Educational Measurement and Evaluation Studies

Print ISSN: 2476-2865 Online ISSN: 2783-0942

National Organization for Educational Testing

Evaluation of Chemistry Questions in Universities and Higher Education Institutes Enrance Exam 2017 Using Item Response Theory

Alireza Karami-Gazafi[1], Sara Mehrabi[2], Farzad Zandi [3]

1.Assistant Professor, Faculty of Basic Sciences, Shahid Rajaee Teacher Training University, Tehran, Iran; (Corresponding Author), Email: ar.karami@sru.ac.ir

2.Department of Chemistry, Faculty of Basic Sciences, Shahid Rajaee Teacher Training University, Tehran, Iran. Email: sara.mehrabi.chemeducation@gmail.com

3.Associate Professor, Department of Educational Psychology, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran. Email: farzadzandi547@yahoo.com

Article Info

Article Type:

Research Article

Received:

2019.04.06

Revised:

2019.08.01

Accepted:

2019.08.07

Published online:

2019.08.08

ABSTRACT

Objective: The aim of this study is to evaluate the chemistry questions of universities and higher education institutes entrance exam in 2017 by using item-response theory.

Methods: This is an applied and descriptive research. Participants' 5,000 answer sheets in Empirical Science group were selected randomly as sample. All parameters were calculated with NOHARM and IRTPRO and EXCEL softwares.

Results: First, the initial assumptions of IRT theory (unidimensionality and local independence) were investigated. The parameters of each questions, such as difficulty and discrimination coefficient, were calculated based on the classic and the item response theories. The results showed that 7 questions show a good fit with one and two-parameter models and 21 questions are also compatible with three-parameter model and 13 questions do not fit with any IRT models. The analysis with 3-PL-IRT showed the discrimination coefficient of 26 questions are strong ($a > 1.3$), 6 are moderate and 3 are weak ($a < 0.65$). Also 11 questions were very difficult ($b > 1.2$) and 24 questions were appropriate ($-1.2 < b < 1.2$). The guessing parameter for 16 questions, were $c \geq 0.2$, which indicates the high predictability of the test.

Conclusion: The results showed that 3-PL IRT model has a better fit with the test. All questions are efficient in terms of discrimination coefficient and have a high level of difficulty and also guessing parameter is high in this test. The test's questions have the highest Information and the least error for a high level of ability (range -1 to 2).

Keywords: Evaluation, University Entrance Exam, Chemistry Education, Classic Theory, Item-Response Theory.

How to Cite: Karami-Gazafi, Alireza; Mehrabi, Sara; Zandi, Farzad (2021). Evaluation of Chemistry Questions in Universities and Higher Education Institutes Enrance Exam in 2017 Using Item Response Theory. *Educational Measurement and Evaluation Studies*, 11 (34): 113-136 Pages. DOI:10.22034/EMES.2021.248201

Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی

سازمان سنجش آموزش کشور

شاپا چاپی: ۲۸۶۵-۲۴۷۶ شاپا الکترونیکی: ۰۹۴۲-۲۷۸۳

ارزشیابی سؤال‌های شیمی کنکور سراسری رشته علوم تجربی در سال ۱۳۹۶ با استفاده از نظریه سؤال- پاسخ

علیرضا کرمی گزافی[۱] سارا مهرابی[۲] فرزاد زندی[۳]

۱. استادیار گروه شیمی، دانشکده علوم پایه، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران؛ (نویسنده مسئول)، پست الکترونیک: ar.karami@sru.ac.ir

۲. گروه شیمی، دانشکده علوم پایه، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران. پست الکترونیک:@sara.mehrabi.chemeducation gmail.com

۳. دانشیار، گروه روان‌شناسی تربیتی، دانشگاه آزاد اسلامی، سنندج، ایران. پست الکترونیک: farzadzandi547@yahoo.com

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت:
۱۳۹۸/۰۱/۱۷

اصلاح:
۱۳۹۸/۰۵/۱۰

پذیرش:
۱۳۹۸/۰۵/۱۶

انتشار:
۱۳۹۸/۰۵/۱۷

چکیده

هدف: در پژوهش حاضر، سؤال‌های شیمی کنکور سراسری گروه علوم تجربی در سال ۱۳۹۶ با استفاده از نظریه سؤال-پاسخ ارزشیابی شده است.

روش پژوهش: روش پژوهش، کاربردی، توصیفی- پیمایشی است. جامعه آماری، شامل همه داوطلبان کنکور سراسری گروه علوم تجربی در سال ۱۳۹۶ و نمونه آماری ۵۰۰۰ پاسخنامه از میان پاسخنامه‌های داوطلبان شرکت‌کننده در این سال است و محاسبات با نرم‌افزارهایIRTPRO ، NOHARMو EXCEL انجام گرفته است.

یافته‌ها: بر اساس نظریه کلاسیک و سؤال-پاسخ، پارامترهای تک‌تک سؤال‌ها محاسبه شد. نتایج نشان داد ۷ سؤال هم با مدل یک‌پارامتری و هم دوپارامتری و ۲۱ سؤال با مدل سه‌پارامتری برازش مناسبی نشان می‌دهند و ۱۳ سؤال نیز با هیچ‌کدام از مدل‌های نظریه سؤال-پاسخ برازش نشان نمی‌دهند. تحلیل سؤال‌ها با مدل سه‌پارامتری نشان داد که ۲۶ سؤال از نظر ضریب تشخیص، قوی (a>1.3)، ۶ سؤال، متوسط و ۳ سؤال، ضعیف (a<0.65)هستند. همچنین از نظر دشواری، ۱۱ سؤال بسیار سخت (b>1.2) و ۲۴ سؤال مناسب (b<1.2>-1.2) بودند. سؤال c≥0.2 بود که بیانگر حدس‌پذیری بالای سؤال‌های شیمی آزمون است.

نتیجه‌گیری: این آزمون ۳۵ سؤالی، دارای ضریب دشواری ۰/۳ و عامل حدس حدود ۰/۲ است و برای سطوح توانایی در گستره توانایی ۰/۵- تا ۲ بیشترین آگاهی و کمترین خطا را دارد. با توجه به فراوانی داوطلبان در این محدوده از توانایی به نظر می‌رسد این آزمون توانایی لازم برای تشخیص داوطلبان مختلف و انتخاب عادلانه آنها برای دانشگاه‌ها را داشته باشد. آلفای کرونباخ آزمون ۰/۹۳۴ و نشان‌دهنده پایایی عالی آزمون است.

واژگان کلیدی: ارزشیابی، کنکور، آموزش شیمی، کلاسیک، نظریه سؤال-پاسخ،

استناد: کرمی گزافی، علیرضا؛ مهرابی ، سارا؛ زندی، فرزاد (۱۴۰۰). ارزشیابی سؤال‌های شیمی کنکور سراسری رشته علوم تجربی در سال ۱۳۹۶ با استفاده از نظریه سؤال–پاسخ فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی، ۱۱(۳۴)، ۱۱۳-۱۳۶ صفحه، . DOI:10.22034/EMES.2021.248201

ناشر: سازمان سنجش آموزش کشور

مقدمه

امروزه نظام آموزشی به‌عنوان ابزاری مهم در رشد و توسعه همه‌جانبه کشور شناخته می‌شود و ارزشیابی یکی از ملاک‌های اصلی این نظام است. ملکی (۱۳۹۲) هدف از ارزشیابی را بهبود فرایند یادگیری و رشد همه‌جانبه دانش‌آموزان می‌داند. درواقع سنجش و ارزشیابی بخش جدایی‌ناپذیر آموزش به شمار می‌رود که بدون استمرار دقیق آن، رسیدن به هدف‌های مورد نظر به‌صورت مطلوب، ناممکن خواهد بود (نفیسی، ۱۳۷۶). کنکور یکی از آزمون‌های مهمی است که همه‌ساله برگزار می‌شود تا داوطلبان ورود به آموزش عالی را ارزیابی و گزینش کند. در کشور ما به دلیل میزان بالای تقاضای آموزش عالی و عرضه محدود آن، آزمون سراسری دانشگاه‌ها به مهم‌ترین رخداد آموزشـــی (بلکه چالش آموزشی کشور) تبدیل شده است. رخدادی که در آن حجم وسیعی از نیروهای انسانی درگیر هستند و روح و روان بسیاری از افراد جامعه، هزینه و منابع بسیاری را به خود اختصاص داده است. همچنین، موجب آسیب‌های مختلف اجتماعی - فرهنگی، آموزشی، اقتصادی شده و از چالش‌های اصلی کشور مطرح شده است (محمودیان، ۱۳۹۰). از نظر اقتصادی، کنکور سراسری یکی از پرهزینه‌ترین آزمون‌هایی است که سالانه برگزار می‌شود. در سال ۱۳۹۶ هزینه ثبت نام به‌ازای شرکت در یک گروه آزمایشی ۲۰۰/۰۰۰ ریال بود، هزینه استفاده از خدمات پیام کوتاه: ۵/۰۰۰ ریال، هزینه اعلام علاقه‌مندی برای دانشگاه‌های پیام نور، غیر انتفاعی و فرهنگیان ۵۱/۰۰۰ ریال و هزینه شرکت در گروه‌های هنر و زبان هر کدام ۲۰۰/۰۰۰ ریال بود. این هزینه‌ها در حالی است که در سال ۱۳۹۶ تعداد کل داوطلبان شرکت‌کننده ۹۲۹۷۹۱ نفر بوده است و این تنها بخش کوچکی از همه هزینه‌های آشکار و پنهانی است که هر ساله در برگزاری این آزمون مهم صرف می‌شود. افزون بر آن، می‌توان شرکت در آزمون‌های آزمایشی، هزینه‌های سرسام‌آور معلمان خصوصی و رده‌های کنکور، تهیه کتاب، سی دی و انواع وسایل کمک‌آموزشی توسط داوطلبان را نیز نام برد که نشان از وجود یک صنعت پول‌ساز در پشت فرایند پذیرش دانشجو در ایران دارد. از طرفی، تاکنون راه عادلانه، نفوذناپذیر و کم‌تبعیض‌تری نســـبت به کنکور سراسری به‌عنوان جایگزین آن معرفی نشده و علی‌رغم تمامی انتقادات هنوز بهترین شیوه ورود به دانشگاه، کنکور سراسری است. بنابراین، انتخاب سؤال‌های این آزمون برای داوطلبان و نیز جامعه علمی آینده کشور بسیار مهم و سرنوشت‌ساز است؛ به همین دلیل، بررسی اصول و قواعد آزمون‌سازی و ویژگی‌های روان‌سنجی سؤال‌های کنکور مهم تلقی می‌شود تا افراد قوی‌تر و مستعدتر شناسایی و پذیرش شوند و از یک‌سو، عدالت در پذیرش رعایت شـــده باشـــد و از سوی دیگر با ورود افراد قوی‌تر، دانشگاه‌ها و بدنه کارشناسی مراکز علمی و خدماتی استخدام‌کننده این افراد ارتقا یابد. سازمان سنجش آموزش ایران با یک چالش بزرگ و سالانه روبه‌روســـت؛ زیرا هر سال باید تعداد ۳۵ سؤال از سرفصل‌های نسبتاً ثابت، محدود و تکراری کتاب‌های درسی شیمی دبیرستان را طراحی و با برگزاری آزمون، افراد شایسته را گزینش و پذیرش کند. سؤال‌های خیلی آسان و زیاد دشوار مناسب نیستند؛ زیرا در حالت اول، همه داوطلبان آن را پاسخ خواهند داد و در حالت دوم، نمره همه داوطلبان در بخش شیمی صفر خواهد بود و بنابراین نمی‌توان بین داوطلبان هیچ تفاوتی قائل شد. میزان

