع هدف[۴] که در ارتباط با مجموعه‌ای از قیود[۵] خاص بیشینه می‌شود، محقق می‌گردد (وندرلیندن، ۲۰۰۵a). بااین‌وجود، یک محدودیت بالقوه این رویکرد آن است که به نرم‌افزارهای جبر خطی از قبیل CPLEX و LINDO برای به‌دست آوردن راه‌حل بهینه نیاز دارد که کاربرد این روش را کمی دشوار می‌کند و ممکن است، کدها و معادلات آن برای اکثریت کاربران در دسترس نباشد که در این صورت اگر برنامه نیاز به اصلاح و یا تغییر داشته باشد، کنترلی بر آن نداشته باشند و چه‌بسا این احتمال وجود دارد که همیشه راه‌حل قابل‌اجرا و عملی[۶] در دسترس نباشد (چانگ[۷]، ۲۰۰۷؛ روبین[۸] و همکارانش، ۲۰۰۵). همچنین در این رویکرد فرض بر این است که سؤال‌ها از قبل در خزانه موجود هستند و از روی آن‌ها یک خزانه کوچک‌تر تشکیل می‌شود (گو و ریکیسی، ۲۰۰۷). در این رویکرد از ویژگی‌های یک خزانه سؤال موجود به‌عنوان نقطه شروع استفاده می‌شود (ریکیسی، ۲۰۱۰). به دلیل این‌که اجرای روش تست سایه در مقیاس بزرگ، به مسئله‌ای بسیار دشوار تبدیل می‌شود، بنابراین، حل مسائل سرهم کردن تست را با مشکل روبرو می‌کند. به‌منظور دوری از مشکلات روش تست سایه، از مدل حداقل انحرافات وزن‌دار[۹] که نوعی روش برنامه‌نویسی خطی است،

1. Shadow a test (SAT)

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

2. Reese
3. binary linear integer programming
4. objective function
5. constraints
6. feasible
7. Chang
8. Robin
9. weighted deviations model (WDM)

به‌عنوان جایگزین استفاده می‌شود. این روش برای سرهم کردن تست‌های چندگانه و تست‌های سنجش انطباقی به روش مؤثرتری عمل می‌کند. در این روش، ابتدا پیش‌بینی جستجوی راه‌حل برای تست کامل صورت می‌گیرد و همزمان هم قابل‌حل بودن و هم بهینه بودن تست را در نظر می‌گیرد. این روش جزء روش‌های شهودی[1] حل مسائل سرهم کردن تست است (وندرلیندن، ۲۰۰۵a، ۲۰۰۵b). این روش در اصل توسط استوکینگ و سوانسون[2]، ۱۹۹۳ به دلیل علاقه و نگرانی آن‌ها در مورد کیفیت ضعیف خزانه‌های سؤال در سرهم کردن تست‌های متوالی در مقیاس بزرگ ایجاد شد. مدل حداقل انحرافات وزن‌دار به‌صراحت ویژگی‌های آماری و غیر آماری سؤال را با تعادل مطلوبی بین ویژگی‌های اندازه‌گیری و ساختاری در نظر می‌گیرد. این ویژگی‌ها به‌وسیله وزن‌هایی که توسط طراحان تست انتخاب می‌شود، در مدل وارد می‌گردد. این روش برخلاف روش تست سایه، ویژگی‌های محتوایی را به‌عنوان اهداف[3] (به‌جای قیود) فرمول‌بندی می‌کند. انحراف از اهداف محتوایی وزن داده می‌شود و در تابع هدف به همراه فاصله آگاهی سؤال از مقدار هدف[4] قرار می‌گیرد (استوکینگ، سوانسون و پیرمن[5]، ۱۹۹۳). البته این روش ابزاری بوده است که در بسیاری از رویکردهای برنامه‌نویسی ریاضی استفاده شده است و مدل‌های غیرقابل‌حل را قابل‌اجرا می‌کرده است (بروک، کندریک و مروس[6]، ۱۹۹۸). در CAT، مدل حداقل انحرافات وزن‌دار سؤال‌هایی را انتخاب می‌کند که به‌طور متوالی کوچک‌ترین مجموع انحرافات وزن‌دار را دارد.

رویکرد دوم، رویکرد اکتشافی[7] ریکیسی است (ریکیسی، ۲۰۰۳)، این رویکرد برخلاف روش برنامه‌نویسی ریاضی، بسیار سراست است. همچنین، در مطالعات گوناگون در مورد طراحی خزانه‌های سؤال بهینه برای CAT استفاده شده است (ریکیسی، ۲۰۰۳،

1. Heuristics
2. Stocking & Swanson
3. goal
4. target
5. Pearlman
6. Brooke, Kendrick & Meeraus
7. heuristic

ریکیسی و هی، ۲۰۰۴، ۲۰۰۵، ۲۰۰۹a، ۲۰۰۹b؛ گو[1] و ریکیسی، ۲۰۰۷). در این رویکرد، استفاده از برنامه‌ریزی اعداد صحیح کنار گذاشته شده است و در آن فرض نمی‌شود که سؤال‌ها از قبل در خزانه وجود دارد. در عوض، در این رویکرد سؤال‌ها برحسب پارامترهای IRT شبیه‌سازی می‌شود تا با برآوردهای اخیر توانایی مطابقت داشته باشد و میزان آگاهی به‌اندازه کافی بهینه‌ای را ایجاد کند. در روش ریکیسی ابتدا، خزانه سؤال هدف بر اساس صفات غیر آماری از قبیل محتوا به خزانه‌های کوچک‌تری تقسیم‌بندی می‌شود. سپس فرآیند CAT شبیه‌سازی می‌شود، به‌طوری‌که خزانه‌های سؤال کوچک‌تر به‌طور همزمان ساخته می‌شوند. شبیه‌سازی با یک آزمودنی که به‌طور تصادفی از توزیع مورد انتظار جامعه فرضی در دامنه مشخص‌شده انتخاب می‌شود، آغازشده و CAT برای او اجرا می‌گردد. هر سؤال به نحوی شبیه‌سازی می‌شود که سؤال بهینه‌ای بر اساس برآورد جدید توانایی آزمودنی باشد. فرایند مشابهی برای آزمودنی بعدی نیز تکرار می‌شود، سپس، به همین ترتیب، برای کل نمونه موردنظر این فرآیند ادامه می‌یابد و سؤال‌ها برای نمونه بزرگی از آزمودنی‌ها شبیه‌سازی می‌شود و به خزانه سؤال اضافه می‌گردند و بدین ترتیب بر اساس این روش که به آن روش "bin-and-union" نیز گفته می‌شود، خزانه سؤال بهینه ساخته می‌شود (ریکیسی، ۲۰۰۳، ۲۰۰۹؛ ریکیسی و هی، ۲۰۰۴، ۲۰۰۹a). برخلاف مسئله‌ی سرهم کردن[2] خزانه سؤال در رویکرد اول که در آن یک خزانه سؤال از یک خزانه بزرگ[3] در دسترس بر طبق ویژگی‌های مطلوب سرهم می‌شود (وندرلیندن، آریل و ولدکمپ، ۲۰۰۶، وندرلیندن، ۲۰۰۰a، ۲۰۰۰b، ۲۰۰۰c،۲۰۰۵a، ۲۰۰۵b)، در مسئله طراحی خزانه سؤال در رویکرد دوم، فرض بر این است که هیچ سؤال واقعی در دسترس نیست. ازاین‌رو، ازآنجایی‌که در عمل نیز، زمانی که یک خزانه سؤال طراحی می‌شود، هیچ سؤال واقعی در دسترس نیست، طبیعتاً طراحی یک خزانه سؤالی که بهینه باشد، هدف مطلوبی است (هی، ریکیسی، ۲۰۱۰). در رویکرد اکتشافی، طراحی خزانه

1. Gu
2. assemble
3. master

سؤال بهینه به این صورت نیست که از قبل خزانه بزرگی در دسترس باشد، بلکه هیچ سؤال واقعی در دسترس نیست و در آن از روش مونت‌کارلو برای تعیین ویژگی‌های یک خزانه سؤال بهینه استفاده می‌شود (گو و ریکیسی، ۲۰۰۷).

ادبیات پژوهشی متعدد در مورد طراحی خزانه‌های سؤال بهینه، هر یک از این رویکردها را به‌صورت مجزا به کار برده‌اند، حال اگر روشی ایجاد شود که از مزایای هر دو رویکرد به‌طور همزمان استفاده کند، می‌تواند هم از طریق کمّی سازی ویژگی‌های آماری و غیر آماری به‌عنوان قیود تست که از مزایای روش برنامه‌نویسی ریاضی است و هم از تمام مزایای رویکرد ریکیسی استفاده کند و خزانه‌ای بهینه ایجاد کند. حال سؤال مهمی که مطرح می‌شود، این است که برای طراحی یک خزانه سؤال بهینه چه تلاشی باید صورت گیرد؟ بدیهی است که در طراحی یک خزانه سؤال باید ویژگی‌های آماری و غیر آماری سؤال در نظر گرفته شود. برای مثال، توزیع پارامترهای سؤال مطلوب باید چگونه باشد؟ سؤال‌ها موجود در خزانه سؤال CAT باید چه صفاتی داشته باشند؟ از طرف دیگر، پرسش‌هایی از قبیل؛ چه چیزی باعث می‌شود که اندازه خزانه سؤال کافی باشد یا به عبارتی به چند سؤال در خزانه نیاز داریم؟ نیز باید در نظر گرفته شود؛ بنابراین، به‌طور خلاصه، زمانی که یک خزانه سؤال بهینه طراحی می‌شود، باید حداقل سه عنصر اساسی در نظر گرفته شود، یعنی، ویژگی‌های آماری، ویژگی‌های غیر آماری و اندازه خزانه سؤال. ویژگی‌های آماری شامل پارامترهای سؤال می‌باشند، ویژگی‌های غیر آماری شامل ویژگی‌های محتوایی، توزیع کلید و مهارت‌های شناختی و غیره می‌باشند (هی و ریکیسی، ۲۰۱۰). ازاین‌رو، در این مطالعه برای طراحی خزانه سؤال بهینه از تلفیق رویکرد ریکیسی و مدل حداقل انحرافات وزن‌دار استفاده می‌شود و راهنمایی‌های مقدماتی برای تهیه‌ی یک خزانه سؤال بهینه برای یک آزمون سنجش انطباقی کامپیوتری که تحت مدل سه پارامتری مدرج شده است و روش کنترل مواجهه سیمپسون-هتر برای کنترل مواجهه بیش‌ازحد در ساخت خزانه سؤال بهینه وارد می‌شود، ارائه خواهد شد. همچنین، عملکرد این خزانه‌ها با یک خزانه سؤال عملیاتی مقایسه می‌شود.

روش‌شناسی

این پژوهش از دو قسمت نسبتاً مرتبط تشکیل شده است. در قسمت اول شبیه‌سازی‌های مونت‌کارلو ریکیسی (۲۰۰۳) و روش برنامه‌نویسی ریاضی WDM، برای طراحی خزانه‌های سؤال بهینه آزمون ریاضی استفاده می‌شود. در قسمت دوم، عملکرد خزانه‌های سؤال بهینه با خزانه سؤال عملیاتی[1] بر اساس مجموعه‌ای از ملاک‌های ارزیابی مورد مقایسه قرار گرفته است. در این پژوهش از طریق سه روش تصادفی[2] (R)، روش تصادفی آمیخته و پیش‌بینی[3] (MRP) و روش حداقل آگاهی آزمون[4] (MTI) پارامترهای سؤال خزانه‌های سؤال بهینه شبیه‌سازی شده‌اند. تلفیق روش ریکیسی (۲۰۰۳) با رویکرد برنامه‌نویسی خطی حداقل انحرافات وزن‌دار، برای طراحی خزانه‌های سؤال بهینه مدرج‌شده با مدل سه پارامتری لُجستیک به‌کار رفته است. تعادل محتوایی در هر یک از آزمون‌های CAT، از طریق مدل انتخاب سؤال حداقل انحرافات وزن‌دار (WDM) وارد برنامه‌نویسی شده است. روش کنترل مواجهه سیمپسون-هتر نیز به‌منظور کنترل مواجهه در مدل وارد شده است (سیمپسون-هتر، ۱۹۸۵). در این پژوهش، یک خزانه عملیاتی در سه حوزه‌ی محتوایی حسابان-دیفرانسیل، هندسه و جبر ساخته‌شده تا حوزه‌های تجربی معنادار توانایی را اندازه بگیرد. تست‌های عملیاتی برای هر آزمودنی، شامل ۶۰ سؤال توانایی ریاضی است و ازلحاظ محتوایی و ویژگی‌های آماری به‌عنوان مبنای طراحی خزانه سؤال بهینه در این پژوهش در نظر گرفته شده‌اند. در این قسمت، علاوه بر شرح موارد مربوط به روش‌شناسی تحقیق، شرح مختصری در مورد روش برنامه‌نویسی ریاضی (WDM) ارائه می‌شود[5].

1. Operational
2. Random Procedure (R)
3. Mixed Random and Prediction Procedure (MRP)
4. Minimum Test Information Procedure (MTI)

۵. برای اطلاع از روش "bin-and-union" رویکرد ریکیس، مفهوم bin در مدل سه پارامتری و ایجاد سؤال‌های بهینه در مدل سه پارامتری به مقاله‌ای با عنوان " طراحی خزانه‌های سؤال بهینه برای سنجش انطباقی کامپیوتری با در نظر گرفتن امنیت آزمون" که توسط نویسندگان این مقاله در مجله " مطالعات اندازه‌گیری و ارزشیابی آموزشی" دوره پنجم، شماره دهم به چاپ رسیده است، مراجعه کنید.

کاربرد مدل انتخاب سؤال حداقل انحرافات وزن‌دار (WDM) در تعیین قیود محتوایی. در این پژوهش از روش برنامه‌نویسی خطی حداقل انحرافات وزن‌دار برای تعیین محتواها و ایجاد تعادل محتوایی در خزانه‌های سؤال استفاده شده است. ابتدا، محتوای آزمون CAT، توسط متخصصان موضوعی مشخص شد و پس از توافق کامل میان ۵ متخصص، محتواها به کدهای کمّی تبدیل شدند. محتواها به سه مجموعه‌ی اصلی (حسابان-دیفرانسیل، هندسه و جبر) تقسیم‌بندی و به دنبال آن هر یک از مجموعه‌ها به زیرمجموعه‌های معین (به ترتیب، ۱۸، ۱۶ و ۱۱ محتوا) تقسیم‌بندی شدند. سپس از طریق روش برنامه‌نویسی ریاضی کدهای مربوط به هر یک از محتواها، وارد برنامه طراحی خزانه سؤال بهینه شدند. از طریق این روش تست‌های سنجش انطباقی برای ۶۰۰۰ نفر سرهم شدند. در این روش، ابتدا پیش‌بینی جستجوی راه‌حل برای تست کامل صورت می‌گیرد و همزمان هم قابل‌حل بودن و هم بهینه بودن تست در نظر گرفته می‌شود. این روش جزء روش‌های شهودی حل مسائل سرهم کردن تست است. با کاربرد مدل حداقل انحرافات وزن‌دار به‌صراحت ویژگی‌های آماری و غیر آماری سؤال‌ها با تعادل مطلوبی بین ویژگی‌های اندازه‌گیری و ساختاری در نظر گرفته می‌شود. این ویژگی‌ها از طریق وزن‌هایی که توسط طراحان تست انتخاب می‌شود، در مدل وارد می‌گردد. به‌عبارت‌دیگر، ویژگی‌های محتوایی به‌عنوان اهداف فرمول‌بندی می‌شوند. انحراف از اهداف محتوایی وزن داده می‌شود و در تابع هدف به همراه فاصله آگاهی سؤال از مقدار هدف قرار داده می‌شود. انتخاب سؤال‌ها در CAT، بر اساس رویکرد WDM طوری تنظیم می‌شود که سؤال‌هایی انتخاب شوند که به‌طور متوالی کوچک‌ترین مجموع انحرافات وزن‌دار را داشته باشند. برای انتخاب یک سؤال از سه‌گام پیروی می‌شود: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، x_i متغیر تصمیم‌گیری و $i=1,..N,$ سؤال‌ها را نشان می‌دهد. اگر سؤال در تست وارد شود، $x_i=1$ و اگر سؤال از تست خارج شود $x_i=0$. همچنین، در این مدل $j=1,...,J$ صفات تست به همراه قیود غیر روان‌سنجی را نشان

می‌دهد و حدود پایین و بالای تعداد سؤال‌هایی که در آزمون دارای چنین ویژگی‌هایی هستند را به ترتیب با L_j و U_j مشخص می‌کند، 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، W_j وزن اختصاص داده‌شده به هر قید، d_{L_j} و d_{U_j} به ترتیب کسری حد پایین و مازاد حد بالا، e_{L_j} و e_{U_j}، به ترتیب، اضافی حد پایین و کسری حد بالا و d_θ انحراف از آگاهی هدف برای یک آزمودنی را نشان می‌دهد. دو جدول ۱ و ۲ به‌صورت خلاصه اطلاعات مربوط به توابع هدف و قیود مربوط به آن را نشان می‌دهد. قیود تست به‌عنوان ویژگی‌های غیر آماری یا غیر روان‌سنجی، به همراه ویژگی‌های آماری وارد شبیه‌سازی‌های روش اکتشافی مرحله قبل می‌شود. سپس، انحرافات از این قیدها برای هر یک از ۶۰۰۰ تعداد CAT که از کل خزانه بهینه سرهم می‌شود، محاسبه می‌گردد. به‌طورکلی، این مرحله تلفیقی از دو رویکرد برنامه‌نویسی ریاضی و رویکرد اکتشافی است.