پاسخگویی به سؤال‌های همراه با حدس نیز از ویژگی‌های رتبه‌بندی یک سؤال است و سختی و هنر طراحان سؤال و سازمان سنجش آموزش کشور در این مرحله مشخص می‌شود. اما مسئله جدی‌تر آن است که فرایند طراحی سؤال با ضریب تشخیص بالا، هر سال از سه کتاب شیمی با سرفصل محدود باید تکرار شود و البته این مسئله و چالش مهم، بزرگ و مرتبط با سرنوشت داوطلبان و آینده علمی کشور، برای چندین درس اختصاصی و عمومی دیگر نیز وجود دارد. این مشکلات، اهمیت و سختی کار طراحان سؤال در سازمان سنجش آموزش کشور را نمایان می‌سازد. تفسیر نمره‌های آزمون، بهبود سؤال‌های آزمون و شیوه‌های آزمودن، بازخوردهای سازنده‌ای در اختیار فراگیران، معلمان، آزمون‌سازان و سیاست‌گذاران قرار می‌دهد (عباسی و همکاران، ۱۳۹۲). این تحلیل، همچنین اطلاعات تشخیصی لازم را برای بررسی کیفیت یادگیری دانش‌آموزان و مشکلات آموزشی معلمان فراهم می‌آورد (سپاسی، ۱۳۸۲). بر این اساس، ارزشیابی سؤال‌های هر سال در کنکور سراسری یک راهنمای بسیار خوب برای طراحان سؤال برای سال‌های بعد است تا سؤال‌های مناسب و نامناسب را شناسایی کرده و از تجربه آنها در طراحی سال‌های آینده استفاده کنند؛ از این‌رو هدف از اجرای این پژوهش، ارزشیابی سؤال‌های شیمی کنکور سراسری سال ۱۳۹۶ با استفاده از نظریه جدید و توانمند سؤال-پاسخ است تا با تحلیل دقیق و کامل آزمون، دشواری، ضریب تشخیص، عامل حدس، توانایی و میزان آگاهی سؤال‌ها و آزمون برآورد شود که درنهایت بتوان با این اطلاعات، یک بانک سؤال خوب تهیه کرد و از این سؤال‌ها یا سؤال‌های مشابه برای آزمون‌های آزمایشی، سؤال‌های امتحانات نهایی، کنکور سراسری، شناسایی نقاط ضعف دانش‌آموزان و اصلاح روش‌های تدریس یا بازنگری کتاب‌های درسی بهره‌برداری کرد.

مهم‌ترین نکته در آزمون‌هایی مانند کنکور سراسری، ساخت آزمون و طراحی سؤال‌های مناسب و حرفه‌ای است و اصول و زیربنای ساخت آزمون‌ها را نظریه‌ها و مدل‌هایی تشکیل می‌دهند که خود این نظریه‌ها دارای کارکردها، توانمندی‌ها و پیش‌فرض‌های متفاوتی هستند که از این میان می‌توان به نظریه کلاسیک آزمون و نظریه سؤال-پاسخ اشاره کرد. اصطلاح کلاسیک، علاوه بر اینکه به زمان شکل‌گیری مدل‌های این نظریه اشاره دارد، به تقابل آن با نظریه‌های جدید روان‌سنجی، یعنی مدل‌های نظریه سؤال-پاسخ نیز مربوط می‌شود (ین[1]، ۲۰۰۲). گرچه نظریه کلاسیک آزمون، مدت‌زمان طولانی به جامعه روان‌سنجی خدمت کرده است؛ اما برخی مطالعات ازجمله گالیکسن و همکاران، محدودیت‌هایی را در این نظریه و در آزمون‌های ساخته شده بر اساس آن نشان می‌دهد (دلاور و همکاران، ۱۳۸۵). ساختار و معادله نظریه کلاسیک، بیانگر آن است که مفروضه‌های آن در سطح نمره انفرادی آزمودنی تدوین شده‌اند، اما هرگز در مدل‌های آن نمره‌های انفرادی تحلیل نمی‌شوند بلکه تمرکز اصلی بیشتر بر ویژگی‌های نمره‌های آزمون در ارتباط با گروه نمونه (مجموعه‌ای از افراد) است. محدودیت دیگر نظریه کلاسیک این است که بیشتر آزمون‌مدار است تا پرسش‌مدار؛ درنتیجه تعیین سهم سؤال در پایایی و در پی آن، خطای معیار اندازه‌گیری آزمون امکان‌پذیر نیست (عباسی و همکاران، ۱۳۹۲).

1. Yen

نقاط ضعف نظریه کلاسیک باعث می‌شود تفسیر مناسبی از آزمون‌ها صورت نگیرد؛ بنابراین برای داشتن یک تحلیل آزمون دقیق‌تر و همچنین افزایش کارایی دانش‌آموختگان باید از مدل‌ها و تکنیک‌هایی در نمره‌گذاری آزمون‌ها استفاده کرد که بتواند نقاط ضعف و قوت آنها را نشان دهد. بدون هیچ تردیدی می‌توان گفت که نظریه سؤال-پاسخ جهش بلندی به سوی آرمان‌های سنجش و اندازه‌گیری است. ریشه نظریه سؤال پاسخ را می‌توان به بینت[1] و ترستون[2] در اواخر قرن ۱۸ و اوایل قرن ۱۹ اطلاق کرد (لیندن[3]، ۲۰۱۰). نظریه سؤال- پاسخ، یک نظریه جامع آماری درباره عملکرد سؤال آزمون و آزمودنی و چگونگی سنجش توانایی‌ها است. این نظریه، نقطه مقابل نظریه کلاسیک اســـت و بر نمره وابسته به گروه نمونه و آزمون نیست. همچنین، برآوردهای بهتری از توانایی آزمودنی به دست می‌دهد و اطلاعات بیشتری درباره آزمودنی و سؤال‌ها فراهم می‌آورد. مدل‌های IRT توابع ریاضی هستند که احتمال یک برون‌داد پیوسته، مانند پاسخ درست به یک سؤال را بر اساس پارامترهای آزمودنی (توانایی θ) و پارامترهای سؤال (دشواری، ضریب تشخیص و حدس) مشخص می‌کند. در حال حاضر، IRT یکی از نیرومندترین ابزارهایی است که برای تهیه و تجزیه‌وتحلیل تست‌ها به کار می‌رود و آن‌چنان گسترده و فراگیر شـــده که به‌جرئت می‌توان گفت دوره کلاســـیک آزمون که متکی بر جمع جبری نمره‌هاست به سر آمده است تا آنجایی که بسیاری از پژوهشگران ازجمله لامزدن[4] (۱۹۷۶) و دیویسون[5] (۱۹۸۲) توصیه کرده‌اند که مدل‌های کلاسیک باید کنار روند و مدل‌های نظریه صفت مکنون جایگزین آن شود (بخشی‌فرد، ۱۳۸۱).

مبانی نظری و پیشینه پژوهش

فراهانی (۱۳۷۵)، با بررسی داده‌های حاصل از اجرای سه آزمون ریاضیات، فیزیک و درس فنی که برای ۴۳۰ نفر از داوطلبان کنکور داخلی وزارت نیرو اجرا شده بود به مقایسه نظریات کلاسیک و جدید از لحاظ برآورد پارامترهای سؤال و توانایی پرداخت. نتایج، تفاوت مدل‌ها را در برآورد پارامتر توانایی نشان داد که مدل سه‌پارامتری نسبت به مدل کلاســـیک و حتی مدل‌های یک و دو پارامتری IRT برآوردهـــای متفاوت‌تر و دقیق‌تری ارائه می‌دهد). واله (۱۳۹۲) نیز در پژوهش خود با عنوان «ارزیابی جامع و تعیین ویژگی‌های روان‌ســـنجی آزمون اســـتعداد و آمادگی تحصیلی ویژه رشته مدیریت در کنکور سراسری ورودی مقطع کارشناسی ارشد سال‌های ۹۰-۱۳۸۹» با اســـتفاده از نظریه کلاسیک و پرسش پاسخ نشان داد که مدل کلاسیک و مدل‌های IRT برآوردهای متفاوتی برای پارامترهای ســـؤال به دست می‌دهند و برآورد پارامترهای ســـؤال بر پایه مدل‌هایIRT ، به‌ویژه مدل سه پارامتری، دقیق‌تر از برآورد پارامترهای ســـؤال بر اساس مدل کلاسیک و مدل‌های ساده‌تر IRT است. همچنین محمدزاده رومیانی (۱۳۷۵)، داده‌های حاصل از اجرای ســـه آزمون زیست‌شناسی، فیزیک و ادبیات مرحله اول کنکور سراسری را تحلیل کرد. این مطالعه همراه با پژوهش فراهانی (۱۳۷۵) در مورد داده‌های حاصل از اجرای ســـه آزمون ریاضیات، فیزیک و درس فنی کنکور داخلی وزارت نیرو و نیز ایزانلو (۱۳۸۹) در ارزشیابی داده‌های

1. Binet
2. Thurstone
3. Linden
4. Lumsden
5. Davison

خرده‌مقیاس گرامر در بخش تخصصی دفترچه زبان انگلیسی آزمون سراسری در سال ۱۳۸۴ نشان دادند که مدل‌های IRT نسبت به مدل کلاسیک برآوردهای دقیق‌تری ارائه می‌دهد. دلاور و همکاران (۱۳۸۵)، با اجرای پژوهشی از نظر اندازه‌گیری شاخص‌های دقت آزمون بر روی ۱۰۰۰ نفر نشان دادند که نظریه سؤال-پاسخ از نظر اندازه‌گیری شاخص‌های دقت آزمون، بر نظریه کلاسیک برتری دارد. در نظریه پاسخ – سؤال، برای هر سطح توانایی، آگاهی و خطای استاندارد جداگانه‌ای به دست می‌آید که بر اساس آن می‌توان فهمید آزمون در کدام سطح توانایی، دارای بیشترین دقت و کمترین خطاست؛ مزیتی که در نظریه کلاسیک اندازه‌گیری نیست.

اَشکارا مدل‌های مختلف نظریه سؤال-پاسخ از نظر دقت با یکدیگر برابر نیستند؛ به‌عنوان مثال، احمدی آذر (۱۳۸۷) با استفاده از مدل‌های نظریه سؤال-پاسخ، خصوصیات روان‌سنجی آزمون‌های کنکور سراسری رشته ریاضی را بررسی کرد و دریافت که در تمام آزمون‌های کنکور مدل سه‌پارامتری برازش بهتری با مجموع داده‌ها دارد و بررسی ضرایب تشخیص، دشواری و همچنین توانایی افراد نشان داد که در نظریه سؤال-پاسخ سؤال‌های بیشتری از نظر دشواری یا قدرت تشخیص، نامناسب هستند که این نشان‌دهنده دقت بیشتر این نظریه نسبت به نظریه کلاسیک است. دیوجی[1] (۱۹۸۶)، کاربرد مدل یک پارامتری راش را برای سؤال‌های چندگزینه‌ای بررسی کرده است؛ اما به‌رغم استفاده از مدل راش برای برآورد پارامترها در سؤال‌های چندگزینه‌ای، به نظر دیوجی به علت نبود پارامتر حدس و ضریب تشخیص یکسان در مدل راش، این کاربردها درست نیست. استیج[2] (۱۹۹۸) در پژوهش خود، به مقایسه کارایی دو نظریه کلاسیک و سؤال-پاسخ در آزمون استعداد تحصیلی سوئد اقدام کرد. این آزمون، دارای ۵ آزمون فرعی و حاوی ۱۶ سؤال چهارگزینه‌ای است و برای انتخاب دانشجویان آموزش عالی در سوئد استفاده می‌شود. او نتیجه گرفت که مدل منطقی سه‌پارامتری با داده‌ها تناسب بهتری نشان می‌دهد. مکاین و اوفلیا[3] (۲۰۱۱) در دانشگاه مانیلا با استفاده از یک آزمون محقق‌ساخته ۱۵ سؤالی برای سنجش استعداد یادگیری درس جبر و به‌کارگیری مدل یک‌پارامتری راش پیشنهاد دادند که در پژوهش‌ها بهتر است از مدل‌های دو و سه پارامتری استفاده شود.

اسدی و همکاران (۱۳۹۱) از مدل‌های تک‌پارامتری و سه‌پارامتری در نظریه سؤال-پاسخ برای سنجش توانایی دانش‌آموزان دبیرستان دخترانه توسط ماتریس‌های پیش‌رونده ریون استفاده کردند و به مقایسه تفاوت‌های بین برآورد پارامتر توانایی آزمودنی و برآورد پارامتر توانایی سؤال در مقیاس هوش پیش‌رونده ریون در دو مدل راش و سه‌پارامتری بر اساس تئوری سؤال-پاسخ بر روی ۴۹۸ دانش‌آموز پرداختند. نتایج این بررسی نشان داد: بین برآوردهای توانایی بر اساس دو مدل تک‌پارامتری و سه‌پارامتری تفاوت معنی‌دار وجود دارد اما بین نمره‌های توانایی‌ها در هر دو مدل همبستگی وجود دارد؛ از روی نمره‌های خام افراد می‌توان توانایی آنان را در مدل سه‌پارامتری پیش‌بینی کرد؛ از مقایسه نمره میانگین دو مدل، این نتیجه حاصل می‌شود که مدل تک‌پارامتری

1. Divgi
2. Stage
3. Macayan & Ofalia

نسبت به مدل سه‌پارامتری مقاوم‌تر است

همچنین، بقایی مقدم و روشن‌ضمیر (۱۳۸۹) با ارزشیابی یک آزمون درک مطلب زبان انگلیسی با مدل پرسش-پاسخ چندبعدی دریافتند که مدل چهاربعدی، برازش بهتری نسبت به مدل دوبعدی؛ و مدل دوبعدی برازش بهتری نسبت به مدل تک‌بعدی دارد. نقی‌زاده (۱۳۹۵) در بررسی تعدادی سؤال از مباحث کتاب شیمی سوم متوسطه نشان داد مدل‌های دوپارامتری و یک‌پارامتری در نظریه سؤال-پاسخ با سؤال‌های آزمون تطابق بهتری دارند. کرمی گزافی در بررسی سؤال‌های منتخب از کتاب‌های مؤسسات مبتکران و قلم‌چی از شیمی دوم متوسطه و فیزیک اول متوسطه نتیجه گرفت که از ۱۶۰ سؤال انتخاب‌شده فقط ۹۷ سؤال با مدل IRT دوپارامتری تطابق بهتری دارند و البته سؤالات مناسب‌تری نیز هستند. بنابراین می‌توان از این سؤال‌ها برای ایجاد بانک سؤال در آزمون‌های بعدی استفاده کرد. از طرفی، وجود عامل حدسی در حدود ۴۰ درصد برای برخی سؤال‌ها نشان داد که این سؤال‌ها کاملاً مناسب نیستند (کرمی گزافی، ۱۳۹۳ و ۱۳۹۴). روحانی (۱۳۹۱)، سؤال‌های کارشناسی گروه علوم انسانی دانشگاه آزاد اسلامی در سال‌های ۱۳۸۷ و ۱۳۸۹ را مطالعه کرد و نشان داد فقط در ۲۱ درس از ۳۶ درس گروه علوم انسانی مفروضه نظریه سؤال-پاسخ برقرار است. تمام درس‌های بررسی‌شده توسط نظریه سؤال-پاسخ با مدل سه‌پارامتری برازش بهتری دارند. محاسبه پارامترهای سؤال‌ها بر اساس دو نظریه کلاسیک و سؤال-پاسخ نشان می‌دهد که این دو نظریه، همسو بوده و نتایج یکدیگر را تکمیل و تأیید می‌کنند.

مطالعاتی نیز روی تأثیر تعداد گزینه‌های سؤال در نظریه سؤال-پاسخ صورت گرفته است؛ فلسفی‌نژاد و همکاران (۱۳۹۰)، تأثیر تعداد گزینه‌های سؤال در ویژگی‌های روان‌سنجی آزمون و توانایی برآوردشده در مدل‌های پرسش پاسخ را بررسی کردند. آنها نشان دادند که تعداد گزینه‌ها بر پارامترهای سؤال اثر ندارد و تأثیر تعداد گزینه‌ها بر ویژگی‌های روان‌سنجی برآوردشده آزمودنی‌ها، در آزمون‌های مختلف یکسان است. شیزوکا و همکاران[1] (۲۰۰۶) در تحقیقی در دانشگاه کانسای ژاپن از IRT برای بررسی آثار کاهش تعداد گزینه‌های سؤال‌ها بر مشخصه‌های روان‌سنجی آزمون ورودی دانشگاه کانسای بهره گرفت. او یکی از گزینه‌های نادرست را در آزمون چهارگزینه‌ای، حذف کرد و با تبدیل آن به یک آزمون سه‌گزینه‌ای، روی گروه دیگری اجرا کرد. مقایسه پاسخ دو آزمون نشان داد که درجه سهولت و پارامتر تشخیص به‌طور معنی‌داری تغییر نکرد.

نظریه سؤال-پاسخ تحت تأثیر حجم نمونه است و با کاهش حجم نمونه، دقت و صحت آن کاهش می‌یابد. لرد[2] (۱۹۸۰)، طی یک مطالعه و ضمن مقایسه مدل‌های یک و دوپارامتری IRT در برآورد نمره حقیقی آزمودنی‌ها، تلاش کرد تأثیر حجم نمونه را بررسی کند. داده‌های مطالعه شامل پاسخ ۳۰۰۰ دانش‌آموز کلاس ششم به آزمون خزانه لغات متروپولیتن تجزیه‌وتحلیل شده است. نتایج مطالعه نشان داد وقتی حجم نمونه کوچک باشد، پارامتر ضریب تشخیص سؤال‌ها و پارامتر مجانب یا حدس سؤال‌ها را نمی‌توان به‌دقت تعیین کرد. شومیر و

1. Shizuka
2. Lord

همکاران[1] (۲۰۱۰) با اســـتفاده از تجزیه‌وتحلیل IRT تأثیر جمله‌بندی سؤال‌های امتحانی را بر عملکرد دانش آموزان بررســـی کردند و به این نتیجه رسیدند که تغییرات کوچکی در جمله‌بندی سؤال‌های امتحانی می‌تواند تفاوت معنی‌داری در عملکرد دانش‌آموزان در پاسخ به سؤال‌های امتحانی داشته باشد.

انصارین (۱۳۷۱)، تفاوت برآورد توانایی در مدل دوپارامتری را مطالعه کرد. او با اســـتفاده از داده‌های حاصل از اجرای آزمون هوش تهران-استنفرد- بینه (TSB) و تحلیل منحنی ویژه سؤال‌ها دریافت که نمره‌های خام یکسان دارای برآورد یکسانی از توانایی و موقعیت آزمودنی بر پیوستار مکنون نبودند. هومن (۱۳۶۸) نیز در پژوهشی با استفاده از آزمون تهران-استنفرد-بینه (TSB) برآورد پارامترهای دشواری و توانایی و درواقع توانمندی مدل راش را در برآورد پارامترها در شـــرایط نقض مفروضات بررسی کرد و نتیجه گرفت که مدل راش برای برآورد پارامتر دشواری سؤال‌های با ضریب تشخیص متفاوت مناسب نیست اما برای برآورد توانایی افراد مناسب و خوب است. اســـتفاده از مدل IRT همبســـتگی بین نمره‌های توانایی آزمودنی‌ها در دو آزمون هوش جداگانه نشان داد بین توانایی افراد در دو آزمون جداگانه همبستگی وجود دارد (بخشی‌فرد، ۱۳۸۱).

پژوهشگران برای شناسایی کارکردهای متفاوت سؤال، سؤال‌ها گرامر در بخش تخصصی دفترچه زبان انگلیسی آزمون سراســـری را با دو روش IRT و کلاسیک بررسی کردند. آنها به این نتیجه رسیدند که نظریه کلاسیک به دلیل وجود تغییرپذیری شـــاخص‌های آن برای شناســـایی کارکرد متفاوت سؤال‌ها مفید نیستند. در مقابل، روش‌های موجود در بافت نظریه سؤال-پاســـخ برای انجام این کار مفیدتر هستند (ایزانلو و حبیبی عسگرآباد، ۱۳۸۷). همچنین، عباســـی و همکاران (۱۳۹۲) با مقایســـه مدل‌های کلاسیک و خصیصه مکنون، در ارزیابی آزمون‌های تخصصی ورود به دوره‌های کارورزی رشـــته پزشکی نشان دادند که تحلیل‌های حاصل از مدل‌های خصیصه مکنون می‌توانند در جهت رفع محدودیت‌های نظریه کلاسیک آزمون مورد استفاده قرار گیرند.