جدول ۱. اطلاعات مربوط به قیود و وزن‌های آزمون‌ها در مورد به حداقل رساندن انحرافات از قیود

تابع هدف: به حداقل رساندن میزان انحرافات وزن‌دار	$\min imize \sum_{j=1}^{J} W_j d_{L_j} + \sum_{j=1}^{J} W_j d_{U_j} + W_\theta d_\theta \rightarrow (objective)$
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 ۲. اطلاعات مربوط به قیود و وزن‌های آزمون‌های CAT 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 $\sum_{i=1}^{I} x_i = 60 \rightarrow test-length$	Test lenght	N1	۲۵	۲۵
		N2	۲۰	۲۰
		N3	۱۵	۱۵
تعداد زیرمجموعه‌های تست $\sum_{i=1}^{25} z_1 = 18, \sum_{i=1}^{20} z_2 = 16, \sum_{i=1}^{15} z_3 = 11$	Number of test sets	$z_1 = \text{arithmetic}$	۱۸	۱۸
		$z_2 = \text{geometry}$	۱۶	۱۶
		$z_3 = \text{algebra}$	۱۱	۱۱
تعداد سؤال‌ها در زیرمجموعه‌های تست $\sum_{i\epsilon v_1} x_1 \le n^{(u)}{}_1 z_1, \sum_{i\epsilon v_1} x_1 \ge n_1^{(l)} z_1, \sum_{i\epsilon v_1} x_2 \le n_2^{(u)} z_2, \sum_{i\epsilon v_1} x_2 \ge n_2^{l} z_2, \sum_{i\epsilon v_1} x_3 \le n_3^{(u)} z_3, \sum_{i\epsilon v_1} x_3 \ge n_3^{(l)} z_3$	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 تجزیه‌وتحلیل: h_2 ترکیب: h_3	۱	۷

ابزارهای پژوهش حاضر از قرار زیر است:

۱). نحوه طراحی خزانه سؤال عملیاتی. در این پژوهش، خزانه سؤال عملیاتی مربوط به آزمون ریاضی است. این آزمون توانایی حل مسائل حساب، هندسه و جبر را برای دانش‌آموزانی که در سال آخر دبیرستان تحصیل و خود را برای کنکور سراسری ریاضی فیزیک آماده می‌کرده‌اند، طراحی شده است و به‌عنوان مبنایی برای محک و ارزیابی خزانه‌های سؤال شبیه‌سازی‌شده این مطالعه، به‌کاررفته است. در این پژوهش، برنامه CAT از طریق زبان PHP نوشته و از پایگاه داده MYSQL برای ذخیره‌سازی سؤال‌ها در خزانه استفاده شده است. سؤال‌های طراحی‌شده برای خزانه سؤال عملیاتی توسط نرم‌افزار BILOG مدرج شده و سؤال‌هایی که با مدل سه پارامتری لُجستیک برازش داشتند در خزانه ذخیره شدند (دی‌آیالا، ۲۰۰۹). همچنین، مفروضات IRT در مورد هریک از

سؤال‌ها بررسی شد. سؤال‌هایی که تک‌بعدی بودن و استقلال موضعی را دارا بودند، در خزانه ذخیره شدند (همبلتون و همکاران، ۱۹۹۱؛ دی‌آیالا، ۲۰۰۹). البته سؤال‌ها طوری طراحی شده بودند که هرکدام مفهوم کاملاً مستقلی نسبت به سؤال‌های دیگر بسنجند، ازاین‌رو، اغلب آن‌ها مفروضات IRT را برقرار می‌کردند؛ بنابراین، به هر یک از آن‌ها وزن محتوایی جداگانه‌ای بر اساس نظر متخصصین محتوایی و موضوعی داده شد. روش انتخاب سؤال در فرایند 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[1]، ۱۹۷۵) استفاده گردید، به‌طوری‌که پیشین مورد انتظار از توزیع نرمال پیروی می‌کرد. پس از این‌که دو پاسخ صحیح و غلط در الگوی پاسخ آزمودنی مشاهده شد، برای برآورد توانایی از شیوه بیشینه درست‌نمایی (MLE) استفاده گردید. برآورد Θ بعد از سؤال آخر، به‌عنوان نمره آزمودنی به‌حساب آمد. همچنین روش سیمپسون-هتر برای کاهش بیش مواجهه سؤال‌هایی که از میزان آگاهی بالایی برخوردار بودند استفاده شد و نرخ مواجهه هدف برابر با ۱÷۳=۰/۳۳ قرار داده شد. هر آزمون بعد از تعداد ثابتی سؤال، یعنی ۶۰ سؤال، با قیود محتوایی مشخص که در جدول ۱ و ۲ آورده شده است، به اتمام می‌رسد. در این مطالعه تعادل محتوایی به‌عنوان یک عامل مهم در ساخت خزانه سؤال بهینه در نظر گرفته شده است؛ بنابراین، تعامل بین دو عامل کنترل محتوایی سؤال‌ها و کنترل مواجهه سؤال‌های ارائه‌شده به آزمودنی‌ها موردبررسی قرارگرفته شده است. توانایی اولیه برای هر فرد روی صفر تنظیم شد و برنامه CAT به نحوی برنامه‌ریزی شد که برای همه افراد، سؤال یکسانی که پارامتر دشواری آن صفر ($b = 0$) باشد، اجرا کند. خزانه سؤال عملیاتی شامل ۹۲۱ سؤال (۴۵۵ سؤال حساب دیفرانسیل، ۲۵۸ سؤال هندسه و ۲۰۸ سؤال جبر) بود که توسط ۱۵ طراح ساخته و توسط ۱۰ طراح دیگر ارزیابی موضوعی و تخصصی گردید. سپس

1. Owen

به‌وسیله ۳ روان‌سنج که در درس ریاضی نیز تخصص داشتند بررسی و پس از رفع مشکلات محتوایی و مشکلات احتمالی گزینه‌ها، سؤال‌ها بر روی ۵۰۰ نفر از آزمودنی‌هایی که به همان جامعه تعلق و در اجرای مداد-کاغذی آن نیز شرکت داشته‌اند مدرج‌سازی گردید. این خزانه عملیاتی به شکلی طراحی شد تا ملاک‌های توصیف‌شده توسط وندرلیندن (۲۰۰۰a) را داشته باشد[۱].

۲). روش شبیه‌سازی خزانه سؤال بهینه. برنامه‌های شبیه‌سازی‌شده از طریق نسخه اصلی برنامه MATLAB (MathWorks, 2014)، به‌منظور شبیه‌سازی الگوی خزانه سؤال و ارزیابی خزانه‌های سؤال شبیه‌سازی‌شده و عملیاتی ایجاد شد. همچنین، برای جستجوی بهترین راه‌حل بهینه برای تعیین قیود محتوایی آزمون، از نرم‌افزار جبر خطی GAMS استفاده شد. شبیه‌سازی الگوی خزانه سؤال بهینه در گام‌های زیر انجام گردید:

گام اول: مدل یابی کردن شیوه‌های CAT: ازآنجاکه هدف این پژوهش ساخت خزانه سؤال بهینه برای سنجش مهارت ریاضی بوده است، در شبیه‌سازی خزانه‌های بهینه نیز تمام ویژگی‌های روان‌سنجی آزمون عملیاتی، به‌دقت وارد گردید. طول آزمون مانند آزمون‌های عملیاتی، ثابت (۶۰ سؤال) قرار داده شد. تعادل محتوایی در برنامه از طریق واردکردن بهترین وزن‌های محتوایی که توسط حل معادلات جبری به‌دست آمد، در برنامه‌ی شبیه‌ساز وارد شد. روش انتخاب سؤال‌ها روش MI به همراه جدول آگاهی در نظر گرفته شد. به‌منظور 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[۲] استفاده شد، پس از ایجاد حداقل دو پاسخ صحیح و غلط در الگوی پاسخ، از روش بیشینه درست‌نمایی استفاده گردید. همچنین، برای بررسی تعامل

---

۱. این ملاک‌ها عبارت‌اند از: الف) خزانه باید به‌اندازه کافی بزرگ باشد تا این اجازه را به ما بدهد که چندین هزار خرده آزمون همپوش از سؤال‌ها به دست آید ب) سؤال‌ها دامنه کاملی از سطوح دشواری نسبت به جامعه موردنظر 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.

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

عامل کنترل مواجهه (S-H) با تعادل محتوایی آزمون، یک شبیه‌سازی بدون واردکردن عامل S-H و شبیه‌سازی دیگر با واردکردن عامل S-H (با وجود تمام شرایط مساوی) انجام گرفت.

گام دوم: ایجاد وزن‌های محتوایی: در این مرحله تمام قیود آماری و غیر آماری سؤال‌ها، توسط متخصصان موضوعی به‌صورت کمّی تعیین شد و بهترین راه‌حل بهینه از طریق برنامه‌نویسی خطی با نرم‌افزار GAMS جستجو گردید. پس از این‌که بهترین وزن‌های محتوایی مشخص شد، در برنامه‌ی شبیه‌سازی CAT با رویکرد ریکیسی تلفیق شد.

گام سوم: ایجاد جامعه آزمون دهندگان: ازآنجاکه، خزانه سؤال عملیاتی برای آزمودنی‌هایی که فرض می‌شد، دارای توزیع توانایی نرمال هستند طراحی گردید، در شبیه‌سازی خزانه سؤال بهینه نیز از توزیع نرمال پیروی شد. دو توزیع حجم نمونه در شبیه‌سازی خزانه سؤال بهینه به کار رفت: به این معنی که خزانه‌های سؤال بهینه با یک نمونه شبیه‌سازی و با نمونه‌ای دیگر ارزیابی شدند. الف) تعداد ۶۰۰۰ توانایی (θ) از توزیع N(0,1) به‌طور تصادفی انتخاب شد و به‌عنوان توانایی واقعی آزمودنی‌ها وارد تحلیل گردید. این نمونه به‌منظور تعیین ویژگی‌های خزانه سؤال بهینه به کار رفت ب). پیوستار توانایی در دامنه ۴- تا ۴+ با فواصل ۰/۱۲۵ به ۶۵ طبقه تقسیم و در هر یک از این سطوح ۵۰۰ آزمودنی قرار گرفت (در کل ۳۲۵۰۰ آزمودنی). این نمونه به‌منظور ارزیابی عملکرد کلی شبیه‌سازی‌ها و محاسبه آماره‌های مشروط توانایی مورداستفاده قرار گرفت.

گام چهارم: ایجاد پارامترهای سؤال: برای هر آزمون 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 از طریق سه روش R، MRP و MTI تولید شدند. در هر سه روش، پارامتر c بر اساس توزیع بتا[1] تولید شد. پارامترهای a بسته به

۱. 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.

برآورد اخیر توانایی و روش ایجاد پارامتر (R، MRP و MTI)، ایجاد گردیدند و پارامترهای b طوری تولید شدند که سؤال حداکثر میزان آگاهی در برآورد توانایی جدید را ایجاد کند.

گام پنجم: ایجاد داده‌های پاسخ: پاسخ‌های آزمودنی‌ها به دنبال هر سؤالی که بر طبق مدل سه پارامتری لُجستیک ایجاد شد، تولید گردید. در مدل سه پارامتری لُجستیک احتمال اینکه آزمودنی j به سؤال i پاسخ صحیح دهد، به‌صورت زیر محاسبه می‌شود:

$$P_i(\theta_j) \equiv c_i + (1 - c_i)(1 + \exp[-1.7a_i(\theta_j - b_i)])^{-1} \quad (۱)$$

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

گام ششم: تعدیل پس از شبیه‌سازی: برای هر ترکیبی از روش‌ها و متغیرهای مستقل ۱۰ تکرار صورت گرفت تا جایی که برآورد نسبتاً ثابتی از خزانه سؤال بهینه به دست آمد. از ۱۰ تکرار الگوها و تعدادهای مواجهه سؤال، قبل انجام تعدیل پس از شبیه‌سازی، میانگین گرفته شد.

جامعه آزمون CAT عملیاتی: این آزمون برای تمام دانش‌آموزان مقطع پیش‌دانشگاهی که خود را برای کنکور سراسری ریاضی سال ۱۳۹۳ آماده می‌کردند، قابلیت اجرا داشت. البته با این فرض که توزیع این جامعه نرمال با میانگین صفر و انحراف معیار یک بوده است.

نمونه آزمون CAT عملیاتی: این آزمون بر روی ۳۵۰ نفر از دانش‌آموزانی که خود را برای آزمون کنکور سراسری ریاضی سال ۱۳۹۳ آماده می‌کردند و در کنکورهای

آزمایشی مرحله‌ای نیز شرکت داشتند، در فواصل فروردین‌ماه تا خرداد ۱۳۹۳ (قبل از کنکور ۹۳) به‌صورت آنلاین اجرا شد. این نمونه به‌صورت تصادفی سیستماتیک (ازآنجاکه لیست کاملی از اسامی شرکت‌کنندگان کنکورهای آزمایشی مرحله‌ای در دسترس بود و اسامی به‌صورت تصادفی لیست شده بودند، از این روش استفاده شد) انتخاب شد که از یک توزیع نرمال توانایی، با میانگین ۰/۱۷- و انحراف استاندارد ۰/۹۵ پیروی می‌کرد.

متغیرهای مستقل پژوهش. در همه طرح‌های خزانه سؤال دو متغیر مستقل دست‌کاری شدند. روش طراحی پارامترهای سؤال، روش کنترل مواجهه. نرخ مواجهه هدف بر اساس روش سیمپسون-هتر برابر با (۰/۳۳) قرار داده شد. این نرخ در خزانه سؤال عملیاتی نیز برابر با ۰/۳۳ قرار گرفت (برای جزئیات بیشتر جدول ۳ را ببینید).