نظریه سؤال-پاسخ، علاوه بر ارزشیابی آزمون، روش و ابزار مناسبی نیز برای پیش‌بینی توانایی و موفقیت افراد در آزمون‌های آتی اســـت. رودنر[2] (۲۰۰۹)، مطالعه‌ای جامع روی آزمون‌های GMAT[3] برگزارشـــده در ایالات متحده آمریکا، طی ســـال‌های ۲۰۰۴-۱۹۹۶ با استفاده از IRT انجام داد و میزان قابلیت پیش‌بینی آزمون در مورد موفقیت داوطلبان در طول دوره آموزشی مدیریت MBA[4] و همچنین مقایسه مدل انطباقی کامپیوتری و فرم مدادکاغذی آزمون GMAT را بررسی کرد و مشخص شد که به‌کارگیری نظریه سؤال-پاسخ در آزمون، پیش‌بینی کننده خوبی برای موفقیت افراد در ۵ سال دوره آموزشی مدیریت خواهد بود. روایی و اعتبار، توانایی آزمودنی‌ها، توابع آگاهی آزمون، احتمال سوگیری سؤال‌ها، مشخصات بانک سؤال‌ها، خطای استاندارد سؤال‌ها و دیگر ویژگی‌های روان‌سنجی نیز در این پژوهش، تحلیل و بررسی شدند.

با توجه به اهمیت آزمون کنکور سراسری در ارزشیابی و گزینش داوطلبان ورود به دانشگاه در ایران و همچنین

1. Schurmeier
2. Rudner
3. Graduate Management Admission Test
4. Master of Business Administration

توانایی روش سؤال-پاسخ، در این پژوهش، سؤال‌های شیمی در آزمون کنکور سراسری سال۱۳۹۶ رشته علوم تجربی با روش IRT بررسی و تحلیل می‌شود. سؤال‌های پژوهش عبارت‌اند از:

- مقدار ضریب دشواری، ضریب تشخیص و عامل حدس هر سؤال در نظریه IRT و کلاسیک چند است؟
- کدام مدل یک، دو یا سه پارامتری IRT برای ارزشیابی داده‌های حاصل از اجرای آزمون کنکور سراسری برازش بهتری دارند؟
- هر یک از سؤال‌های آزمون و همچنین کل سؤال‌های شیمی کنکور سراسری ۱۳۹۶ با کدام مدل IRT سازگارتر هستند؟
- کدام‌یک از سؤال‌های شیمی آزمون سراسری ۱۳۹۶ برای بانک سؤال مناسب‌تر هستند؟
- میزان آگاهی هر سؤال و آگاهی کل آزمون چقدر است؟

روش پژوهش

جامعه و نمونه آماری: پژوهش حاضر از نظر هدف کاربردی و با توجه به نوع و ماهیت موضوع توصیفی-پیمایشی است. جامعه آماری پژوهش، پاسخنامه همه داوطلبان شرکت‌کننده در آزمون سراسری ورود به دانشگاه سال ۱۳۹۶ در رشته علوم تجربی بود. با توجه به اینکه نظریه‌های روان‌سنجی، به‌خصوص سؤال-پاسخ، نظریه‌های نمونه بزرگ هستند، برای برآورد باثبات پارامترها، نمونه‌ای با حجم بالا لازم است. معمولاً نمونه‌های بیشتر از ۱۰۰۰ نمونه‌های بزرگ محسوب می‌شوند (یونسی، ۱۳۸۶). بنابراین برای تحلیل ۳۵ سؤال بخش شیمی آزمون سال ۱۳۹۶، از بین همه داوطلبان شرکت‌کننده در کنکور تجربی، ۵۰۰۰ پاسخنامه به‌صورت تصادفی به‌عنوان نمونه برای انجام تحلیل انتخاب شدند.

ابزار گردآوری داده‌ها: ابزار این پژوهش ۳۵ سؤال شیمی (دفترچه کد A-۲۲۰ از سؤال ۲۳۶ تا سؤال ۲۷۰) در آزمون رشته علوم تجربی سال ۱۳۹۶ است که سازمان سنجش آموزش کشور به‌صورت سراسری برگزار کرده است. این سؤال‌ها در ۶ کد مختلف بین آزمون‌شوندگان توزیع می‌شود. این سؤال‌ها چهارگزینه‌ای بوده و هدف این آزمون، گزینش افراد برای رشته‌های مختلف دانشگاهی بر اساس سوابق تحصیلی و نمره‌های افراد در هر یک از بخش‌های مختلف این آزمون است. با توجه به اینکه سؤال‌ها توسط متخصصان و طراحان مجرب کشور طراحی و ارزیابی می‌شود بنابراین روایی محتوایی آن قبلاً توسط این متخصصان سنجش شده است. پایایی این آزمون برای ۳۵ سؤال بررسی شده و مقدار آلفای کلی برابر ۰/۹۳۴ است که بیانگر پایایی خوب سؤال‌هاست.

فرایند اجرای پژوهش: پاسخ‌های ۵۰۰۰ آزمودنی به سؤال‌های کنکور سراسری سال ۱۳۹۶ در بخش شیمی رشته علوم تجربی، از سازمان سنجش و آموزش کشور در قالب دو فایل Excel و SPSS تحویل گرفته شد که در فایل Excel پاسخ هر آزمودنی با عبارت‌های «صحیح» و «غلط» ثبت شده بود. سپس پاسخ‌های «صحیح» هر فرد به عدد یک و پاسخ‌های «غلط یا بدون پاسخ» به عدد صفر تبدیل شدند و نمره کل هر داوطلب محاسبه

شد. نمره‌های آزمون‌شوندگان برحسب نمره کل آنها (صفر تا ۳۵) طبقه‌بندی‌شده است که به نمره صفر عدد ۳- و به نمره ۳۵ عدد ۳+ اختصاص داده شـــده است. ســـپس جدول این اطلاعات به‌عنوان فایل ورودی برای نرم‌افزار SPSSو IRTPRO آماده شد و مورد استفاده و تحلیل قرار گرفت.

روش تجزیه‌وتحلیل داده‌ها: ابتدا مفروض تک‌بعدی بودن با استفاده از نرم‌افزار NOHARM و شاخص‌های RMSR[1] و تاناکا بررســـی شـــد که به ترتیب، مقادیر ۰/۰۰۳۵ و ۰/۹۷۳ به دست آمد. مقدار مناسب برای این شـــاخص‌ها به ترتیب، کوچک‌تر از ۰/۰۵ و بزرگ‌تر از ۰/۹۵ اســـت که با توجه به مناسب بودن این ضرایب و تک‌بعدی بودن ســـؤال‌ها، نیازی به بررســـی مفروضه استقلال موضعی نیست. سپس داده‌های هر سؤال با سه مدل یک، دو و ســـه پارامتری IRT توســـط نرم‌افزار IRTPRO تحلیل شد و شاخص‌هایی مانند درجه آزادی، خی دو، شـــاخص معنی‌داری sig و مقدار -Log-Likelihood۲ برای هر ســـه مدل نیز محاسبه شد. با مقایسه شـــاخص -Log-Likelihood۲ و استفاده از آزمون خی دو، بهترین مدل برای برازش داده‌های آزمون برای هر سؤال مشخص شد و پس از مشخص شدن بهترین مدل برای هر سؤال، منحنی ویژگی سؤال‌ها (ICC)، منحنی آگاهی سؤال‌ها و منحنی آگاهی آزمون و همچنین منحنی خطای استاندارد آزمون به وسیله نرم‌افزار IRTPRO ترســـیم شد. میزان آگاهی هر سؤال در مدل ســـه‌پارامتری (بهترین مدل از نظر برازش + با داده‌ها) با استفاده از نرم‌افزارIRTPRO محاسبه شد. تابع آگاهی برای سؤال nم به‌صورت Ii(θ) در مدل سه‌پارامتری به‌صورت زیر تعریف می‌شود:

$$I_{i(\theta)} = a_{i(\theta)}{}^2 \frac{(P_{i(\theta)} - c_i)^2}{(1 - c_i)^2} \frac{q_{i(\theta)}}{p_{i(\theta)}} \qquad q_{i(\theta)} = 1 - p_{i(\theta)}$$

در این فرمول:

a_i: پارامتر ضریب تشخیص برای سؤال i

$p_{(\theta i)}$: احتمال پاسخ درست به سؤال توسط آزمودنی با توانایی θ

$q_{(\theta i)}$: احتمال پاسخ غلط به سؤال توسط آزمودنی با توانایی θ

θ: سطح توانایی مورد نظر

c_i: عامل حدس در سؤال iم است.

در صورت برقراری مفروضه استقلال موضعی می‌توان با جمع کردن توابع آگاهی سؤال‌ها، تابع آگاهی یک آزمون را از رابطه زیر به دست آورد.

$$I_{(\theta)} = \sum_{i=1}^{N} I_{i(\theta)}$$

همچنین شاخص‌های آماری درجه دشواری و ضریب تشخیص (همبستگی دو رشته‌ای نقطه‌ای) برای سؤال‌ها در نظریه کلاسیک نیز بررسی شد.

1. Root mean square of residual

جدول (۱) معیار رتبه‌بندی سؤال‌ها بر اساس مقدار پارامتر تشخیص سؤال در مدل منطقی

دامنه	۰	۰/۰۱ -۰ /۳۴	۰/۳۵ – ۰/۶۴	۰/۶۵ – ۱/۳۴	۱/۳۵ – ۱/۶۹	۱/< ۷۰	+ بی‌نهایت
رتبه	هیچ	خیلی ضعیف	ضعیف	متوسط	قوی	خیلی قوی	کامل

یافته‌ها

تمامی ســـؤال‌ها به‌صورت جداگانه با نرم‌افزار ارزشیابی شدند و در جدول (۲) شاخص‌های مدل یک، دو و سه پارامتری شامل دشواری، ضریب تشخیص ، عامل حدس، خی دو، درجه آزادی و سطح معنی‌داری و همچنین شاخص‌های مدل کلاسیک برای همه سؤال‌ها آمده است. جدول (۲) نشان می‌دهد که:

- ۷ سؤال، شامل سؤال‌های ۲۳۸، ۲۴۲، ۲۴۳، ۲۵۱، ۲۵۷، ۲۶۷ و ۲۶۸ با مدل یک‌پارامتری برازش دارند زیرا دارای سطح معنی‌داری بیشتر از ۰/۰۵ هستند.
- ۹ سؤال، شامل سؤال‌های ۲۳۸، ۲۴۲، ۲۴۳، ۲۴۴، ۲۵۱، ۲۵۷، ۲۶۳، ۲۶۷ و ۲۶۸ با مدل دوپارامتری برازش دارند زیرا دارای سطح معنی‌داری بیشتر از ۰/۰۵ هستند.
- ۲۱ سؤال شامل سؤال‌های ۲۳۸، ۲۳۹، ۲۴۱، ۲۴۲، ۲۴۳، ۲۴۴، ۲۴۵، ۲۴۸، ۲۵۱، ۲۵۳، ۲۵۶، ۲۵۷، ۲۵۹، ۲۶۰، ۲۶۱، ۲۶۲، ۲۶۳، ۲۶۴، ۲۶۸، ۲۶۹ و ۲۷۰ با مدل سه‌پارامتری برازش مناسبی دارند.
- در مدل ســـه‌پارامتری، آسان‌ترین سؤال ۲۴۸ و دشوارترین سؤال ۲۶۶ هستند که پارامتر b آنها به ترتیب کمترین (۰/۲۷-) و بیشترین (۸/۴۵) است.
- در مدل سه‌پارامتری، کمترین میزان ضریب تشخیص (a) مربوط به سؤال۲۳۹ با ضریب ۰/۳۲ و بیشترین میزان آن مربوط به سؤال ۲۵۶ با ضریب ۷/۰۱ است.
- در مدل سه‌پارامتری، ســـؤال‌های ۲۳۹، ۲۴۵ و۲۶۶ از نظر قدرت تشخیص، سؤال‌های مناسبی محسوب نمی‌شوند اما سایر سؤال‌ها ضریب تشخیص بیشتر از ۰/۶۵ دارند و مناسب هستند.
- در مدل سه پارامتری، کمترین میزان پارامتر حدس متعلق به سؤال ۲۵۲ با ضریب ۰/۰۹ و بیشترین میزان آن مربوط به سؤال ۲۵۳ با ضریب ۰/۲۹ است.
- ۱۳ سؤال، شامل سؤال‌های ۲۳۶، ۲۳۷، ۲۴۰، ۲۴۶، ۲۴۷، ۲۴۹، ۲۵۰، ۲۵۲، ۲۵۴، ۲۵۵، ۲۵۸، ۲۶۵ و ۲۶۶ با هیچ‌یک از مدل‌های ذکرشده برازش نداشته‌اند.
- ۷ سؤال، هم با مدل یک‌پارامتری و هم با دوپارامتری برازش دارند.
- ۶ سؤال، هم‌زمان با هر سه مدل برازش نشان داده‌اند.

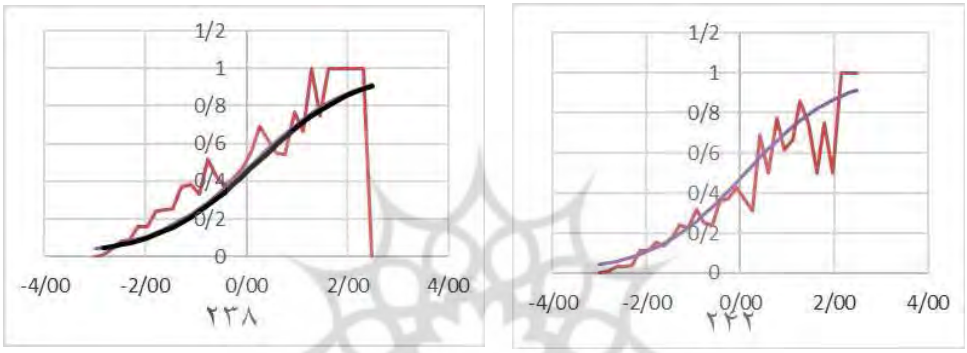
جدول (۲) ضرایب دشواری، تشخیص و حدس به‌دست‌آمده از تحلیل سؤال‌ها با نظریه کلاسیک و سؤال-پاسخ

	کلاسیک		IRT										
شماره سؤال	b	a	تک‌پارامتری			دوپارامتری			سه‌پارامتری				
			b	خی دو	معنی‌داری	b	a	خی دو	معنی‌داری	b	a	c	خی دو	معنی‌داری
۲۳۶	۰/۲۷۲	۰/۴۹۶	-۰/۱۶	۲۶/۱۷	۰/۰۰۱	-۰/۱۰	۱/۶۲	۲۶/۱۷	۰/۰۰۱	۰/۱۰	۱/۹۱	۰/۱۰	۱۶/۱۷	۰/۰۲۳
۲۳۷	۰/۱۱۱	۰/۴۷۵	-۰/۵۶	۳۸/۲۷	۰/۰۰۰	-۰/۵۲	۱/۸۷	۳۸/۲۷	۰/۰۰۰	-۰/۰۹	۳/۱۸	۰/۲۴	۲۵/۴۵	۰,۰۰۰
۲۳۸	۰/۱۶۰	۰/۵۳۱	۰/۱۳	۶/۶۴	۰/۴۶۹	۰/۱۳	۱/۰۲	۶/۶۴	۰/۴۶۹	۰/۵۹	۱/۵۸	۰/۱۹	۸/۴۸	۰/۲۹۴
۲۳۹	۰/۱۵۲	۰/۵۲۶	۰/۵۹	۳۸/۵۳	۰/۰۰۰	۳/۳۱	۰/۱۵	۳۸/۵۳	۰/۰۰۰	۴/۵۸	۰/۳۲	۰/۲۳	۱۳/۰۵	۰/۰۷۰
۲۴۰	۰/۲۷۹	۰/۵۴۰	-۰/۵۲	۳۷/۴۹	۰/۰۰۰	-۰/۳۳	۱/۹۸	۳۷/۴۹	۰/۰۰۰	-۰/۱۴	۲/۲۵	۰/۱۰	۱۵/۶۴	۰/۰۲۸
۲۴۱	۰/۱۹۸	۰/۵۰۷	۰/۳۰	۱۶/۷۳	۰/۰۱۹	۰/۳۵	۰/۸۳	۱۶/۷۳	۰/۰۱۹	۰/۹۷	۲/۲۳	۰/۲۶	۹/۰۳	۰/۲۵۰
۲۴۲	۰/۱۱۹	۰/۴۸۵	۰/۱۱	۸/۰۹	۰/۴۲۶	۰/۰۱	۱/۵۶	۸/۰۹	۰/۴۲۶	۰/۴۱	۳/۰۳	۰/۱۸	۱۱/۹۷	۰/۱۰۱
۲۴۳	۰/۱۹۴	۰/۵۸۷	۰/۲۸	۵/۷۶	۰/۵۶۹	۰/۱۹	۱/۶۳	۵/۷۶	۰/۵۶۹	۰/۴۱	۲/۲۰	۰/۱۰	۴/۲۶	۰/۶۴۳
۲۴۴	۰/۰۶۶	۰/۴۸۳	۰/۷۲	۲۰/۴۹	۲۰/۴۹	۴/۴۲	۰/۲۱	۲۰/۴۹	۰/۰۰۵	۲/۷۰	۱/۱۵	۰/۲۳	۸/۱۸	۰/۳۱۸
۲۴۵	۰/۱۳۹	۰/۵۲۰	۰/۵۱	۴۳/۰۹	۴۳/۰۹	۱/۸۱	۰/۲۴	۴۳/۰۹	۰/۰۰۰	۲/۹۸	۰/۳۶	۰/۱۸	۹/۸۱	۰/۱۹۹
۲۴۶	۰/۲۲۸	۰/۶۵۱	-۰/۳۹	۱۶/۸۲	۱۶/۸۲	-۰/۲۹	۱/۶۶	۱۶/۸۲	۰/۰۱۸	۰/۲۲	۲/۹۹	۰/۲۳	۱۴/۳۳	۰/۰۴۵
۲۴۷	۰/۱۶۸	۰/۶۰۷	-۰/۳۲	۲۷/۱۲	۲۷/۱۲	-۰/۲۴	۱/۶۹	۲۷/۱۲	۰/۰۰۰	-۰/۰۲	۱/۹۵	۰/۱۱	۱۸/۵۲	۰/۰۰۹
۲۴۸	۰/۲۸۵	۰/۶۲۱	-۰/۷۴	۲۱/۲۳	۲۱/۲۳	-۰/۴۹	۱/۸۱	۲۱/۲۳	۰/۰۰۶	-۰/۲۷	۲/۰۳	۰/۱۳	۱۳/۶۸	۰/۰۵۸
۲۴۹	۰/۲۵۱	۰/۵۸۷	-۱/۰۵	۴۰/۵۷	۴۰/۵۷	-۰/۶۵	۲/۳۹	۴۰/۵۷	۰/۰۰۰	-۰/۵۱	۲/۴۹	۰/۱۰	۱۷/۹۲	۰/۰۱۲
۲۵۰	۰/۰۷۶	۰/۵۳۷	-۰/۰۰	۲۴/۷۱	۲۴/۷۱	-۰/۰۹	۱/۲۸	۲۴/۷۱	۰/۰۰۱	۰/۴۷	۲/۶۴	۰/۲۳	۱۵/۴۴	۰/۰۳۰
۲۵۱	۰/۲۵۵	۰/۵۷۶	-۰/۰۵	۱۳/۸۲	۱۳/۸۲	-۰/۰۵	۱/۲۲	۱۳/۸۲	۰/۰۸۶	۰/۴۵	۲/۲۴	۰/۲۲	۱۱/۲۲	۰/۱۲۹

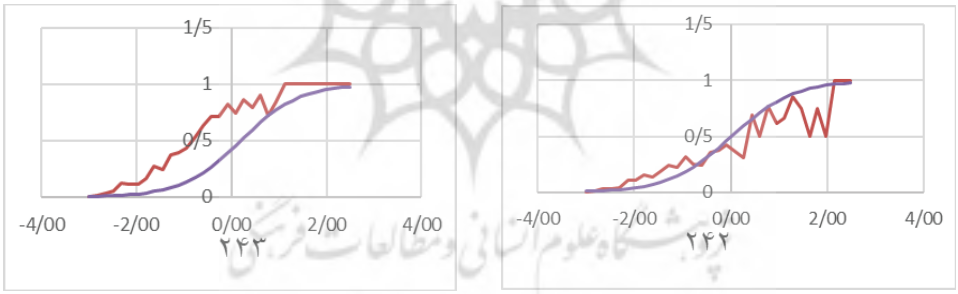
ادامه جدول (۲)