جدول ۳. طرح شبیه‌سازی خزانه‌های سؤال بهینه

طول آزمون	۶۰	
توزیع توانایی	$N(0,1)$	
کنترل مواجهه	عدم کنترل مواجهه روش سیمپسون-هتر (نرخ مواجهه هدف ۰/۳۳)	
روش طراحی خزانه سؤال	مدل تصادفی (R) مدل تصادفی آمیخته و پیش‌بینی (MRP) مدل کمینه آگاهی آزمون (MTI)	
پهنای bin	b-bin:۰/۲ a-bin:$\Delta\alpha^{*}=\Delta 2I_{max}=$۰/۴	
تعادل محتوایی	آزمون سه محتوایی	حساب-دیفرانسیل هندسه جبر

ارزیابی خزانه سؤال شبیه‌سازی‌شده. در این پژوهش عملکرد خزانه‌های سؤال بهینه شبیه‌سازی‌شده با خزانه‌های سؤال عملیاتی بر اساس مجموعه‌ای از ملاک‌های تجربی ارزیابی و مقایسه شد. توزیع نمونه‌گیری دوم مطالعه (توزیع ۳۲۵۰۰ نفری) در ارزیابی

خزانه سؤال مورداستفاده قرار گرفت. ملاک‌های ارزیابی خزانه سؤال به شرح زیر است (چانگ و ییِنگ[1]، ۱۹۹۹؛ ریکیسی 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 ۲۰ (I=۲۰، ...، ۲، ۱=i) بودند، آگاهی آزمون، به‌عنوان 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 j ام (۶۵، ... ،۲ ،۱=j) آگاهی آزمون در زیر ارائه‌شده، نمادهای استفاده‌شده در این رابطه قبلاً تعریف شده‌اند:

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

۲- **خطای استاندارد شرطی اندازه‌گیری[4] (CSEM):** این شاخص میزان خطای اندازه‌گیری برآورد توانایی را در هر یک از سطوح توانایی واقعی (θj) محاسبه می‌کند:

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

اگر θ_j ،و توانایی ام (۶۵، ... ،۲ ،۱=j) در پیوستار ۴- تا ۴+ (یعنی؛ ۴+، ۳/۸۷۵+، ... ۳/۸۷۵-، ۴-=θ_j) را نشان دهد، i هر یک از آزمودنی‌ها در θ_j و N_i=۵۰۰، تعداد کل تکرارهای CAT اجراشده 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.

1. Chang&Ying
2. Conditional test information
3. Efficacy index
4. Conditional standard error of measurement

۳- **اریب و میانگین مجذور خطا**[1] **(MSE):** در معادله‌های ۴ و ۵، N تعداد کل آزمودنی‌ها در تمام نقاط ثابت توانایی (۶۵ نقطه ثابت توانایی از ۴- تا ۴) یعنی، برابر با $(\sum_{j=1}^{65}\sum_{i=1}^{500} N_{ij} = 32500)$ آزمودنی است و $\hat{\theta}_i$ برآورد کننده iامین آزمودنی با سطح توانایی واقعی θ_i است:

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i) \qquad (۴)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2 \qquad (۵)$$

۴- **اریب شرطی**[2] **و میانگین مجذور خطای شرطی**[3] **(CMSE):** در اصل در معادله‌های ۶ و ۷ مقدار θ_j همان مقادیر ثابت (یعنی؛ ۴+، ۳/۸۷۵+، ... ۳/۸۷۵-، ۴-=θ_j) است، که اینجا برآوردهای $\hat{\theta}_{ij}$ در هر سطح توانایی، از مقدار ثابت θ_j کم می‌شود:

$$Conditional\ Bias(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j) \qquad (۶)$$

$$CMSE(\theta_j) = \frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{\theta}_{ij} - \theta_j)^2 \qquad (۷)$$

۵- **کجی توزیع نرخ مواجهه سؤال**[4]**:** آماره کای دو که توسط چانگ و ییِنگ (۱۹۹۹) ارائه‌شده، برای اندازه‌گیری میزان کجی توزیع مواجهه سؤال به‌کاررفته و برابر است با:

$$X^2 = \sum_{i=1}^{n}\frac{(r_i - \frac{L}{n})^2}{\frac{L}{n}} \qquad (۸)$$

1. Bias and mean square error
2. Conditional Bias
3. Conditional mean square error
4. Skewness of item exposure rate distribution

در این معادله r_i، نسبت نرخ مشاهده‌شده iامین سؤال، L=۲۰، طول آزمون و n تعداد سؤال‌های خزانه است. معادله ۸ اختلاف بین نرخ مواجهه سؤال مشاهده‌شده و ایده‌آل را محاسبه می‌کند و مقدار اثربخشی استفاده از خزانه سؤال را نیز تعیین می‌نماید. مقدار کای دو کوچک نشان می‌دهد که بیشتر سؤال‌ها استفاده شده‌اند.

۶- **درصد سؤال‌های بیش مواجهه شده**[1]: نرخ مواجهه یک سؤال را می‌توان به‌عنوان نسب تعداد دفعات اجراهای سؤال به تعداد کل آزمودنی‌ها در نظر گرفت. درمجموع، سطح متوسط نرخ مواجهه سؤال مناسب است. نرخ بالای مواجهه یک سؤال بدین معناست که خطر فاش شدن سؤال برای آزمودنی‌های بعدی افزایش می‌یابد. اگر چنین باشد، هم امنیت و هم روایی[2] آزمون به دلیل نرخ بالای مواجهه سؤال مورد تهدید قرار می‌گیرد؛ بنابراین، درصد سؤال‌های بیش مواجهه شده، به‌عنوان ملاک مهمی برای ارزیابی موفقیت برنامه CAT در نظر گرفته می‌شود (هَو و چانگ[3]، ۲۰۰۱).

۷- **درصد سؤال‌های کم مواجهه شده**[4]: نرخ کم مواجهه شدن یک سؤال بدین معناست که یک سؤال بندرت در برنامه CAT مورداستفاده قرار گیرد. خزانه سؤالی که سؤال‌های بسیار زیادی با نرخ مواجهه خیلی پایینی دارد، دارای فایده کمی است. دو موضوع به‌صرفه بودن طراحی سؤال‌ها و مناسب بودن شیوه انتخاب آن‌ها، به دلیل نرخ مواجهه کم سؤال به چالش کشیده می‌شوند. نرخ مواجهه پایین‌تر از ۰/۰۲ به‌عنوان سؤال کم مواجهه شده در نظرگرفته می‌شود (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 به‌طور تصادفی نمونه‌گیری شدند، ارائه می‌شود، تقسیم‌بر طول مورد انتظار آزمون[6]. به‌طور ایده‌آل، تعداد سؤال‌های مشترک بین دو

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

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

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

در این رابط N، تعداد CATهای (با طول ثابت) اجراشده، L، تعداد سؤال‌های در هر یک از CATها، n تعداد سؤال‌های خزانه و m_i تعداد دفعاتی است که سؤال i برای همه N تعداد CAT اجرا شده است.

سؤال‌های زیر در این پژوهش بررسی می‌شوند:

۱) آیا روش شبیه‌سازی مونت‌کارلو ریکیسی (۲۰۰۳) می‌تواند با روش برنامه‌نویسی ریاضی (WDM) تلفیق شود؟

۲) عملکرد خزانه سؤال بهینه برای CAT زمانی که در الگوریتم انتخاب سؤال، علاوه بر تعادل محتوایی، مواجهه بیش‌ازحد سؤال وارد می‌شود در مقابل زمانی که مواجهه بیش‌ازحد وارد نمی‌شود، چگونه است؟

۳) آیا خزانه‌های سؤال بهینه‌ای که طراحی می‌شوند بهتر از خزانه سؤال عملیاتی برحسب ملاک تجربی عمل می‌کنند؟

یافته‌های پژوهش

در این قسمت نتایج مربوط به شبیه‌سازی خزانه‌های بهینه و مقایسه عملکرد آن‌ها با خزانه سؤال عملیاتی شرح داده می‌شود. همان‌طور که در جدول ۴ نمایش داده شده است، در این پژوهش برای ساخت خزانه‌های سؤال بهینه‌ای که در آن تعادل محتوایی نیز از اهمیت برخوردار است، دو عامل (سه روش ایجاد سؤال بهینه و اعمال و عدم اعمال شیوه کنترل

1. Chen & Ankenmann & Spray

مواجهه سیمپسون-هتر) دست‌کاری شده‌اند، بنابراین، تعداد شش الگوی خزانه سؤال بهینه به وجود آمده است. به خزانه‌های سؤال بهینه‌ای که بر اساس روش bin-and-union[1] ایجاد شدند، به‌اختصار ROP[2] گفته می‌شود (هی و ریکیسی، ۲۰۱۱). همچنین، خزانه سؤال عملیاتی به‌اختصار با OP[3] نشان داده می‌شود (گو و ریکیسی، ۲۰۰۷). پهنای b–bin ها در تمام خزانه‌ها برابر با ۰/۲ و میزان تغییر آگاهی در پارامتر a برابر با ۰/۴ قرار داده شد.

جدول ٤. ترکیب عامل روش ایجاد سؤال بهینه و شیوه کنترل مواجهه برای ساخت خزانه‌های سؤال بهینه با تعادل محتوایی

شیوه کنترل مواجهه سیمپسون-هتر	روش ایجاد سؤال بهینه			
Exposure control-) Sympson-Hetter procedure	item generation methods	R	MRP	MTI
	without- Exposure control	ROP_1	ROP_2	ROP_3
	Exposure control	ROP_4	ROP_5	ROP_6

خزانه‌های سؤال بهینه بدون کنترل مواجهه بیش‌ازحد سؤال. در این مرحله خزانه‌های بهینه با توجه به ایجاد تعادل محتوایی بر اساس روش برنامه‌نویسی WDM و بدون عامل کنترل مواجهه سیمپسون-هتر (S-H) شبیه‌سازی شده‌اند. روش ایجاد سؤال بهینه از طریق سه روش R، MRP، MTI با پهنای b-bin = 0.2 و میزان آگاهی یا پهنای پارامتر a برابر با a-bin: $\Delta a2=2\Delta IMaximum = 0.4$ بوده است. جدول ۵، اندازه‌ها و خلاصه آماره‌های مربوط به پارامترهای سؤال در خزانه‌های سؤال بهینه و عملیاتی را در سه محتوای مشخص ارائه می‌کند. نتایج نشان می‌دهد که خزانه‌های سؤال بهینه شامل حداقل تعداد سؤال می‌باشند. البته یکی از دلایل آن می‌تواند این قضیه باشد که در ساخت آن‌ها هیچ نوع کنترل مواجهه‌ای صورت نگرفته است. توزیع پارامترهای سؤال در ماتریس‌های مربوط به هر یک از خزانه‌ها نشان می‌دهد[4] که همه خزانه‌های بهینه دارای سؤال‌هایی با دامنه

۱. برای یادآوری به صفحه ۶ در مقدمه مراجعه کنید.

2. Range-optimal item pool

3. Operational item pool

۴. به دلیل محدودیت صفحات از آوردن ماتریس پارامترها خودداری می‌شود و تنها به ذکر نتایج استخراج‌شده از این ماتریس‌ها اکتفا می‌شود.

محدودی از سطوح دشواری هستند. دلیل این امر این است که زمانی که قیود محتوایی در تعامل با ویژگی‌های روان‌سنجی قرار می‌گیرند، خزانه‌های سؤال بهینه شبیه‌سازی دارای ویژگی‌های روان‌سنجی دقیق‌تری می‌شوند، به‌طوری‌که دامنه دشواری سؤال‌ها محدودتر می‌شود. خزانه‌های سؤال بهینه دارای میانگین دشواری بالاتری نسبت به خزانه‌های عملیاتی هستند. خزانه بهینه MTI در هر سه محتوا دارای حداقل تعداد سؤال است، ولی تفاوت زیادی با خزانه‌های MRP ندارد. بااین‌وجود، 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 سؤال، 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، این نتیجه به دلیل ماهیت روشی است که پارامترهای سؤال را ایجاد می‌کند. در این روش، پارامترها در سراسر ماتریس پراکنده می‌شوند. توزیع پارامتر دشواری و تشخیص سؤال‌ها در این روش بسیار مشابه خزانه عملیاتی است و در تمام محتواها دارای مقادیر پارامتر متنوع‌تری است. امّا سؤال‌های دشوار در خزانه‌های بهینه MRP دارای پارامتر ضریب تشخیص بالاتری هستند و برعکس سؤال‌ها 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 کمتری هستند. بررسی نتایج عملکرد این خزانه‌ها در جدول ۶ آورده شده است. برآورد توانایی در هر سه خزانه بهینه و عملیاتی، دارای سطح معینی از اریب مثبت است، ولی مقدار این اریب‌ها بسیار کوچک است. میانگین مجذور خطا (MSE) در خزانه‌های سؤال بهینه کوچک‌تر از خزانه سؤال عملیاتی است. در این میان خزانه‌های سؤال بهینه MTI عملکرد بهتری در شاخص (MSE) نشان می‌دهند، زیرا تعامل میزان حداقل آگاهی در هر سطح با ویژگی‌های محتوایی موردتوجه قرار گرفته است و بنابراین، توانایی را با دقت بیشتری برآورد می‌کنند. خزانه‌های سؤال بهینه دارای نرخ همپوشی پایین‌تری هستند، با وجود این‌که دارای سؤال‌ها کمتری می‌باشند. نمودار ۱ نشان می‌دهد که نرخ

همپوشی تست در سطوح توانایی زیر ۲- در خزانه‌های عملیاتی و بهینه تقریباً مشابه است. ولی در بقیه سطوح نرخ همپوشی خزانه‌های بهینه کمتر از عملیاتی است. خزانه‌های بهینه در مقدار همپوشی تست‌ها شباهت زیادی به یکدیگر دارند. دلیل این شباهت به خاطر ویژگی مشترک همه آن‌ها در ایجاد پارامتر b است، زیرا پارامتر b ارتباط مستقیمی با انتخاب و سرهم شدن تست‌ها دارد. در این پژوهش خزانه سؤال عملیاتی همانند آزمون‌های مشابه دیگر در نوع خود، در دو انتهای سطوح توانایی دارای نرخ مواجهه کم نیست، زیرا این برنامه طوری طرح‌ریزی‌شده که نقاط برش در سطوح بالای توانایی باشند پس طبیعتاً به سؤال‌های دشوارتری نیاز دارد. ازاین‌رو، تنها در سطوح پایین توانایی نرخ همپوشی تست کمتر است. همچنین، خزانه‌های بهینه درصد خیلی کوچک‌تری از کم مواجهه شدن سؤال‌ها را نسبت به خزانه عملیاتی دارند. نرخ سؤال‌های بیش مواجهه شده در هر سه خزانه تقریباً مشابه و بیشتر از خزانه عملیاتی است. نمودارهای ۲ تا ۵ درصدهای مواجهه سؤال در هر یک از سطوح توانایی را برای خزانه‌های عملیاتی و بهینه نشان می‌دهد. در همه خزانه‌های سؤال، سؤال‌ها خیلی آسان و خیلی دشوار که به ترتیب در سطوح پایین و بالای توانایی ارائه می‌شوند، دارای نرخ‌های مواجهه کوچک‌تری هستند. نتایج نشان می‌دهد که در همه خزانه‌ها، سؤال‌های با سطوح دشواری متوسط بیشترین قابلیت استفاده را دارند. در این نمودارها نرخ‌های مواجهه سؤال، به تفکیک محتواها در هر چهار خزانه گزارش شده است. شکل نمودارها الگوی جالبی از مواجهه را در هریک از خزانه‌ها نشان می‌دهد. در خزانه‌های عملیاتی، سؤال‌ها کم مواجهه شده در سرتاسر طول پیوستار توانایی پراکنده شدند. خزانه‌های بهینه R دارای الگویی با توزیع نرمال در میزان مواجهه سؤال‌ها هستند. خزانه‌های MRP در سطوح بالای توانایی دارای نرخ مواجهه کمتری هستند. در پایان، خزانه‌های بهینه MTI تنها در یک بازه‌ی کوچک در متوسط پارامتر توانایی دارای مواجهه بیش‌ازحد می‌باشند و در سطوح دیگر میزان مواجهه کم می‌شود. جدول ۷ نرخ‌های مواجهه سؤال‌ها را به تفکیک محتواها در هر چهار خزانه سؤال نشان می‌دهد. نتایج جدول ۷ گویای این مطلب است که در محتوای اول، دوم و سوم خزانه‌های MRP دارای بیشترین نرخ مواجهه هستند، زیرا ویژگی‌های پارامترهای سؤال‌ها در خزانه‌های عملیاتی (مانند میزان

ضریب همبستگی بین پارامترهای a و b در تعامل با پارامتر توانایی) با ویژگی‌های این روش ایجاد سؤال تعامل برقرار کرده و سؤال‌هایی با ضریب تشخیص بالا در سطوح بالای توانایی ایجاد می‌کنند که با توجه به ماهیت آزمون (میانگین دشواری آزمون بالاتر از متوسط است)، این نتیجه به‌دست می‌آید. همچنان که در نمودار ۶ ملاحظه می‌شود، میانگین آگاهی خزانه‌های سؤال در خزانه‌های R، MTI و خزانه سؤال عملیاتی در تمام سطوح کاملاً مشابه است. خزانه MRP در سطوح توانایی زیر ۲ کاملاً مشابه خزانه‌های دیگر است ولی در سطوح توانایی بالای ۲ متوسط میزان آگاهی بسیار بیشتر می‌شود که به دلیل ماهیت روش ایجاد پارامترهای سؤال است. نتایج مربوط به آگاهی تست‌ها نمایانگر این است که اگر در ساخت خزانه‌های بهینه سؤال به ویژگی‌های کاربردی آزمون‌های عملیاتی توجه شود، نتایج خزانه‌های بهینه و عملیاتی بسیار به یکدیگر شبیه می‌شود، با این تفاوت که اندازه خزانه‌های سؤال بهینه بسیار کمتر از خزانه سؤال عملیاتی است.

نمودار ۷، ۸ و ۹ به ترتیب، خطای استاندارد شرطی اندازه‌گیری (CSEM)، اریب شرطی و میانگین مجذور خطا (CMSE) را در هر چهار خزانه سؤال نشان می‌دهد. خطاهای استاندارد اندازه‌گیری الگویی مشابه با خزانه‌های عملیاتی دارند. ولی در تمام سطوح توانایی کمتر از خزانه عملیاتی است. نمودار ۸ نشان می‌دهد که در خزانه عملیاتی میزان اریب در اغلب سطوح توانایی بیشتر از خزانه‌های بهینه است. خزانه MTI نسبت به خزانه‌های دیگر از میزان اریب کمتری برخوردار است. دلیل این نتیجه تعامل بین ویژگی‌های محتوایی و حداقل میزان آگاهی آزمون است. همچنین، نمودار ۹ میانگین مجذور خطا را در سطوح متفاوت توانایی نشان می‌دهد. نتایج این نمودار نشان می‌دهد که MSE هر سه خزانه بهینه کوچک‌تر از خزانه سؤال عملیاتی است. به‌خصوص خزانه MTI از حداقل مقدار MSE برخوردار است.

جدول ۵. 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(arithmetic)													
OP	۴۵۵	۱/۰۸۹	۰/۲۸۴۴	۳/۰۴۵	۰/۱۶۶	-۰/۰۳۹	۱/۰۷۹	۳/۹۸۱	-۳/۵۹۶	۰/۱۴۵	۰/۸۰۱	۰/۴۱۷۹	۰/۰۰۰۵
ROP_1	۲۳۰	۱/۷۹	۰/۲۷۷	۲/۷۸	۰/۴۶	۰/۰۱۶	۱/۱۵	۳/۶۵	-۳/۸۲	۰/۱۴۶	۰/۷	۰/۴۱۰۲	۰/۰۰۱
ROP_2	۱۷۸	۲/۰۱۲	۰/۲۸۱	۳/۱۲	۰/۹۵	۰/۰۲۰۱	۱/۰۶۳	۳/۸۲۵	-۳/۱۱۸	۰/۱۴۸	۰/۷۲	۰/۴۳	۰/۰۲۴
ROP_3	۱۷۴	۱/۵۶۲	۰/۲۵۳	۲/۳۲۴	۰/۹۸۷	۰/۰۱۴۶	۱/۰۰۸	۳/۸۴۱	-۳/۵۹۲	۰/۱۴۲	۰/۰۶۴	۰/۴۶	۰/۰۰۱
Content 2 (geometry)													
OP	۲۵۸	۱/۲۰۶	۰/۲۲۴۵	۲/۹۳	۰/۲۴۵	-۰/۰۴۸۲	۱/۰۹۴	۳/۸۴	-۳/۴۵۸	۰/۱۸۴	۰/۰۹۷	۰/۴۷۶	۰/۰۹۱
ROP_1	۱۹۹	۱/۷۳۲	۰/۲۶۳	۲/۶۵۴	۰/۵۲۶	-۰/۰۸۷۵	۰/۹۷۶	۳/۵۵	-۳/۸۵۴	۰/۱۷۱	۰/۰۷۴	۰/۴۲	۰/۰۱
ROP_2	۱۶۸	۱/۸۸۲	۰/۲۷۸	۲/۹۳۵	۰/۹۲۱	۰/۰۱۲۸	۱/۰۴۷	۳/۷۴	-۳/۴۹	۰/۱۸۳	۰/۰۶۸	۰/۴۶	۰/۰۰۱۵
ROP_3	۱۶۷	۱/۷۵۹	۰/۲۷۲	۲/۵۱	۱/۱۵	-۰/۰۱۳۵	۰/۹۹۳	۳/۱۰۴	-۳/۱۰۴	۰/۱۷۸	۰/۰۸۱	۰/۴۸	۰/۰۰۲
Content 3 (algebra)													
OP	۲۰۸	۱/۳۵۶	۰/۲۴۷	۲/۸۸۹	۰/۵۳۸	۰/۰۳۲۴	۱/۰۱۸	۳/۶۸۹	-۳/۴۰۹	۰/۱۷۴	۰/۷۵۴	۰/۴۸۹	۰/۰۰۴
ROP_1	۱۸۳	۱/۷۸۳	۰/۲۶۸	۲/۷۸۹	۰/۲۳۵	۰/۰۲۳۸	۰/۹۹۸	۳/۵۳۲	-۳/۴۵۶	۰/۱۸۱	۰/۰۶۲	۰/۴۹۷	۰/۰۰۳
ROP_2	۱۵۷	۲/۱۴۸	۰/۲۹۱	۳/۰۴۶	۰/۹۵	۰/۰۱۸۴	۰/۹۹۴	۳/۵۸۹	-۳/۱۴	۰/۱۸۹	۰/۰۷۵	۰/۴۶۵	۰/۰۰۲
ROP_3	۱۵۵	۱/۷۹۴	۰/۲۷۹	۲/۴۸۹	۰/۹۰۲	-۰/۰۱۰۹	۱/۰۰۷	۳/۵۹۸	-۳/۰۸۹	۰/۱۸۴	۰/۷۴۸	۰/۴۵	۰/۰۰۱

جدول ٦: خلاصه‌ی آماره‌های عملکرد خزانه سؤال، بدون S-H (b-bin=0.2)، با تعادل محتوا

آماره‌ها	OP	R	MRP	MTI
Bias	۰/۰۰۵۲	۰/۰۰۲۱	۰/۰۰۰۷۳	۰/۰۰۱۹۹
MSE	۰/۰۱۷۴۵	۰/۰۱۲۶۷	۰/۰۱۱۳۸	۰/۰۰۷۶
کجی نرخ مواجهه	۹۸/۸۵۶	۵۱/۴۰۲	۴۸/۴۳۶۵	۴۳/۴۹۴
نرخ همپوشی سؤال	۰/۵۱۷۳	۰/۳۷۶۵	۰/۳۸۴۴	۰/۳۹۴۱
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۸/۱۶۹٪	۹/۶۳۴٪	۱۱/۵۲۴٪	۱۰/۰۰۳٪
درصد سؤال‌های با نرخ مواجهه کوچک‌تر از 0.02	۵۱/۵۲۶٪	۲۵/۵۳۱٪	۲۱/۱۵٪	۱۹/۱۸۹٪
درصد تست‌هایی که از قیود تست تخطی دارند	۵۴/۸٪	۳/۰۰۲٪	۲/۰۰۱۳٪	۲/۰۰۱۵٪
اندازه خزانه سؤال	۹۲۱	۶۱۲	۵۰۳	۴۹۶

جدول ۸ و ۹ نتایج مربوط به میزان تخطی از قیود را در تست‌ها برحسب پارامترهای توانایی نشان می‌دهد. در CAT عملیاتی، ۳۵۰ تست سرهم شدند و در CAT های شبیه‌سازی‌شده، ۶۰۰۰ تست سرهم شدند، برای قابل‌مقایسه کردن میزان تخطی‌ها، درصد فراوانی نسبی آن‌ها محاسبه شد. مطابق با نتایج این دو جدول، خزانه عملیاتی در تمام سطوح توانایی میزان تخطی از قیود بیشتری نسبت به خزانه‌های بهینه دارد. در خزانه‌های R میزان تخطی‌ها در دو دامنه توانایی بیشتر است و تقریباً الگویی مشابه با خزانه عملیاتی دارد. در خزانه MRP و MTI در سطوح پایین توانایی میزان تخطی از قیود بیشتر است. در خزانه MRP در سطوح بالای توانایی به دلیل وجود سؤال‌ها بیشتر، تخطی از قیود در تست‌هایی که سرهم می‌شود، به حداقل خود می‌رسد.

جدول ۷. درصد سؤال‌ها بیش مواجهه شده و کم مواجهه شده

در هر یک از محتواها برای خزانه‌های سؤال بدون S-H

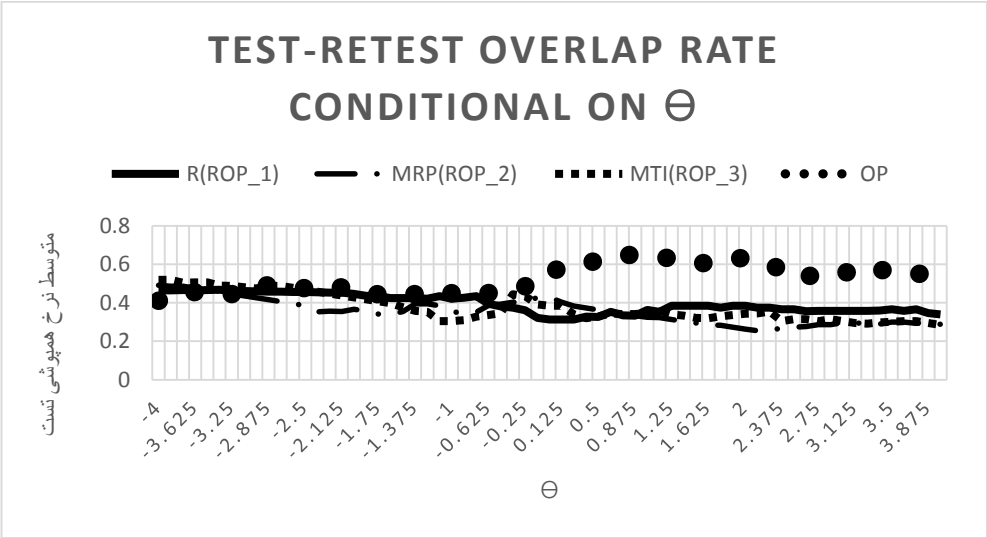
آماره	محتوای اول				محتوای دوم				محتوای سوم			
خزانه‌های سؤال	OP	R	MRP	MTI	OP	R	MRP	MTI	OP	R	MRP	MTI
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۹/۸۱٪	۹/۹۲۱٪	۱۳/۴۶٪	۱۱/۷۷٪	۷/۵۸٪	۹/۵۸٪	۹/۹۷٪	۸/۵۲۱٪	۷/۰۸٪	۷/۹۳٪	۹/۵۹٪	۸/۳۱۶٪
اندازه خزانه	۴۵۵	۳۳۰	۱۷۸	۱۷۴	۲۵۸	۱۹۹	۱۶۸	۱۶۷	۲۰۸	۱۸۳	۱۵۷	۱۵۵

جدول ۸. تعداد تست‌ها بر اساس تخطی از قیود محتوایی در CAT ها، بدون S-H

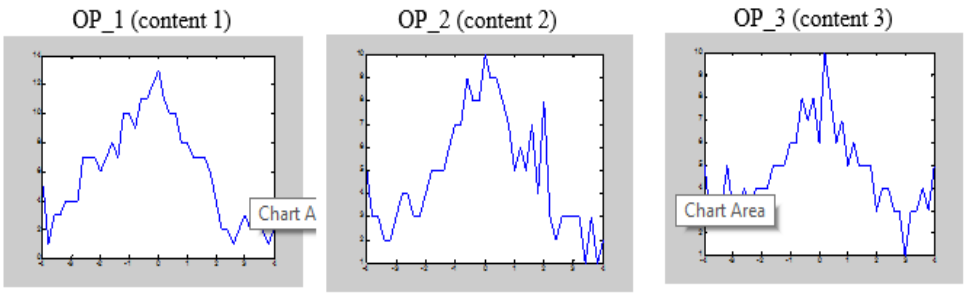
سطوح توانایی / خزانه سؤال	CAT total	OP	R (ROP_1)	MRP (ROP_2)	MTI (ROP_3)	کل
-۴	۲۵	۱۱	۲۵	۲۳	۲۴	۱۲۴
-۳/۵	۱۸	۱۴	۱۸	۲۰	۲۲	۱۲۴
-۳	۱۵	۱۸	۱۵	۱۴	۱۷	۱۲۴
-۲/۵	۷	۱۴	۷	۱۹	۱۳	۱۲۴
-۲	۵	۱۴	۵	۱۳	۱۰	۱۲۴
-۱/۵	۰	۱۴	۰	۱۰	۰	۱۲۴
-۱	۰	۳	۰	۰	۸	۱۲۴
-۰/۵	۳	۱	۳	۱	۳	۱۲۴
۰	۰	۱	۰	۰	۰	۱۲۴
۰/۵	۰	۲	۰	۰	۳	۱۲۴
۱	۱۱	۴	۱۱	۶	۲	۱۲۴
۱/۵	۵	۷	۵	۴	۳	۱۲۴
۲	۷	۱	۷	۳	۱	۱۲۴
۲/۵	۷	۱۱	۷	۲	۰	۱۲۴
۳	۱۲	۲۵	۱۲	۱	۴	۱۲۴
۳/۵	۱۶	۲۵	۱۶	۱	۶	۱۲۴
۴	۲۷	۲۸	۲۷	۲	۸	۱۲۴
کل	۱۵۸	۱۹۳	۱۵۸	۱۱۹	۱۲۴	