شماره سؤال	کلاسیک		IRT									
	b	a	تک‌پارامتری			دوپارامتری				سه‌پارامتری				
			b	خی دو	معنی‌داری	b	a	خی دو	معنی‌داری	b	a	c	خی دو	معنی‌داری
۲۵۲	۰/۲۸۵	۰/۴۶۶	-۰/۸۳	۷۷/۷۶	۷۷/۷۶	-۰/۵۴	۲/۰۹	۷۷/۷۶	۰/۰۰۰	-۰/۴۰	۲/۲۳	۰/۰۹	۴۲/۶۰	۰/۰۰۰
۲۵۳	۰/۰۷۹	۰/۵۴۵	۰/۳۲	۲۶/۷۹	۲۶/۷۹	۰/۸۹	۰/۵۱	۲۶/۷۹	۰/۰۰۰	۱/۳۰	۲/۹۰	۰/۲۹	۹/۸۳	۰/۲۸۰
۲۵۴	۰/۱۰۸	۰/۵۵۳	۰/۳۲	۲۵/۳۳	۲۵/۳۳	۰/۶۱	۰/۵۹	۲۵/۳۳	۰/۰۰۰	۱/۲۹	۱/۲۸	۰/۲۵	۱۵/۳۴	۰/۰۳۱
۲۵۵	۰/۰۴۷	۰/۵۳۱	۰/۶۰	۳۱/۰۲	۳۱/۰۲	۰/۹۱	۰/۷۷	۳۱/۰۲	۰/۰۰۰	۱/۲۳	۱/۳۸	۰/۱۶	۱۸/۱۵	۰/۰۰۵
۲۵۶	۰/۱۶۵	۰/۶۲۱	-۰/۳۴	۱۸/۹۳	۱۸/۹۳	-۰/۳۶	۱/۹۳	۱۸/۹۳	۰/۰۰۸	۰/۲۳	۷/۰۱	۰/۲۶	۱۰/۱۴	۰/۱۸۰
۲۵۷	۰/۰۳۶	۰/۵۵۳	۰/۶۲	۱۱/۷۴	۱۱/۷۴	۱/۳۵	۰/۶۱	۱۱/۷۴	۰/۱۰۹	۱/۶۷	۰/۹۷	۰/۱۴	۱۰/۵۴	۰/۱۵۹
۲۵۸	۰/۱۱۲	۰/۵۹۳	-۰/۱۳	۲۵/۱۴	۲۵/۱۴	-۰/۲۰	۱/۸۲	۲۵/۱۴	۰/۰۰۱	۰/۲۹	۳/۸۰	۰/۲۰	۱۷/۶۳	۰/۰۱۳
۲۵۹	۰/۱۷۶	۰/۵۳۵	۰/۴۴	۱۹/۸۲	۱۹/۸۲	۰/۴۴	۱/۰۲	۱۹/۸۲	۰/۰۰۶	۰/۸۱	۱/۶۹	۰/۱۶	۱۳/۵۱	۰/۰۶۰
۲۶۰	۰/۱۷۵	۰/۵۴۲	۰/۰۴	۱۹/۷۶	۱۹/۷۶	۰/۰۴	۰/۹۵	۱۹/۷۶	۰/۰۰۶	۰/۵۶	۱/۴۳	۰/۲۰	۹/۵۰	۰/۱۴۷
۲۶۱	۰/۱۴۸	۰/۴۱۶	۰/۹۲	۴۵/۵۶	۴۵/۵۶	۲/۳۳	۰/۳۵	۴۵/۵۶	۰/۰۰۰	۲/۰۹	۱/۳۵	۰/۲۳	۱۲/۸۵	۰/۰۷۵
۲۶۲	۰/۱۹۰	۰/۵۵۹	۰/۲۳	۲۷/۸۲	۲۷/۸۲	۰/۳۱	۰/۶۳	۲۷/۸۲	۰/۰۰۰	۱/۰۲	۰/۹۶	۰/۲۱	۹/۷۸	۰/۱۳۳
۲۶۳	۰/۱۸۰	۰/۵۳۲	-۰/۲۷	۱۵/۸۹	۱۵/۸۹	-۰/۲۷	۱/۵۰	۱۵/۸۹	۰/۰۴۴	۰/۰۹	۲/۰۲	۰/۱۷	۱۱/۳۴	۰/۱۲۴
۲۶۴	۰/۰۶۲	۰/۵۵۱	۱/۰۳	۲۹/۸۸	۲۹/۸۸	۲/۱۸	۰/۴۴	۲۹/۸۸	۰/۰۰۰	۲/۴۰	۰/۷۳	۰/۱۴	۹/۰۸	۰/۲۴۶
۲۶۵	۰/۱۰۵	۰/۵۸۲	-۰/۱۲	۲۳/۴۹	۲۳/۴۹	-۰/۲۰	۱/۸۵	۲۳/۴۹	۰/۰۰۳	۰/۲۰	۲/۹۶	۰/۱۸	۱۶/۷۳	۰/۰۱۹
۲۶۶	۰/۰۳۸	۰/۵۴۰	۰/۹۲	۳۴/۱۳	۳۴/۱۳	۱۶/۱۱	۰/۰۷	۳۴/۱۳	۰/۰۰۰	۸/۴۵	۰/۳۷	۰/۲۱	۲۱/۱۹	۰/۰۰۳
۲۶۷	۰/۱۹۸	۰/۵۰۶	۰/۱۰	۱۳/۸۳	۱۳/۸۳	۰/۰۹	۱/۱۰	۱۳/۸۳	۰/۰۵۴	۰/۴۱	۱/۴۵	۰/۱۴	۱۵/۵۱	۰/۰۲۹
۲۶۸	۰/۰۸	۰/۴۴۲	۰/۴۸	۱۱/۳۶	۱۱/۳۶	۰/۴۸	۱/۰۲	۱۱/۳۶	۰/۱۲۳	۰/۷۶	۱/۴۲	۰/۱۲	۹/۸۹	۰/۱۹۴
۲۶۹	۰/۰۷۲	۰/۵۴۱	۰/۳۸	۱۵/۵۷	۱۵/۵۷	۰/۶۸	۰/۶۵	۱۵/۵۷	۰/۰۲۳	۱/۲۵	۱/۲۷	۰/۲۱	۹/۱۲	۰/۲۴۴
۲۷۰	۰/۱۴۵	۰/۵۲۹	۰/۱۴	۱۴/۷۷	۱۴/۷۷	۰/۰۳	۱/۵۸	۱۴/۷۷	۰/۰۳۹	۰/۳۲	۲/۳۶	۰/۱۳	۱۱/۴۴	۰/۱۲۰

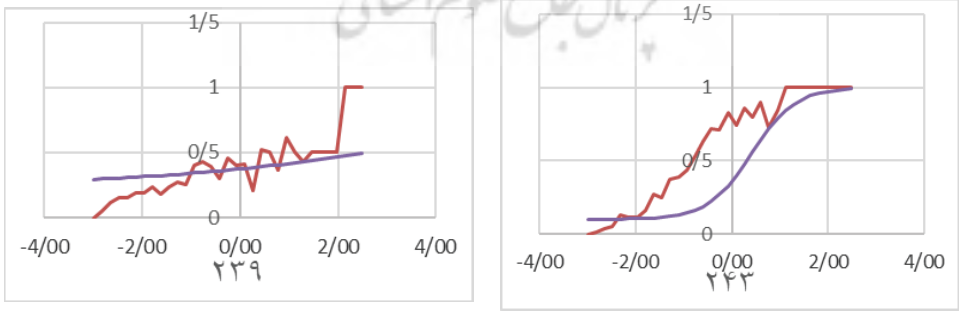
در زیر منحنی ویژه ســـؤال برای تعدادی از ســـؤال‌ها در مدل‌های مختلف رسم شده است. شکل‌های (۱ و ۲) به ترتیب، بهترین برازش با مدل تک‌پارامتری و دوپارامتری را نشـــان می‌دهند. در شـــکل (۳) سؤال‌ها با مدل سه‌پارامتری برازش شده‌اند که در آنها آسان‌ترین، سخت‌ترین، سؤال‌های با بیشترین و کمترین ضریب تشخیص و عامل حدس نشان داده شده‌اند. در این نمودارها محور افقی توانایی (θ) و محور عمودی p(θ) و p(e)را نشان می‌دهد. نمودارهای p(θ) به‌صورت خط پیوســـته قرمزرنگ و دندانه‌دار و برازش با مدل به‌صورت خط پیوسته بنفش‌رنگ نشان داده شده است.