جدول ۹. درصد تست‌ها بر اساس تخطی از قیود محتوایی در CAT ها، بدون S-H

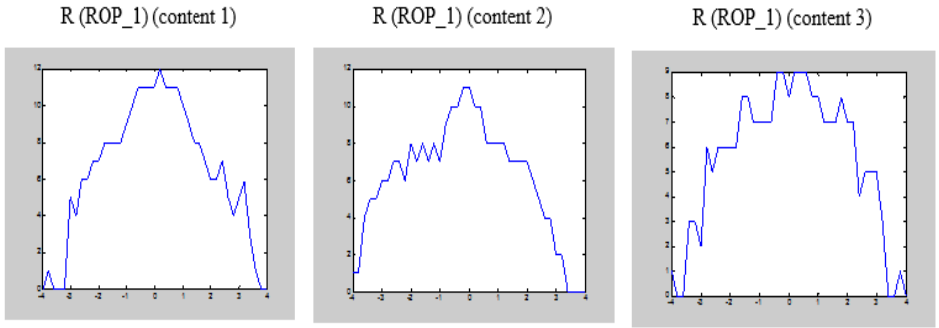
سطوح توانایی / خزانه سؤال	CAT total	OP	R (ROP_1)	MRP (ROP_2)	MTI (ROP_3)	کل
-۴	۶۰۰۰	۰/۰۳	۰/۰۰۴	۰/۰۰۴	۰/۰۰۴	۰/۰۳
-۳/۵	۶۰۰۰	۰/۰۴	۰/۰۰۳	۰/۰۰۳	۰/۰۰۴	۰/۰۳
-۳	۶۰۰۰	۰/۰۵	۰/۰۰۲۵	۰/۰۰۲	۰/۰۰۳	۰/۰۳
-۲/۵	۶۰۰۰	۰/۰۴	۰/۰۰۱	۰/۰۰۳	۰/۰۰۲	۰/۰۳
-۲	۶۰۰۰	۰/۰۴	۰/۰۰۰۸	۰/۰۰۳	۰/۰۰۱	۰/۰۳
-۱/۵	۶۰۰۰	۰/۰۴	۰	۰/۰۰۱	۰	۰/۰۳
-۱	۶۰۰۰	۰/۰۰۸	۰	۰	۰	۰/۰۳
-۰/۵	۶۰۰۰	۰/۰۰۱	۰	۰	۰	۰/۰۳
۰	۶۰۰۰	۰/۰۰۱	۰	۰	۰	۰/۰۳
۰/۵	۶۰۰۰	۰/۰۰۴	۰	۰	۰	۰/۰۳
۱	۶۰۰۰	۰/۰۱	/۰۰۲	۰/۰۰۱	۰	۰/۰۳
۱/۵	۶۰۰۰	۰/۰۲	۰/۰۰۰۸	۰	۰	۰/۰۳
۲	۶۰۰۰	۰/۰۰۴	۰/۰۰۱	۰	۰	۰/۰۳
۲/۵	۶۰۰۰	۰/۰۳	۰/۰۰۱	۰	۰	۰/۰۳
۳	۶۰۰۰	۰/۰۷	۰/۰۰۲	۰	۰	۰/۰۳
۳/۵	۶۰۰۰	۰/۰۷	۰/۰۰۳	۰	۰/۰۰۱	۰/۰۳
۴	۶۰۰۰	۰/۰۸	۰/۰۰۵	۰	۰/۰۰۱	۰/۰۳
کل	۶۰۰۰	۰/۵۴۸	۰/۰۳	۰/۰۲	۰/۰۲	



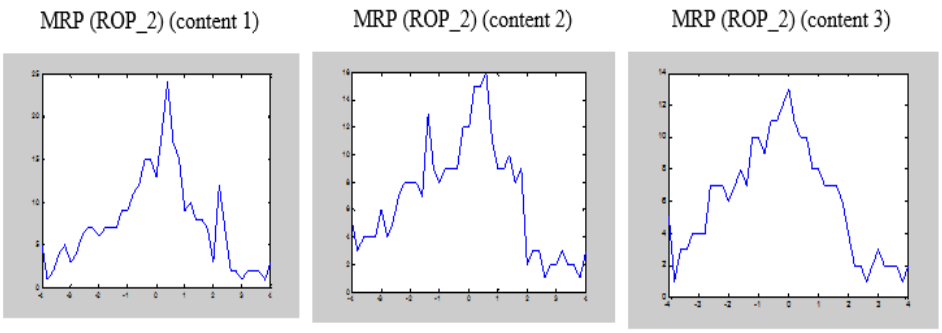
نمودار ۱. نرخ همپوشی تست مشروط به Θ بدون **S-H** (***b*-bin: 0.2**) با تعادل محتوایی



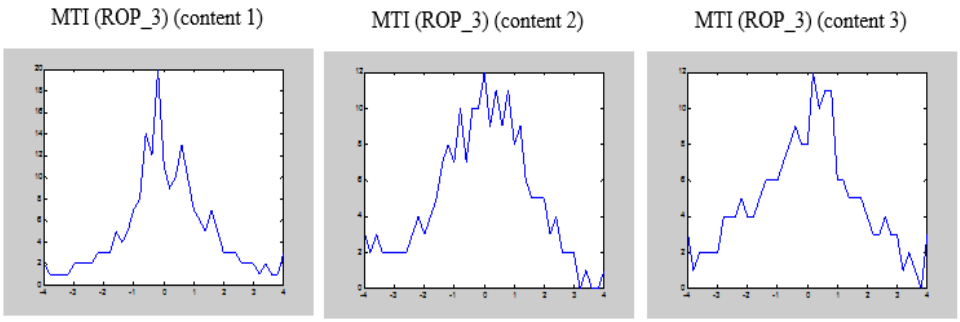
نمودار ۲. درصد سؤال‌ها بیش مواجهه شده در خزانه عملیاتی برحسب محتواهای سه‌گانه



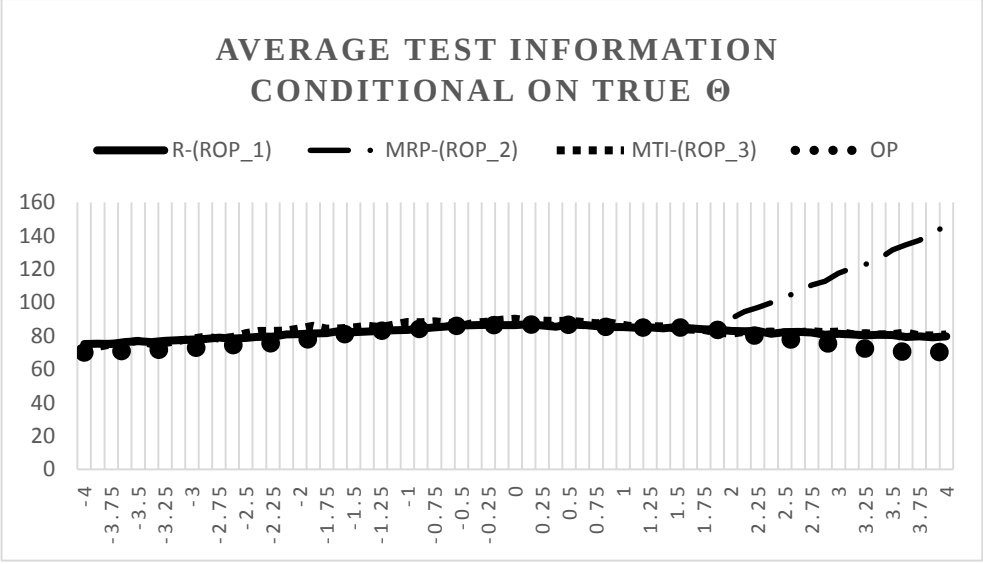
نمودار ۳. نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده در سطوح توانایی بر اساس روش R



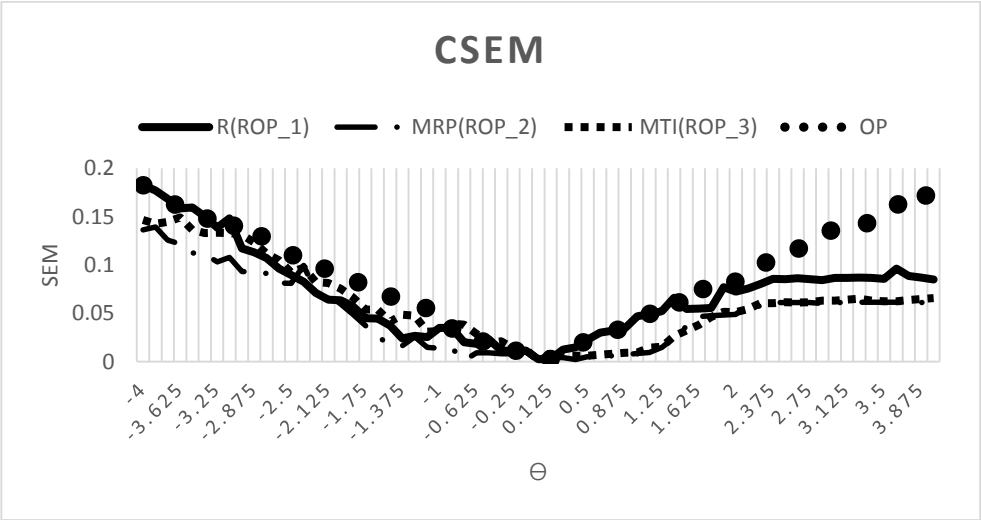
نمودار ٤. نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده در سطوح توانایی بر اساس روش MRP



نمودار ٥. نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده بر اساس روش **MTI**

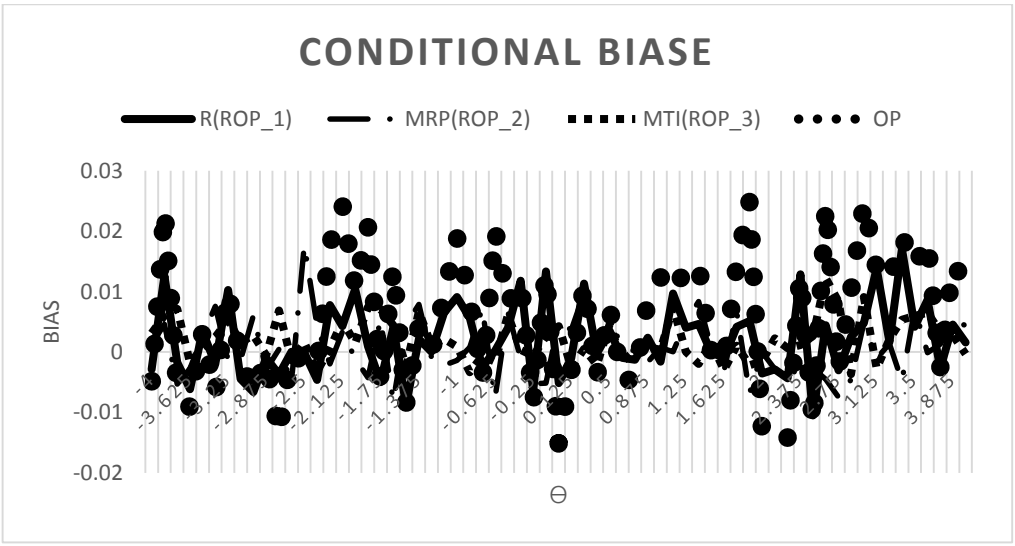


نمودار ٦. متوسط آگاهی تست مشروط به توانایی واقعی در خزانه‌های سؤال بدون S-H (*b*-bin: 0.2)، با تعادل محتوا



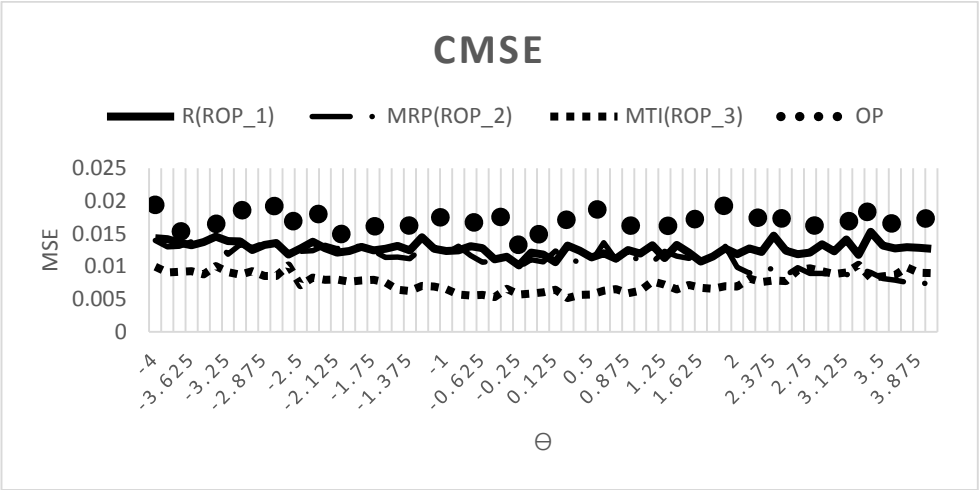
نمودار ۷. خطای استاندارد اندازه‌گیری (CSEM) در خزانه‌های سؤال

بدون S-H (*b*-bin: 0.2)، با تعادل محتوایی



نمودار ۸. اریب شرطی (conditional Bias) در خزانه‌های سؤال

بدون S-H (*b*-bin: 0.2)، با تعادل محتوایی



نمودار ۹. میانگین مجذور خطا (CMSE) در خزانه‌های سؤال
بدون S-H (*b*-bin: 0.2)، با تعادل محتوایی

خزانه‌های سؤال بهینه با کنترل مواجهه بیش‌ازحد سؤال. در این مرحله، خزانه‌های سؤال با روش کنترل مواجهه S-H، شبیه‌سازی شدند. جدول ۱۰ اندازه‌ها و خلاصه آماره‌های مربوط به پارامترهای سؤال در خزانه‌های سؤال بهینه و عملیاتی را در هر سه محتوای ارائه می‌کند. نتایج نشان می‌دهد که با وجود کنترل مواجهه سؤال در ساخت خزانه‌های سؤال بهینه، هنوز این خزانه‌ها شامل حداقل تعداد سؤال می‌باشند. در این مرحله هم همانند خزانه‌هایی که بدون کنترل مواجهه ایجاد شدند، هر سه خزانه بهینه دارای سؤال‌های با دامنه محدودی از 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. خزانه‌های MRP (ROP_5) نسبت به MRP (ROP_2) دارای تعداد سؤال‌های بسیار بیشتری هستند. درصورتی‌که مقایسه خزانه‌های R (ROP_4) نسبت به R (ROP_1) و MTI (ROP_6) نسبت به MTI (ROP_3) به این اندازه تفاوت نشان نمی‌دهد. در این مرحله نیز خزانه‌های MTI دارای حداقل میانگین پارامتر a می‌باشند و خزانه‌های بهینه MRP دارای بیشترین مقدار پارامتر a هستند. نوع توزیع پارامترهای سؤال در هر سه خزانه سؤال بهینه مشابه خزانه‌های سؤال در مرحله قبل است، با این تفاوت که میانگین ضریب

تشخیص در هر سه خزانه بالاتر است. بررسی نتایج عملکرد این خزانه‌ها در جدول ۱۱ ارائه شده است. برآورد توانایی در هر سه خزانه بهینه و عملیاتی، دارای سطح معینی از اریب مثبت است، ولی مقدار این اریب‌ها بسیار ناچیز است. میانگین مجذور خطا (MSE) در خزانه‌های سؤال بهینه کوچک‌تر از خزانه سؤال عملیاتی است. در میان خزانه‌های سؤال بهینه، MRP عملکرد بهتری در این شاخص نشان می‌دهد، زیرا دارای سؤال‌هایی با ضریب تشخیص بالاتری هستند؛ بنابراین، برآورد توانایی را با دقت بیشتری برآورد می‌کند. همچنین، خزانه‌های سؤال بهینه دارای نرخ همپوشی تست پایین‌تری هستند، با وجود این‌که دارای سؤال‌ها کمتری می‌باشند. نمودار ۱۰ نشان می‌دهد که نرخ همپوشی تست در تمام سطوح توانایی کمتر از خزانه‌های عملیاتی است. این نرخ در سرتاسر پارامتر توانایی به‌صورت یکنواخت توزیع شده است. در این مرحله نیز سه خزانه بهینه شباهت زیادی در همپوشی تست‌ها دارند. همچنین، خزانه‌های بهینه درصد خیلی کوچک‌تری از بیش مواجهه شدن سؤال‌ها را نسبت به خزانه عملیاتی دارند. نرخ سؤال‌ها بیش مواجهه شده در خزانه‌های MTI کمتر از خزانه‌های دیگر است. نمودارهای ۱۱ تا ۱۳ درصدهای مواجهه سؤال در هر یک از سطوح توانایی را نشان می‌دهد. الگوی این نمودارها نشان می‌دهد که توزیع مواجهه سؤال‌ها در این مرحله یکنواخت‌تر از مرحله‌ی قبل است. جدول ۱۲ نرخ‌های مواجهه سؤال‌ها را به تفکیک محتواها در هر چهار خزانه سؤال نشان می‌دهد. نتایج این جدول گویای این مطلب است که در محتوای اول، دوم و سوم خزانه‌های R دارای بیشترین نرخ مواجهه هستند و با این‌که دارای بیشترین تعداد سؤال است، 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 گویای این قضیه است که با وجود این‌که خزانه‌های MTI دارای تعداد سؤال کمتری هستند، نرخ‌های مواجهه کمتری دارند. نرخ سؤال‌ها کم مواجهه شده در این نوع خزانه‌ها نیز کمتر از بقیه است، همه این نتایج گویای بهینه بودن این نوع خزانه‌ها ازنظر استفاده بهینه از سؤال‌ها است. همچنان که در نمودار ۱۴ ملاحظه می‌شود، میانگین آگاهی خزانه‌های سؤال در خزانه‌های بهینه و خزانه سؤال عملیاتی در تمام سطوح کاملاً مشابه

است. مطابق نتایج جداول ۱۳ و ۱۴ خزانه عملیاتی در تمام سطوح درصد فراوانی نسبی تخطی از قیود بیشتری نسبت به خزانه‌های بهینه دارد. در خزانه‌های R میزان تخطی‌ها در دو دامنه‌ی توانایی بیشتر است و تقریباً الگویی مشابه با خزانه عملیاتی دارد. در خزانه MRP و MTI در سطوح پایین توانایی میزان تخطی از قیود بیشتر است. در خزانه MRP در سطوح بالای توانایی به دلیل وجود سؤال‌ها بیشتر، تخطی از قیود در تست‌هایی که سرهم می‌شود، به حداقل خود می‌رسد. در هر سه خزانه سؤال بهینه، میزان تخطی از قیود محتوایی تست‌ها بیشتر از زمانی است که کنترل مواجهه سؤال وارد نشده است، این نتیجه دلیلی بر این امر است که واردکردن 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. نمودارهای ۱۵، ۱۶ و ۱۷ به ترتیب خطای استاندارد شرطی اندازه‌گیری (CSEM)، اریب شرطی و میانگین مجذور خطا (CMSE) را در هر چهار خزانه سؤال نشان می‌دهند. خطاهای استاندارد اندازه‌گیری در سطوح توانایی زیر صفر الگویی مشابه با خزانه‌های عملیاتی دارند، بااین‌وجود مقادیر آن کمتر از خزانه عملیاتی است. ولی در سطوح توانایی بالاتر از صفر، به‌خصوص در خزانه‌های MRP و MTI این خطا به صفر نزدیک می‌شود و دارای الگوی متفاوتی می‌شود. نمودار ۱۶ نشان می‌دهد که در خزانه عملیاتی میزان اریب در اغلب سطوح توانایی بیشتر از خزانه‌های بهینه است. خزانه MRP نسبت به خزانه‌های دیگر از میزان اریب کمتری برخوردار است. همچنین، نمودار ۱۷ میانگین مجذور خطا را در سطوح متفاوت توانایی نشان می‌دهد. نتایج این نمودار نشان می‌دهد که MSE هر سه خزانه بهینه کوچک‌تر از خزانه سؤال عملیاتی است. به‌خصوص خزانه MRP از حداقل مقدار MSE برخوردار است.

جدول ۱۰. 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(arithmetic)													
OP	۴۵۵	۱/۰۸۹	۰/۲۸۴۴	۳/۰۴۵	۰/۱۶۶	−۰/۰۳۹	۱/۰۷۹	۳/۹۸۱	−۳/۵۹۶	۰/۱۴۵	۰/۸۰۱	۰/۴۱۷۹	۰/۰۰۰۵
ROP_4	۲۶۱	۱/۹۰۱	۰/۳۰۱	۳/۰۰۴	۰/۹۰۲	۰/۰۳۴۶	۰/۹۹۸	۳/۹۹۸	−۲/۹۸۹	۰/۱۴۸	۰/۰۷۲۳	۰/۴۳	۰/۰۰۱
ROP_5	۲۴۹	۲/۲۸۹	۰/۳۰۱۲	۳/۱۸۶	۰/۹۳۸	۰/۰۵۴۹	۱/۰۹۲	۳/۹۹۴	−۳/۴۹۵	۰/۱۴۲	۰/۰۶۸	۰/۴۲۱	۰/۰۰۴
ROP_6	۲۱۴	۱/۶۱۳	۰/۲۶۷	۲/۵۲۸	۱/۱۸۹	۰/۰۲۱۸۰	۱/۱۱۸	۳/۹۹۶	−۳/۶۴۷	۰/۱۴۴	۰/۰۷۱	۰/۴۶۹	۰/۰۰۱
Content 2 (geometry)													
OP	۲۵۸	۱/۲۰۶	۰/۲۲۴۵	۲/۹۳	۰/۲۴۵	−۰/۰۴۸۲	۱/۰۹۴	۳/۸۴۵	−۳/۴۵۸	۰/۱۸۴	۰/۰۹۷	۰/۴۷۶	۰/۰۹۱
ROP_4	۲۳۹	۱/۸۹۳	۰/۲۹۴	۲/۸۵۴	۰/۹۰۵	−۰/۰۹۴۸	۱/۱۸۲	۳/۵۹۳	−۳/۵۸۷	۰/۱۷۹	۰/۰۹۱	۰/۴۶	۰/۰۰۴
ROP_5	۲۳۶	۲/۰۹۳	۰/۳۰۴	۳/۱۸۳	۱/۰۸۶	۰/۲۱۵	۱/۰۰۳	۳/۹۹۷	−۳/۴۸۲	۰/۱۸۲	۰/۰۸۷	۰/۴۵۸	۰/۰۰۳
ROP_6	۲۰۶	۱/۵۴۲	۰/۲۸۶	۲/۵۰۲	۱/۰۸۷	−۰/۰۳۶۸	۱/۱۲۲	۳/۷۸۲	−۳/۵۷۹	۰/۱۷۶	۰/۰۸۹	۰/۴۴۵	۰/۰۰۷
Content 3 (algebra)													
OP	۲۰۸	۱/۳۵۶	۰/۲۴۷	۲/۸۸۹	۰/۵۳۸	۰/۰۳۲۴	۱/۰۱۸	۳/۶۸۹	−۳/۴۰۹	۰/۱۷۴	۰/۷۵۴	۰/۴۸۹	۰/۰۰۴
ROP_4	۱۹۹	۱/۷۶۹	۰/۲۹۸	۳/۰۰۶	۰/۹۱۸	۰/۰۰۴۲۸	۱/۰۰۱	۳/۵۸۲	−۳/۵۵۲	۰/۱۸۱	۰/۰۶۸	۰/۴۹۲	۰/۰۰۲
ROP_5	۱۹۲	۲/۰۲۳	۰/۲۹۹	۲/۷۸۶	۰/۹۴۸	−۰/۰۰۴۶۱	۱/۰۱۴	۳/۹۹۸	−۳/۱۸۴	۰/۱۸۴	۰/۰۶۷	۰/۴۸۳	۰/۰۰۱
ROP_6	۱۸۸	۱/۷۲۹	۰/۲۷۴	۲/۵۲۳	۱/۱۰۴	−۰/۰۴۹۶	۱/۱۰۲	۳/۹۹۳	−۳/۴۹۵	۰/۱۸۳	۰/۰۶۹	۰/۴۶	۰/۰۰۱۵

جدول ۱۱. خلاصه‌ی آماره‌های عملکرد خزانه سؤال با S-H (b-bin=0.2)، با تعادل محتوا

آماره‌ها	OP	R	MRP	MTI
Bias	۰/۰۰۵۲	۰/۰۰۱۲	۰/۰۰۰۱۱	۰/۰۰۰۷۲
MSE	۰/۰۱۷۴۵	۰/۰۱۲۶۵	۰/۰۰۷۳	۰/۰۱۲۲۴
کجی نرخ مواجهه	۹۸/۸۵۶	۳۰/۲۵۲	۲۶/۷۴۱	۲۶/۹۹۴
نرخ همپوشی سؤال	۰/۵۱۷۳	۰/۲۲۴۳	۰/۲۳۶۷	۰/۲۵۰۱
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۸/۱۶۹٪	۵/۱۵۳٪	۴/۶۳۵٪	۳/۸۶۲٪
درصد سؤال‌های با نرخ مواجهه کوچک‌تر از 0.02	۵۱/۵۲۶٪	۲۳/۰۸۶٪	۲۴/۴۲٪	۱۷/۳۲٪
درصد تست‌هایی که از قیود تست تخطی دارند	۵۴/۸٪	۳/۴٪	۲/۵٪	۲/۶٪
اندازه خزانه سؤال	۹۲۱	۶۹۹	۶۷۷	۶۰۸

جدول ۱۲. درصد سؤال‌ها بیش مواجهه شده و کم مواجهه شده در هر محتوا برای خزانه‌های سؤال با S-H

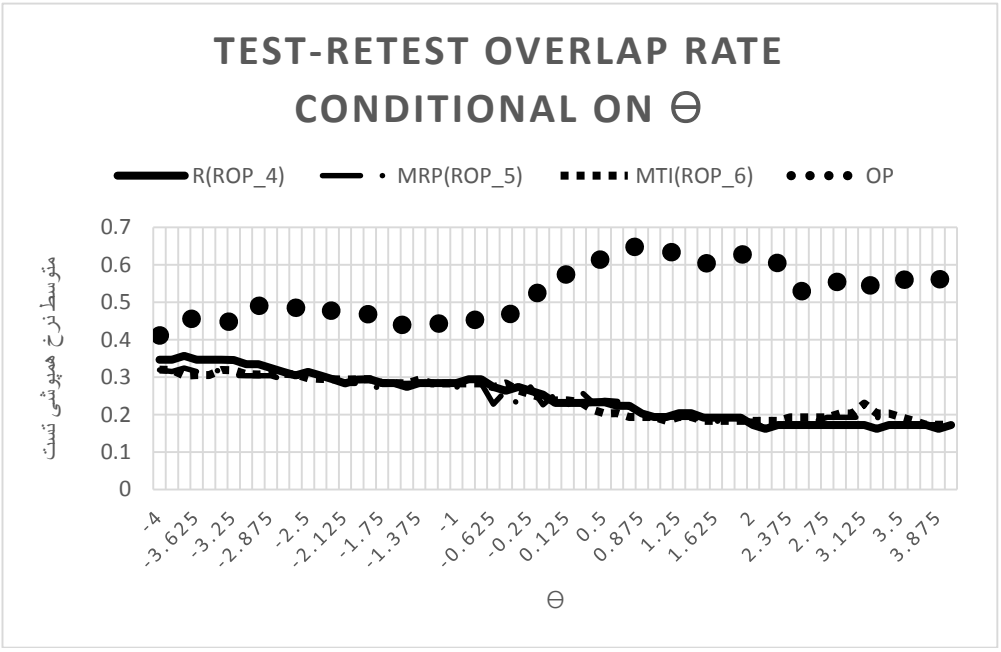
آماره	محتوای ۱				محتوای ۲				محتوای ۳			
خزانه‌های سؤال	OP	R	MRP	MTI	OP	R	MRP	MTI	OP	R	MRP	MTI
درصد سؤال‌های با نرخ مواجهه بزرگ‌تر از $\frac{1}{3}$	۹/۸۱٪	۵/۶۴٪	۴/۸۷٪	۴/۱۸۹٪	۷/۵۸٪	۵/۵۲٪	۴/۹۲٪	۳/۷۵٪	۷/۰۸٪	۵/۲۶٪	۴/۱۶٪	۳/۹۲٪
اندازه خزانه	۴۵۵	۲۶۱	۲۴۹	۲۱۴	۲۵۸	۲۳۹	۲۳۶	۲۰۶	۲۰۸	۱۹۹	۱۹۲	۱۸۸

جدول ۱۳. تعداد تست‌ها بر اساس تخطی از قیود محتوایی در CAT ها، بدون عامل کنترل مواجهه S-H

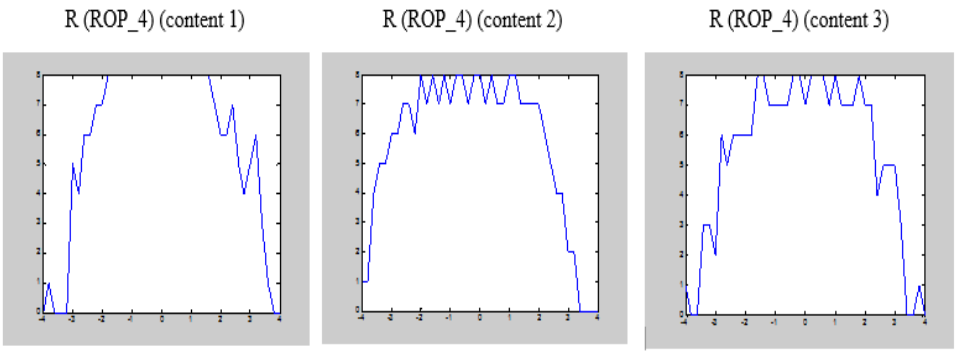
سطوح توانایی / خزانه سؤال	CAT total	-۴	-۳/۵	-۳	-۲/۵	-۲	-۱/۵	-۱	-۰/۵	۰	۰/۵	۱	۱/۵	۲	۲/۵	۳	۳/۵	۴	کل
OP	۳۵۰	۱۱	۱۴	۱۸	۱۴	۱۴	۱۴	۳	۱	۱	۲	۴	۷	۱	۱۱	۲۵	۲۵	۲۸	۱۹۳
R (ROP_4)	۶۰۰۰	۲۹	۲۲	۱۷	۱۲	۸	۳	۳	۳	۶	۲	۱۴	۶	۹	۹	۱۲	۱۸	۳۱	۲۰۴
MRP (ROP_5)	۶۰۰۰	۲۹	۲۵	۱۶	۲۴	۱۶	۱۱	۴	۵	۰	۰	۹	۴	۳	۰	۲	۲	۰	۱۵۰
MTI (ROP_6)	۶۰۰۰	۲۷	۲۴	۱۹	۱۵	۱۲	۶	۱۲	۶	۰	۵	۴	۴	۳	۰	۵	۷	۹	۱۵۶

جدول ١٤. درصد تست‌ها بر اساس تخطی از قیود محتوایی در CAT ها، بدون عامل کنترل مواجهه S-H

سطوح توانایی / خزانه سؤال	CAT total	-۴	-۳/۵	-۳	-۲/۵	-۲	-۱/۵	-۱	-۰/۵	۰	۰/۵	۱	۱/۵	۲	۲/۵	۳	۳/۵	۴	کل
OP	۳۵۰	۰/۰۳	۰/۰۴	۰/۰۵	۰/۰۴	۰/۰۴	۰/۰۴	/۰۰۸	/۰۰۱	/۰۰۱	/۰۰۴	۰/۰۱	۰/۰۲	/۰۰۴	۰/۰۳	۰/۰۷	۰/۰۷	۰/۰۸	/۵۴۸
R (ROP_4)	۶۰۰۰	۰/۰۰۵	۰/۰۰۳۶	۰/۰۰۲۸	۰/۰۰۲	۰/۰۰۱۳	۰/۰۰۰۵	۰/۰۰۰۵	۰/۰۰۱	۰/۰۰۰۵	۰/۰۰۰۳	/۰۰۲۳	۰/۰۰۱	۰/۰۰۱۵	۰/۰۰۱۵	۰/۰۰۲	۰/۰۰۳	۰/۰۰۵	۰/۰۳۴
MRP (ROP_5)	۶۰۰۰	۰/۰۰۵	۰/۰۰۴۱	۰/۰۰۲۶	۰/۰۰۴	۰/۰۰۲۶	۰/۰۰۱۸	۰/۰۰۰۶	۰/۰۰۰۸	۰	۰	۰/۰۰۱۵	۰/۰۰۰۶	۰/۰۰۰۵	۰	۰/۰۰۰۳	۰/۰۰۰۳	۰	۰/۰۲۵
MTI (ROP_6)	۶۰۰۰	۰/۰۰۴۵	۰/۰۰۰۴	۰/۰۰۳۱	۰/۰۰۲۵	۰/۰۰۰۲	۰/۰۰۰۱	۰/۰۰۰۲	۰/۰۰۱	۰	۰/۰۰۰۸	۰/۰۰۰۳	۰/۰۰۰۶	۰/۰۰۰۵	۰	۰/۰۰۰۸	۰/۰۰۱۱	۰/۰۰۱۵	۰/۰۲۶