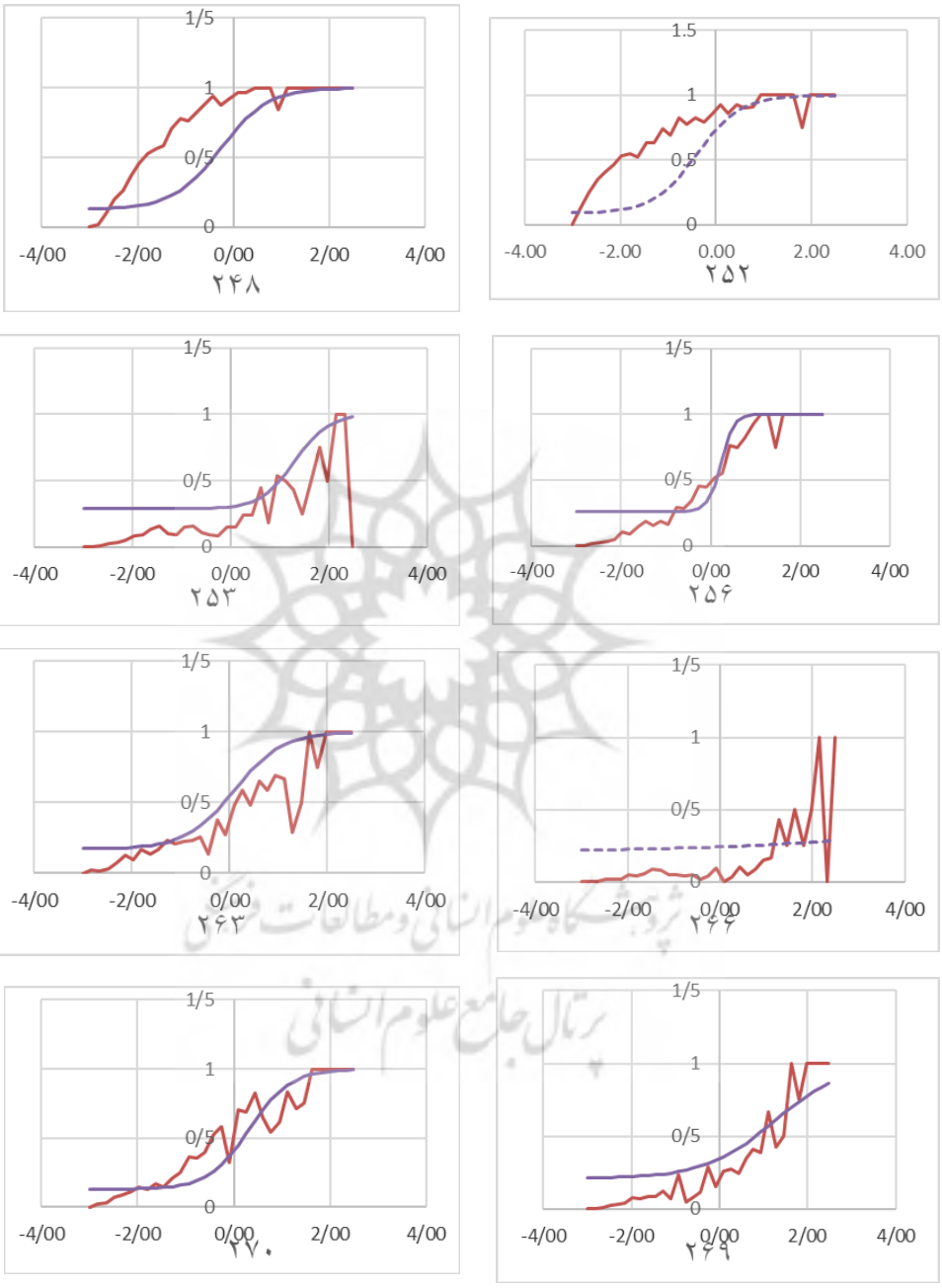


شکل (۱) منحنی ویژگی تعدادی از سؤال‌ها با مدل یک‌پارامتری



شکل (۲) منحنی ویژگی تعدادی از سؤال‌ها با مدل دوپارامتری





شکل (۳) منحنی ویژگی تعدادی از سؤال‌ها با مدل سه‌پارامتری

سؤال ۲۶۶ با b= ۸/۴۵ دشواری بسیار زیادی دارد و نمودار آن تقریباً افقی و شیب کمی (a=۰/۳۷) دارد و با توجه به جدول (۲) می‌توان گفت که قدرت تشخیص سؤال ضعیف است. از طرفی، مجانب پایینی نمودار بزرگ است و عامل حدس در این سؤال ۰/۲۱ است یعنی شانس افراد با توانایی کم نیز برای پاسخ دادن حدسی به این سؤال بالاست. سطح معنی‌داری این سؤال ۰/۰۰۳ است که بسیار کمتر از ۰/۰۵ است و نشانه عدم برازش این سؤال با مدل سه‌پارامتری است. این سؤال با هیچ مدلی برازش نداشته و سؤال نامناسبی است و بهتر است در سال‌های بعد از طرح سؤال‌های مشابه با این سؤال صرف‌نظر شود.

در سؤال ۲۵۶ دشواری متوسط و b=۰/۲۳ است اما شیب بسیار زیادی دارد (a=۷/۰۱) بنابراین، شکل منحنی آن تقریباً s شکل است و مطابق جدول (۲) می‌توان گفت قدرت تشخیص این سؤال خیلی خوب است. مجانب پایینی نمودار بزرگ و عامل حدس در این سؤال ۰/۲۶ است یعنی شانس افراد با توانایی کم نیز برای پاسخ دادن حدسی به این سؤال بالاست. سطح معنی‌داری این سؤال ۰/۱۸ است که بیشتر از ۰/۰۵ است و نشانه برازش خوب این سؤال با مدل سه‌پارامتری است؛ بر این اساس، به‌طورکلی می‌توان گفت این سؤال مناسب است.

بررسی برازش مدل‌های مختلف با استفاده از RMSEA

در جدول (۳)، مقادیر RMSEA در مدل‌های مختلف نشان داده شده است. شاخص RMSEA در هر مدل، هرچه به صفر نزدیک‌تر باشد سؤال‌های با آن مدل برازش بهتری دارند. مقدار این شاخص در مدل سه پارامتری۰/۰۶ و از همه کمتر است بنابراین، سؤال‌های آزمون شیمی سال ۱۳۹۶ با مدل سه‌پارامتری برازش بهتری دارند. نمودارهای منحنی ویژگی سؤال در شکل (۳) نیز نشان‌دهنده این مورد است.

جدول (۳) مقادیر RMSEA در مدل‌های مختلف برای بررسی برازش کل آزمون با هر مدل

مدل	بیشینه درست‌نمایی	X^2	درجه آزادی	احتمال(p)	RMSEA
۱ پارامتری	۵۲۷۵۱/۰۲	۱۸۷۲۹/۶۹	۵۹۴	۰/۰۰۰۱	۰/۰۸
۲ پارامتری	۵۱۲۰۹/۶۵	۱۴۰۴۹/۳۳	۵۶۰	۰/۰۰۰۱	۰/۰۷
۳ پارامتری	۵۰۸۹۳/۰۷	۱۰۸۷۸/۵۰	۵۲۵	۰/۰۰۰۱	۰/۰۶

مقایسه برازش داده‌ها با مدل‌های مختلف

برای مقایسه مدل‌های مختلف IRT و تعیین بهترین مدل از میان مدل‌های یک و دو و سه پارامتری از آزمون خی دو استفاده شده است. در این آزمون، فرضیه به این قرار است که مدل دارای پارامترهای بیشتر نسبت به مدل دارای پارامترهای کمتر، برازش بهتری با داده‌ها دارد. این آزمون به‌صورت زیر استفاده می‌شود.

((مدل ب) ۲Loglikelihood-) - ((مدل الف) ۲Loglikelihood-) = خی دو

جدول (٤) مقایسه برازش داده‌ها با مدل‌های مختلف با استفاده از آزمون خی دو

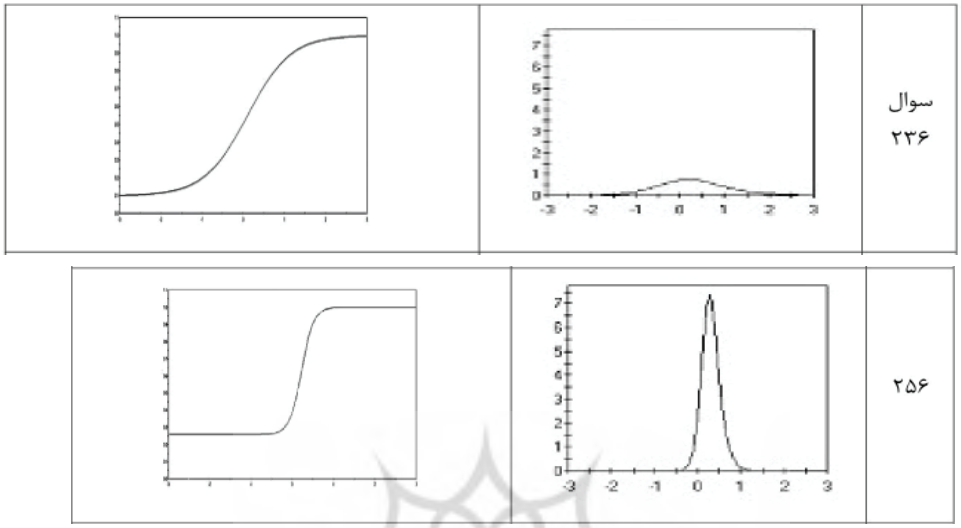
	-Loglikelihood۲ مربوط به هر مدل			خی دو
	یک پارامتری	دو پارامتری	سه پارامتری	
مقایسه یک با دو پارامتری	۵۲۷۵۱/۰۲	۵۱۲۰۹/۶۵		۱۵۴۱/۳۷
مقایسه دو با سه پارامتری		۵۱۲۰۹/۶۵	۵۰۸۹۳/۰۷	۳۱۶/۵۸

مقدار مجاز خی دو در جدول‌های استاندارد برای درجه آزادی ۳۵ و سطح اطمینان ۹۹ در صد ۵۷/۳۴۲ است. مقدار خی دو به‌دست‌آمده در جدول (۴) برای مقایسه مدل یک و دو پارامتری ۱۵۴۱/۳۷ و بزرگ‌تر از مقدار مجاز است؛ به عبارتی برازش داده‌ها با مدل دوپارامتری با برازش آنها با مدل یک‌پارامتری متفاوت است. همچنین، چون مقدار بیشینه درست‌نمایی در مدل دوپارامتری کمتر است پس مدل دوپارامتری نسبت به مدل تک‌پارامتری، برازش بهتری با این مجموعه از داده‌ها دارد. همچنین با استدلال مشابه، می‌توان گفت که مدل سه‌پارامتری از مدل دوپارامتری برای برازش سؤال‌ها بهتر است.

نمودار آگاهی و منحنی ویژه سؤال‌ها در مدل سه‌پارامتری

برای تخمین نقش هر سؤال در میزان آگاهی‌دهندگی آن سؤال روی پیوستار توانایی، از تابع آگاهی سؤال استفاده می‌شود که در شکل (۴) بهترین و ضعیف‌ترین سؤال از نظر آگاهی در منحنی ویژگی این دو سؤال آمده است. تابع آگاهی سؤال در سمت راست و منحنی ویژگی آن در سمت چپ آمده است. در نمودار آگاهی، محور افقی، توانایی و محور عمودی، میزان آگاهی سؤال است و در منحنی ویژگی سؤال، محور افقی، توانایی و محور عمودی، احتمال پاسخ درست به سؤال $\theta(P)$ است.

نمودار آگاهی هر سؤال، میزان آگاهی آن را روی پیوستار توانایی مشخص می‌سازد. همان‌طور که دیده می‌شود هرچه شیب منحنی یا ضریب تشخیص یک سؤال بیشتر باشد، آگاهی آن سؤال نیز افزایش می‌یابد؛ به این معنی که این سؤال‌ها قادر به تفکیک افراد دارای سطوح مختلف توانایی هستند پس سؤال‌های مناسبی نیز خواهند بود.



شکل (٤) نمودار تابع آگاهی و منحنی ویژه سؤال برای دو سؤال در مدل سه‌پارامتری

سؤال‌های ۲۳۹، ۲۴۵ و ۲۶۶ شیب کمتر از ۰/۶۵ دارند و نمودار آگاهی آنها مماس بر محور توانایی است. سؤال‌های ۲۴۴، ۲۵۴، ۲۵۵، ۲۵۷، ۲۶۱، ۲۶۲، ۲۶۴ و۲۶۹ دارای شیب ۱/۳۵ > a > ۰/۶۵ هستند و نمودار آگاهی آنها به‌سختی قابل تشخیص است. سؤال‌های ۲۳۸، ۲۵۹، ۲۶۰، ۲۶۷ و ۲۶۸ دارای شیب ۱/۷ >a> ۱/۳۵ هستند و نمودار آگاهی آنها پهن و بدون قله نوک‌تیز است و دامنه آگاهی آنها نسبتاً زیاد است. هرچه سؤال سخت‌تر باشد و مقدار پارامتر b در آن بیشتر باشد، نمودار آگاهی به سمت راست کشیده می‌شود و هرچه مقدار b کمتر باشد نمودار به سمت چپ کشیده می‌شود. سؤال‌های ۲۳۶، ۲۳۷، ۲۴۰، ۲۴۱، ۲۴۲، ۲۴۳، ۲۴۶، ۲۴۷، ۲۴۸، ۲۴۹، ۲۵۰، ۲۵۱، ۲۵۲، ۲۵۳، ۲۵۶، ۲۵۸، ۲۶۳، ۲۶۵ و ۲۷۰ دارای شیب بزرگ‌تر از ۱/۷ هستند و نمودار آگاهی آنها دارای دامنه کم و مقدار آگاهی زیاد و قله نوک‌تیزتر و مرتفع‌تر است که از میان آنها سؤال ۲۵۶ دارای بیشترین شیب و بیشترین آگاهی است. هر چقدر ضریب تشخیص (شیب) یک سؤال بزرگ‌تر باشد تابع آگاهی آن به‌صورت یک قله مرتفع و باریک‌تر ظاهر خواهد شد و برعکس، سؤال‌های با ضریب تشخیص کمتر، به شکل یک تپه پهن و کم‌ارتفاع درخواهد آمد؛ به عبارتی تابع آگاهی سؤال با ضریب تشخیص آن رابطه مستقیم دارد.