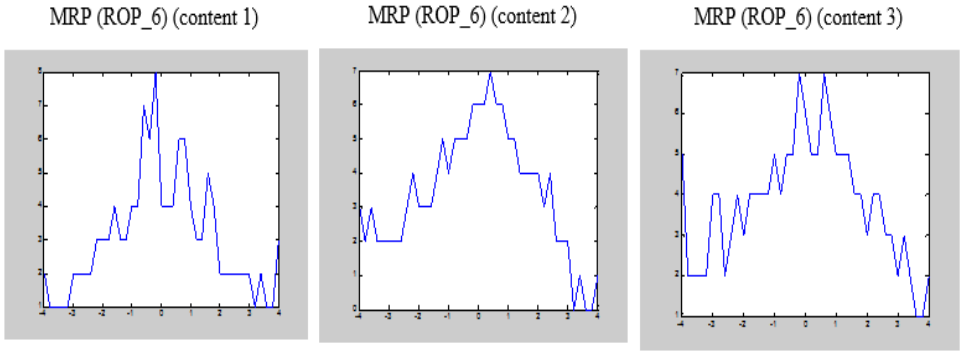


نمودار ۱۰. نرخ همپوشی تست مشروط به با کنترل S-H (*b*-bin: 0.2)، با تعادل محتوا

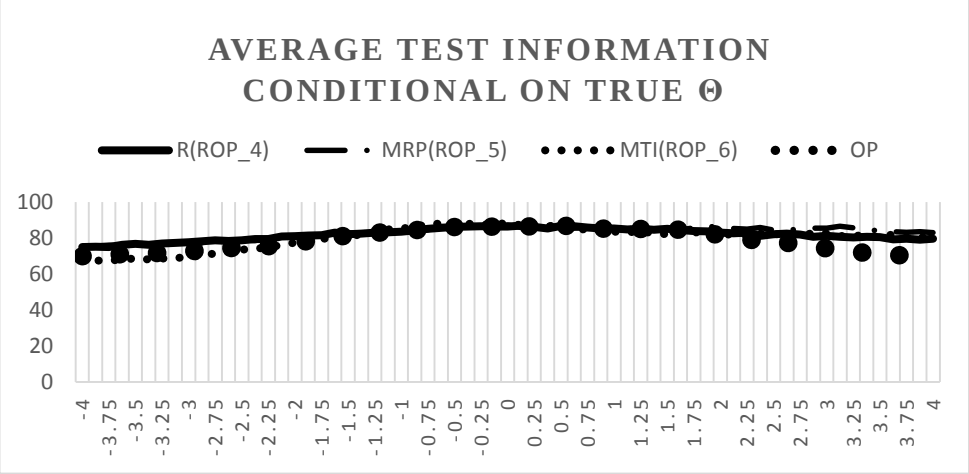


نمودار ۱۱. نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده بر اساس روش R

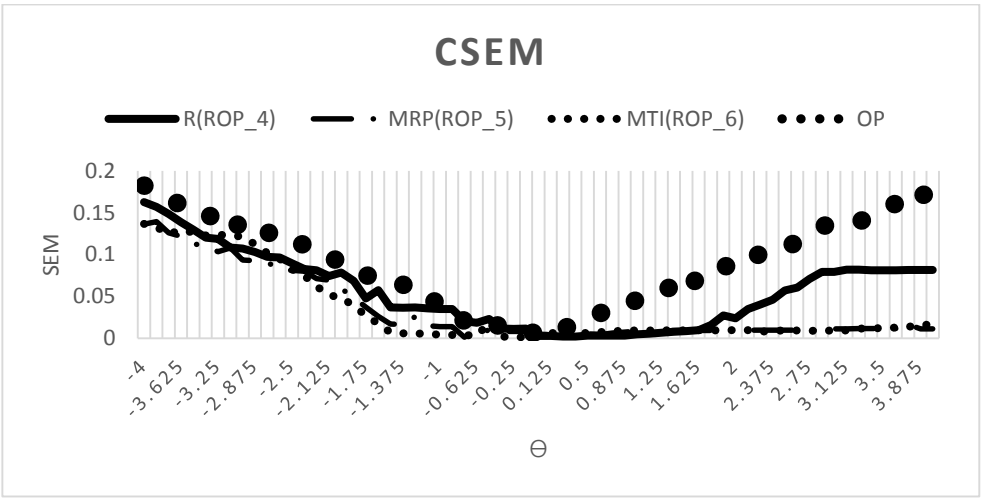
نمودار ۱۲: نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده بر اساس روش MRP



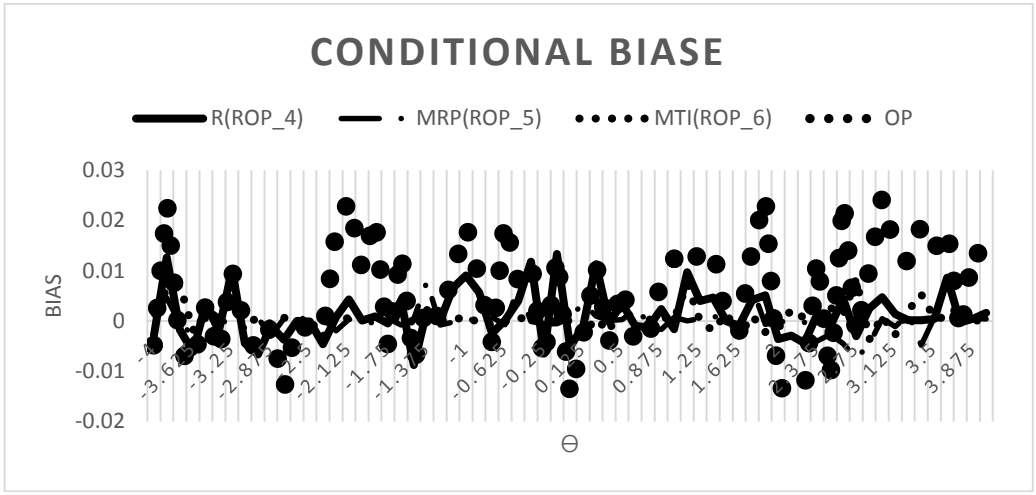
نمودار ۱۳. نمودارهای مربوط به درصد سؤال‌ها بیش مواجهه شده بر اساس روش MTI



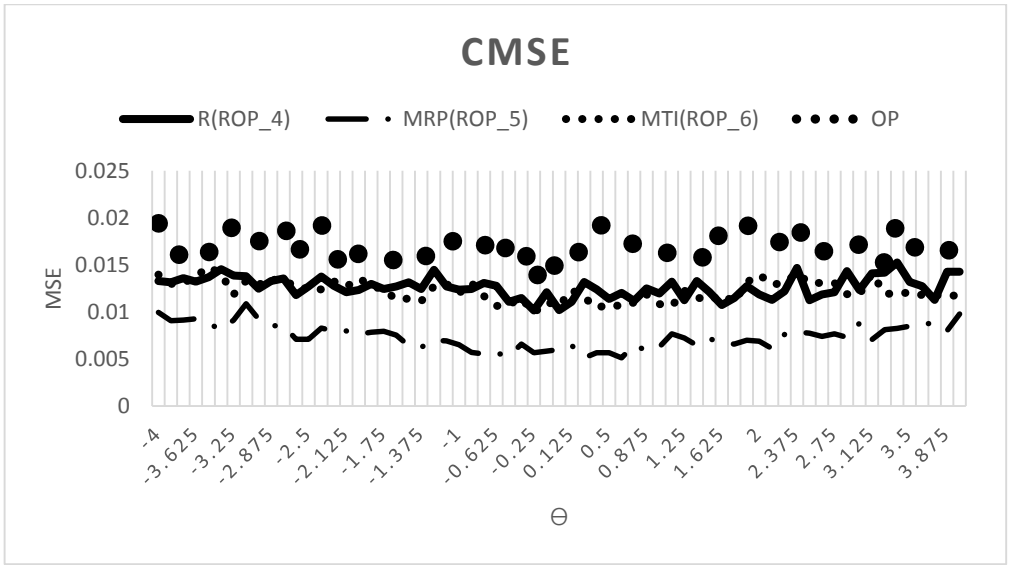
نمودار ۱٤. متوسط آگاهی تست مشروط به توانایی واقعی در خزانه‌های سؤال با S-H (*b*-bin: 0.2)، با تعادل محتوا



نمودار ۱۵. خطای استاندارد اندازه‌گیری (CSEM) در خزانه‌های سؤال با S-H (*b*-bin: 0.2)، با تعادل محتوا



نمودار ۱۶: اریب شرطی (conditional Bias) در خزانه‌های سؤال
با S-H (*b*-bin: 0.2)، با تعادل محتوا



نمودار ۱۷. میانگین مجذور خطا (CMSE) در خزانه‌های سؤال
با S-H (*b*-bin: 0.2)، با تعادل محتوا

نتیجه‌گیری

هدف این پژوهش، ساخت خزانه‌های سؤال بهینه بر اساس رویکرد اکتشافی ریکیسی، با در نظر گرفتن تعادل محتوایی و نرخ مواجهه سؤال‌ها در آزمون بوده است. خزانه‌هایی که با توجه به مدل سه پارامتری لجستیک مدرج شده‌اند. یکی از پیشنهادهای پژوهشی ریکیسی در تحقیقات خود این بود که؛ تعادل محتوایی متنوع یکی از مهم‌ترین مؤلفه‌ها در سنجش انطباقی است که می‌توان آن را در الگوریتم‌های انتخاب سؤال CAT گنجاند (هی و ریکیسی، ۲۰۱۰؛ ۲۰۱۱؛ گو و ریکیسی، ۲۰۰۷)، بنابراین در این پژوهش این عامل وارد برنامه‌ی شبیه‌سازی شد، البته لازم به ذکر است که در تحقیقات آن‌ها این عامل در نظر گرفته می‌شد، ولی از طریق پارتیشن‌بندی کردن خزانه بهینه به خزانه‌های کوچک‌تر این کار صورت می‌گرفت و هیچ‌کدام از این تحقیقات این عامل را به‌صورت کمّی وارد تحلیل نکردند. در این پژوهش تعادل محتوایی سؤال‌ها از طریق مدل حداقل انحرافات وزن‌دار ایجاد شد. به‌طوری‌که از پیش، کدهای محتوایی سؤال‌ها توسط متخصصین موضوعی، مشخص و به همراه ویژگی‌های آماری وارد برنامه شد. در این پژوهش با دست‌کاری دو عامل: روش ایجاد سؤال بهینه (R، MRP، MTI)، کنترل یا عدم کنترل مواجهه بیش‌ازحد سؤال با روش سیمپسون-هتر (S-H)، ۶ مدل طراحی خزانه سؤال بهینه (ROP_1, ROP_2, ROP_3, …, ROP_6) ایجاد شد. همه خزانه‌های سؤال بهینه‌ای که در این پژوهش طراحی شد، صرف‌نظر از عواملی چون کنترل مواجهه و روش ایجاد سؤال بهینه، عملکرد بهتری نسبت به خزانه‌های عملیاتی داشتند. دلیل این امر این است که درمجموع، الگوهای خزانه سؤال بهینه در جستجوی مطلوب‌ترین و مناسب‌ترین ترکیب سؤال‌ها برای تشکیل یک خزانه سؤالی هستند که از طریق آن بتوان تعداد زیادی از تست‌های انطباقی را سرهم کرد. بااین‌وجود، در دنیای واقعی خزانه سؤالی وجود ندارد که به‌طور مطلقی بهینه باشد، زیرا به تعداد عوامل و ترکیب‌های متفاوتی از سؤال‌های موجود در خزانه محدود می‌شود. این دلایل باعث می‌شود که هریک از این نوع خزانه‌ها دارای صحت و دقت اندازه‌گیری متفاوتی باشند و هر یک ازلحاظ بهینه بودن کاملاً از یکدیگر

متفاوت باشند. امّا، در کل، هدف کلی برای الگوهای خزانه سؤال بهینه این است که دارای سه ملاک مهم باشند که توسط وندرلیندن (a۲۰۰۰) ارائه شده است. اولین ملاک این است که به‌اندازه کافی برای سرهم کردن چندین هزار خرده آزمون همپوش، بزرگ باشد. استوکینگ (۱۹۹۴)، مباحث متنوعی در مورد اندازه خزانه سؤال در مورد آزمون‌های ورودی سرنوشت‌ساز که به شکل CAT اجرا می‌شود، مطرح کرد. این قاعده بیان می‌کرد که یک خزانه سؤال CAT برای آزمون‌های سرنوشت‌ساز، ۶ تا ۱۲ برابر طول آزمون CAT باشد. در این پژوهش، اندازه خزانه‌های سؤال بهینه از ۴۹۶ تا ۶۹۹ سؤال در نوسان بود، هیچ‌کدام از خزانه‌ها بیشتر از ۷۲۰ ($60*12=720$) سؤال نداشتند؛ بنابراین، به‌عبارت‌دیگر، روش bin-and-union به‌اندازه زیادی با توصیه استوکینگ برای ساخت یک خزانه سؤال با اندازه کافی برای یک برنامه CAT سازگار است. این نتایج نشان می‌دهد که ملاک اولی که وندرلیندن در مورد خزانه‌های سؤال بهینه مطرح کرد، در مورد خزانه‌های سؤال بهینه در پژوهش حاضر برقرار است. در مورد ملاک دوم که در مورد حدود دامنه سطوح دشواری سؤال‌های خزانه است، باید ذکر کرد که همه خزانه‌های سؤال بهینه طراحی‌شده در این پژوهش در دامنه وسیعی از سطوح دشواری سؤال پراکنده شده‌اند. به‌طوری‌که توزیع پارامترهای b در تمام ROP ها به‌صورت یکنواختی در سراسر مقیاس توانایی توزیع شده است، این توزیع برخلاف توزیع نرمال زیربنایی توانایی آزمودنی‌هایی است که آزمون برای آن‌ها شبیه‌سازی شد. به‌عبارت‌دیگر، پارامترهای b به‌طور یکنواختی در طول مقیاس مهارت توزیع شدند، به‌طوری‌که با توزیع توانایی استفاده‌شده در شبیه‌سازی مطابقت ندارند این ویژگی نیز در تحقیقات هی و ریکیسی (۲۰۱۰) مورد تأیید قرار گرفت، زیرا ماهیت CAT ایجاب می‌کند که توزیع پارامتر b در طول پیوستار توانایی یکنواخت باشد. در کل، ویژگی‌های پارامتر b سؤال پیشنهادشده در این مطالعه با نتایج یوری (۱۹۷۷) و جنسما (۱۹۷۷) نیز سازگار است. ملاک سوم، هزینه ساخت و طراحی سؤال‌ها است که در این مطالعه به‌طور مستقیمی برآورد نشده است. بااین‌وجود، یک مکانیسم مکنون برای برآورد غیرمستقیم هزینه نوشتن سؤال‌ها و تلاش برای به حداقل رساندن آن، در روش طراحی خزانه سؤال ریکیسی مورداستفاده قرار گرفته

است. ایجاد سؤال‌ها بهینه با روش‌های R، MRP و MTI به رابطه بین پارامترهای سؤال و هزینه طراحی سؤال، به‌طور همزمان توجه می‌کند. البته تفاوت این سه شیوه ایجاد پارامترهای سؤال در نگاه اول در مفروضات آن‌ها است. روش MTI بر این فرض استوار است که سؤال‌هایی با ضرایب تشخیص بالا (با مقدار a بالا) هزینه طراحی بالایی دارند؛ بنابراین، روش MTI تلاش می‌کند تا تعداد سؤال‌هایی با ضرایب تشخیص بالا را از طریق شبیه‌سازی کردن سؤال‌هایی که شرط حداقل آگاهی تست را داشته باشند، محدود کند (گو و ریکیسی، ۲۰۰۷). در دو روش دیگر، یعنی روش R و قسمت R روش MRP، بر تصادفی سازی رابطه بین پارامترها توجه دارند؛ بنابراین ازلحاظ هزینه طراحی به‌صرفه نیستند، چون به تعداد بالایی سؤال نیاز خواهند داشت و سؤال‌هایی با توزیع یکنواخت پارامتر دشواری و تشخیص ایجاد می‌کند. امّا روش P یا پیش‌بینی (در روش MRP)، فرض می‌کند که طراحی سؤال‌هایی با مشخصات معین از قبل تعیین‌شده و موردنیاز، کم‌هزینه‌تر از طراحی سؤال‌ها بدون در نظر گرفتن این ویژگی‌ها است. این روش هزینه ایجاد سؤال را از طریق مدل‌یابی کردن مشخصات سؤال (مانند، رابطه بین پارامترهای IRT سؤال‌ها) به حداقل می‌رساند و سؤال‌هایی شبیه‌سازی می‌کند که مشابه با سؤال‌های موجود در خزانه عملیاتی است. همچنین، بهتر است که خزانه‌های سؤال برای آزمون‌های CAT، دقت اندازه‌گیری مشابهی در سراسر دامنه توانایی ایجاد کنند. بررسی دقیق‌تر به خزانه عملیاتی نشان می‌دهد که سؤال‌هایی با ضرایب تشخیص بالاتر در دامنه پارامتر b بین ۱/۷۲ تا ۴ قرار می‌گیرند. در عمل، زمانی که از خزانه عملیاتی به‌طور فراوانی استفاده می‌شود، سؤال‌هایی با ضرایب تشخیص بالا که به‌احتمال زیادی توسط الگوریتم‌های CAT انتخاب می‌شوند، کنار گذاشته می‌شوند، طراحی سؤال‌های با چنین ضرایب تشخیص بالایی برای جایگزینی با سؤال‌ها قبلی بسیار دشوار است؛ بنابراین، این نتیجه ممکن است این شک را ایجاد کند که آیا در این مرحله عملکرد یکسانی بر روی سطوح توانایی یکسان با مراحل قبل، می‌تواند به‌آسانی تکرار شود. بااین‌وجود، خزانه‌های بهینه‌ای که از طریق روش ریکیسی طراحی می‌شوند، دارای سؤال‌ها بیشتری با ویژگی‌های یکنواخت در طول پیوستار توانایی هستند، درنتیجه این نوع خزانه‌ها عملکرد بهتری ازنظر دقت و صحت طبقه‌بندی، در

حداکثر سطوح توانایی مکنون ایجاد می‌کنند. درمجموع، خزانه‌های سؤالی که به‌طور بهینه طراحی می‌شوند، عملکرد بهتری نسبت به خزانه‌های سؤال عملیاتی در شاخص‌های ارزیابی خزانه‌های سؤال ایجاد می‌کنند. این نتایج، مشابه با تحقیقات یوری (۱۹۷۷)، گو و ریکیسی (۲۰۰۷)، هی و ریکیسی (۲۰۱۰) بوده است، در این پژوهش‌ها نیز این ویژگی‌ها به‌عنوان ویژگی‌های برتر خزانه‌های سؤال CAT توصیه شدند. پارامترهای a، در همه ROP ها در حدود ۱/۱ با حداقل مقداری بیشتر از ۸/. به اوج خود می‌رسند. نتایج حاصل از خروجی این شبیه‌سازی‌ها حاکی از آن است که تلفیق روش مونت‌کارلو ریکیسی و روش برنامه‌نویسی حداقل انحرافات وزن‌دار (WDM) امکان‌پذیر است. همچنین، با توجه به این‌که ملاک‌های وندرلیندن (۲۰۰۰a) در مورد خزانه‌های بهینه برای CAT در این پژوهش وجود داشته است، این شبیه‌سازی با موفقیت انجام گرفته است.

همچنین، نتایج نشان می‌دهد که بدون توجه به عامل کنترل مواجهه S-H، خزانه‌های سؤال بهینه‌ای که با کنترل تعادل محتوایی طراحی می‌شوند، دارای دقت اندازه‌گیری خوبی نسبت به خزانه عملیاتی هستند. زمانی که عامل مواجهه سؤال نیز کنترل می‌شود، این میزان دقت اندازه‌گیری بیشتر می‌شود. تعامل دو عامل تعادل محتوایی و کنترل مواجهه S-H نتایج جالبی در مورد دقت اندازه‌گیری به‌دست آورد: میزان اریب در هر چهار خزانه مثبت 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 (ROP_3) دارای بهترین عملکرد ازنظر دقت اندازه‌گیری است. خزانه (ROP_2) MRP و (ROP_1) R دارای خطاهای اندازه‌گیری تقریباً مشابهی در زمانی که عامل S-H در شبیه‌سازی نمی‌شود هستند. امّا زمانی که عامل S-H وارد می‌شود، خزانه (ROP_6) MTI دارای دقت اندازه‌گیری کمتری نسبت به خزانه MTI (ROP_3) است، زیرا سؤال‌هایی که دارای میزان آگاهی متناسب با دامنه توانایی موردنظر و کد محتوایی مناسب هستند، بیشتر از ۰/۳۳ ارائه‌شده و برای کنترل این قضیه، برنامه مجبور است که از سؤال‌های bin های هم‌جوار استفاده کند که باعث می‌شود دقت اندازه‌گیری کاهش یابد. در مقابل، زمانی که عامل S-H وارد می‌شود خزانه MRP (ROP_5) دارای بهترین عملکرد ازنظر

خطای اندازه‌گیری است. درمجموع، هر سه خزانه بهینه در هر دو مرحله بهتر از خزانه‌های سؤال عملیاتی ازنظر اندازه خزانه، دقت اندازه‌گیری و نرخ مواجهه سؤال‌ها عمل می‌کنند. در کل، بررسی دقیق‌تر نمودارهای مربوط به‌دقت اندازه‌گیری در هر یک از سطوح توانایی نشان می‌دهد که خزانه‌های سؤالی که با کنترل مواجهه سؤال طراحی می‌شوند، دارای دقت اندازه‌گیری بیشتری نسبت به خزانه‌هایی که بدون کنترل مواجهه طراحی می‌شوند، هستند. این نتیجه به دلیل این است که سؤال‌های اضافه‌شده به خزانه‌های بهینه با کنترل مواجهه S-H دارای سؤال‌هایی با ضرایب تشخیص بالاتری هستند. در کل، به نظر می‌رسد که بدون در نظر گرفتن عامل S-H، خزانه‌های MTI از سؤال‌های موجود در خزانه استفاده بیشتری می‌کنند و دارای حداقل سؤال‌های کم مواجهه شده است. همچنین، از نرخ همپوشی تست کمی- باوجوداینکه دارای حداقل تعداد سؤال هستند، برخوردار است. درمجموع، بدون توجه به عامل S-H، خزانه‌های بهینه MTI نسبت به خزانه‌های R و MRP دارای سؤال‌ها کمتری هستند و از نرخ مواجهه کمتری نیز برخوردارند و از سؤال‌ها استفاده بیشتری می‌کنند. زمانی که عامل S-H در شبیه‌سازی وارد می‌شود، خزانه(ROP_5) MRP دارای دقت اندازه‌گیری بالاتری است ولی ازنظر اقتصادی به‌صرفه نیست؛ بنابراین، توصیه می‌شود که زمانی که به‌صرفه بودن طراحی خزانه‌های سؤال، تعادل محتوایی و امنیت آزمون ازلحاظ نرخ مواجهه سؤال، عامل بسیار مهمی می‌باشند، برای کاهش تعداد سؤال‌ها موردنیاز در خزانه CAT از روش MTI استفاده شود. البته نکته دیگری که باید به آن توجه کرد این است که زمانی که 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 است، این نتیجه دلیلی بر این امر است که واردکردن کنترل مواجهه سیمپسون-هتر بر انتخاب سؤال‌ها تأثیر می‌گذارد و این امکان وجود دارد که برنامه‌ی CAT، سؤالی را برای اجرا انتخاب کند که از قیود محتوایی تخطی دارد؛ بنابراین، با توجه به این مباحث، می‌توان در مورد چگونگی نحوه عملکرد خزانه‌های سؤال بهینه در هر دو حالت واردکردن و همچنین، وارد نکردن عامل S-H در شبیه‌سازی، نتیجه‌گیری کرد. البته تصمیم‌گیری نهایی در مورد آن به موقعیت استفاده از خزانه برمی‌گردد.

درمجموع، عملکرد خزانه‌های سؤال بهینه در شاخص‌های ارزیابی تجربی بهتر از خزانه سؤال عملیاتی بوده است.

نتایج این پژوهش نشان داد که تعمیم رویکرد ریکیسی در طراحی خزانه‌های سؤال در موقعیت‌هایی که سؤال‌ها با مدل سه پارامتری مدرج می‌شوند و تلفیق آن با رویکرد برنامه‌نویسی ریاضی به‌خوبی عمل می‌کند. در مقایسه با رویکرد برنامه‌نویسی ریاضی، رویکرد ریکیسی شیوه CAT را به‌صورت سرراست‌تری شبیه‌سازی می‌کند. رویکرد ریکیسی با شیوه‌های متفاوت انتخاب سؤال و فرآیند برآورد توانایی مطابقت کامل دارد و دارای انعطاف بیشتری نسبت به رویکرد برنامه‌نویسی ریاضی است. همچنین کاربرد این روش بسیار آسان‌تر است و نیاز به نرم‌افزارهای پیچیده ندارد. در این رویکرد قیود مربوط به صفات غیر آماری از قبیل صفات محتوایی، در مرحله‌ی اول طراحی خزانه وارد برنامه‌نویسی می‌شود و خزانه‌ها را به بخش‌های کوچک‌تری بخش‌بندی می‌کند (هی و ریکیسی، ۲۰۱۰). در مقابل رویکرد برنامه‌نویسی ریاضی ساختاربندی ریاضی بیشتری دارد و تمام قیود آماری و غیر آماری را به‌صورت کمّی درمی‌آورد و سپس بهترین راه‌حل بهینه را از طریق برنامه‌نویسی خطی جستجو می‌کند (وندرلیندن، ۲۰۰۵a). امّا این روش به کاربرد رویکرد انتخاب سؤال "تست سایه" در شبیه‌سازی CAT نیاز دارد. رویکرد ریکیسی روی تصادفی سازی کردن پارامترهای سؤال در شبیه‌سازی تأکید دارد، درحالی‌که رویکرد برنامه‌نویسی ریاضی روی بهینه‌سازی سؤال‌های "ساختگی[1]" از قبل تعریف‌شده تأکید دارد. امّا دو رویکرد در پایان کار شبیه‌سازی به نتایج یکسانی می‌رسند. رویکرد ریکیسی در بعضی جنبه‌ها مشابه رویکرد برنامه‌نویسی ریاضی است. یکی از شباهت‌های مهم بین این دو رویکرد این است که روش شبیه‌سازی خزانه سؤال P (در MRP) و رویکرد برنامه‌نویسی ریاضی در مورد کاهش هزینه‌های طراحی مانند یکدیگر عمل می‌کنند، به‌طوری‌که هر دو فرایند طراحی خزانه سؤال هزینه‌ها را به حداقل می‌رسانند. رویکرد برنامه‌نویسی ریاضی تابع هزینه را که معکوس تعداد سؤال‌ها واقعی با

1. pseudo

ترکیب معینی از صفات شامل پارامترهای IRT سؤال‌ها است، تعریف می‌کند. این رویکرد فرض می‌کند که ایجاد سؤال‌ها واقعی با ترکیب مشخص پارامترهای سؤال، هزینه کمتری دارند. این ایده مشابه روش P است که در آن روش در فرایند شبیه‌سازی به‌احتمال بیشتری سؤال‌ها در طول خط رگرسیون پارامترهای b روی پارامترهای a سؤال‌ها واقعی، ایجاد می‌شود (گو و ریکیسی، ۲۰۰۷). هریک از این دو رویکرد بر همین اساس می‌توانند ایده‌های مشابهی از این نوع را از یکدیگر اقتباس کنند تا طراحی خزانه سؤال را بهبود دهند. در این پژوهش، ویژگی‌های برتر برنامه‌نویسی ریاضی شامل کدگذاری دقیق ریاضی ویژگی‌های غیر آماری در ارتباط با ویژگی‌های آماری با رویکرد ریکیسی تلفیق شده است. برای دوری از دشواری‌های روش انتخاب سؤال تست سایه نیز از مدل انتخاب سؤال حداقل انحرافات وزن‌دار (WDM) استفاده شده است. نتایج نشان داد که از طریق این روش می‌توان در سرهم کردن تست‌های CAT میزان تخطی از قیود محتوایی را به حداقل رساند؛ بنابراین، نتایج این تحقیق نشان داد که تلفیق رویکرد برنامه‌نویسی WDM به همراه رویکرد اکتشافی ریکیسی، نتایج مطلوبی ایجاد می‌کند و با کمّی سازی ویژگی‌های محتوایی به همراه ویژگی‌های روان‌سنجی سؤال‌ها، می‌توان میزان تخطی تست‌ها را از قیود محتوایی کاهش داد. این روش ویژگی‌های آماری و غیر آماری سؤال را با تعادل مطلوبی بین ویژگی‌های اندازه‌گیری و ساختاری در نظر می‌گیرد. در این پژوهش، این ویژگی‌ها به‌وسیله وزن‌هایی که توسط طراحان اولیه تست انتخاب شد، در مدل وارد شد. این روش برخلاف روش تست سایه، ویژگی‌های محتوایی را به‌عنوان اهداف به‌جای قیود فرمول‌بندی می‌کند. انحراف از اهداف محتوایی وزن داده‌شده و در تابع هدف به همراه فاصله آگاهی سؤال از مقدار هدف قرار می‌گیرد. از این طریق می‌توان مدل‌های غیرقابل‌حل را قابل‌اجرا کرد و تست‌هایی با مقدار تخطی کمتری ایجاد کرد. این نتایج هم‌راستا با تحقیقات بروک، کندریک و مروس، ۱۹۹۸؛ استوکینگ، سوانسون و پیرمن، ۱۹۹۳ است.

منابع

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

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.

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

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.

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. (1972). An application of latent trait mental test theory to the Washington Pre-College Testing Program. Unpublished doctoral dissertation. University of Washington, 1972.

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.

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.

Reckase, M. D. (1974). An application of the Rasch simple logistic model to tailored testing. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Reckase, M. D. (1976). The effect of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the spring meeting of the Psychometric Society, Murray Hill, NJ.

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, September). 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., & 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. (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, 2010 (2), 127-141.

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. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.

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

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (No. ETS- RR-94-5): Educational Testing Service, Princeton, NJ.

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.

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. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195-211.

Van der Linden, W. J., & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22 (3), 259-270.

Van der Linden, W. J., & Glas, C. A. W. (2000 a). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13(1), 35-53.

Van der Linden, W. J. (2000 b). 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 c). Optimal assembly of tests with item sets. Applied Psychological Measurement, 24, 225–240.

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., Adelaide, A., & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, 31(1), 81-100.

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 149–162). The Netherlands: Kluwer Academic Publishers.

Wise, S., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. Psicologica, 21, 135-155. Retrieved from http://www.uv.es/psicologica/articulos1y2.00/wise.pdf.

Xing, D., & Hambleton, R. K. (2004). Impacts of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64(1), 5-21.