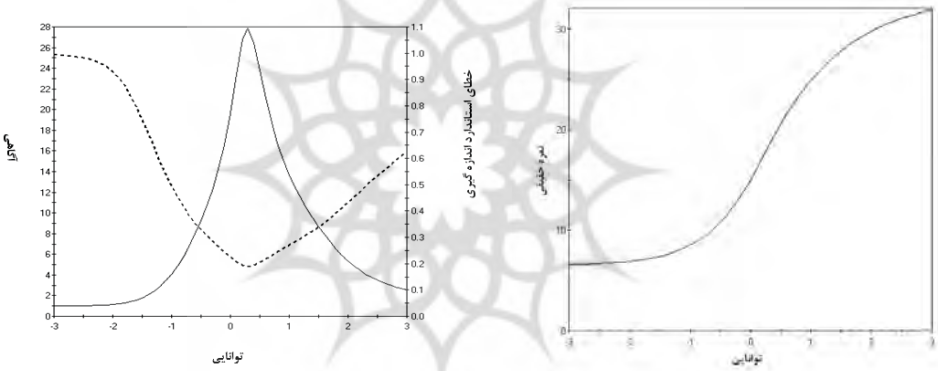
منحنی ویژه کل آزمون در مدل سه‌پارامتری

ازآنجاکه مدل سه‌پارامتری با سؤال‌های بیشتری برازش دارد و شاخص RMSEA کوچک‌تری در مقایسه با سایر مدل‌ها دارد منحنی ویژگی کل آزمون یعنی تمامی ۳۵ سؤال با استفاده از مدل سه پارامتری تحلیل شد که شکل آن در شکل (۵) رسم شده است.

این نمودار به شکل s با یک شیب تند و ضریب دشواری ۰/۳ است؛ بنابراین می‌توان گفت که دشواری آزمون شـــیمی نسبتاً متعادل بوده است زیرا ضریب دشـــواری در یک آزمون با دشواری متوسط، در نقطه صفر قرار می‌گیرد. همچنین شیب تند نمودار نشان می‌دهد مجموعه سؤال‌های آزمون شیمی کنکور سراسری سال ۱۳۹۶ رشته تجربی قدرت تشخیص بسیار خوبی دارد و با توجه به شیب منحنی بیشترین قدرت تشخیص بین سطوح توانایی ۰/۵- تا ۲ است. در ادامه مقاله، تأثیر این ناحیه بر تابع آگاهی آزمون مشخص خواهد شد. همچنین از شکل (۵) می‌توان دریافت که عامل حدس آزمون نسبتاً زیاد است که از معایب این آزمون است.

تابع آگاهی کل آزمون در مدل سه‌پارامتری

تابع آگاهی، اطلاعات مفیدی را در مورد ارزشـــیابی آزمون‌شوندگان در اختیار قرار می‌دهد. در شکل (۶) تابع آگاهی کل آزمون (منحنی پیوسته) و خطای استاندارد اندازه‌گیری (منحنی نقطه‌چین) رسم شده است.



شکل (۵) منحنی ویژه کل آزمون با مدل سه‌پارامتری

شکل (٦) تابع آگاهی کل آزمون

نمودار (۶) در مدل سه‌پارامتری نشان می‌دهد که این آزمون برای داوطلبان کنکوری که توانایی آنها در دامنه ۱- تا ۲/۵ قرار دارد مناسب است زیرا همان‌طور که شکل (۵) نشان می‌دهد رشد و افزایش منحنی ویژه آزمون در این ناحیه است؛ یعنی در این ناحیه از توانایی با تغییر توانایی آزمون‌شوندگان، میزان پاسخ صحیح آنها نیز تغییر می‌کند. به عبارتی، با افزایش قدرت تشخیص آزمون، تابع آگاهی آن نیز افزایش می‌یابد. بررسی آمار توصیفی و فراوانی داوطلبان نیز نشان داد که حدود ۸/۵ درصد از داوطلبان کنکور سراسری این سال در این ناحیه قرار می‌گیرند و درصد پاســـخ درست آنان به سؤال‌های شـــیمی در محدوده ۹۱/۵-۳۲ درصد بوده است. از طرفی، حدود ۹۱ در صد از داوطلبان نیز در ناحیه با توانایی کمتر از ۱- قرار دارند که نمره آنان در سؤال‌های شیمی کمتر از ۳۲ درصد اســـت و آزمون میزان آگاهی بســـیار کمی در مورد آنها در اختیار ما قرار می‌دهد. همچنین

از نمودار پیداست که آزمون بیشترین میزان آگاهی و اطلاعات را در توانایی ۰/۳ فراهم می‌آورد. بیشینه یا قله نمودار آگاهی در نقطه عطف منحنی ویژه آزمون قرار می‌گیرد که تغییرات شیب منحنی در این ناحیه بیشترین مقدار است و در نمودار (۵) نیز نشان داده است.

بحث

یافته‌های پژوهش نشان می‌دهد سؤال ۲۴۸ آسان‌ترین سؤال است زیرا در مورد محاسبات ساده استوکیومتری و واکنشگر محدودکننده بدون نیاز به ماشین‌حساب است. اما سؤال ۲۶۶ دشوارترین سؤال در مورد محاسبه pH است که علاوه بر دانش شیمی به مهارت و دانش ریاضی برای محاسبه لگاریتم اعداد بدون استفاده از ماشین‌حساب نیاز دارد و همین دشواری بسیار زیاد سؤال سبب کاهش شدید ضریب تشخیص این سؤال شده است؛ بنابراین طراحی چنین سؤال‌هایی در سال‌های آینده پیشنهاد نمی‌شود یا اینکه دبیران شیمی نحوه محاسبه لگاریتم بدون استفاده از ماشین‌حساب را به دانش‌آموزان آموزش دهند. سؤال ۲۵۶ در مورد تهیه محلول‌های رقیق از محلول‌های غلیظ است که باید از فرمول ریاضی متفاوت و پیچیده‌تر و لحاظ کردن درصد محلول در مقیاس ۱۰۰ استفاده شود. بنابراین حل این سؤال اندکی سخت‌تر است اما آنچنان نیست که ضریب سختی سؤال را بسیار زیاد کند، پس ضریب تشخیص سؤال در محدوده بسیار مناسبی قرار می‌گیرد و بهترین سؤال این آزمون است. پیشنهاد می‌شود از سؤال‌های محاسباتی دارای چندین فرمول برای حالات مختلف و با عملیات ریاضی نه‌چندان سخت در سال‌های آینده بیشتر استفاده شود. سؤال‌های ۲۴۵ و ۲۳۹ سؤال‌های ترکیبی از موضوعات مختلف هستند و برای پاسخ درست به سؤال باید هم‌زمان به چندین مبحث شیمی تسلط داشت. مثلاً سؤال ۲۴۵ از مباحث گروه‌های عاملی، قلمرو الکترونی و فرمول‌های ساختاری طراحی شده است و در طرح سؤال ۲۳۹ از مباحث آرایش الکترونی، نیروهای بین مولکولی، خاصیت تناوبی انرژی یونش و ارتباط آن با آرایش الکترونی استفاده شده است. این سؤال‌ها به‌حدی دشوار و سخت هستند که تعداد افراد بسیار بسیار کمی توانایی پاسخ به آن را دارند؛ بنابراین به دلیل ضریب دشواری بسیار زیاد، ضریب تشخیص آنها خیلی کوچک است و سؤال‌های مناسبی به شمار نمی‌آیند و باید از طرح سؤال بسیار سخت و پیچیده پرهیز شود. درعوض از سؤال‌های ترکیبی متشکل از دو موضوع استفاده شود. علی‌رغم اینکه ۱۳ سؤال با هیچ‌یک از مدل‌های IRT برازش نداشتند اما بررسی شاخص RMSEA نشان داد که مدل سه‌پارامتری با کل آزمون و سؤال‌های بیشتری برازش دارد و تحلیل منحنی ویژه کل آزمون با استفاده از مدل سه‌پارامتری، ضریب دشواری ۰/۳ را نشان می‌دهد و بیانگر دشواری نسبتاً متعادل آزمون شیمی است. همچنین، شیب منحنی ویژه کل آزمون قدرت تشخیص بسیار خوبی را برای آزمون شیمی نشان می‌دهد که بیشترین قدرت تشخیص، برای سطوح توانایی ۰/۵- تا ۲ است. این منحنی بیانگر عامل حدس نسبتاً زیاد آزمون است که از معایب این آزمون است. بررسی تابع آگاهی کل آزمون نشان می‌دهد این آزمون برای داوطلبان کنکوری که توانایی آنها در دامنه ۱- تا ۲/۵ قرار

دارد مناســـب اســـت که حدود ۸/۵ درصد از داوطلبان کنکور سراسری این سال در این ناحیه قرار می‌گیرند و درصد پاسخ درست آنان به سؤال‌های شیمی در محدوده ۹۱/۵-۳۲ در صد بوده است. بیشترین میزان آگاهی و اطلاعات این آزمون برای داوطلبان با توانایی ۰/۳ است.

نتیجه‌گیری

آزمون شـــیمی کنکور سراسری ۱۳۹۶ با مدل سه‌پارامتری برازش بهتری نشان می‌دهد. از مزایای آن می‌توان گفت که بیشترین آگاهی این آزمون برای داوطلبان دهک بالا یعنی پاسخ‌های صحیح در محدوده ۹۱/۵-۳۲ درصد است و ضریب دشواری و ضریب تشخیص نسبتاً متعادلی دارد اما پارامتر حدس آن نسبتاً زیاد است که از معایب این آزمون است. با توجه به یافته‌های این پژوهش پیشنهاد می‌شود سؤال‌های دارای ضرایب دشواری و تشـــخیص و عامل حدس مناسب در بانک سؤال‌های ســـازمان سنجش کشور قرا گیرند و با تحلیل سالیانه کنکور سراسری، بانک غنی از سؤال‌های مناسب شیمی تهیه شود و در آزمون‌های سال‌های آینده و آزمون‌های آزمایشـــی و استخدامی و المپیادها و ... از این بانک استفاده شود. همچنین به طراحان سؤال کنکور پیشنهاد می‌شود با مطالعه و الگوبرداری از سؤال‌هایی که در این پژوهش و پژوهش‌های مشابه، مناسب تشخیص داده شده‌اند، طراحی سؤال‌ها را در سال‌های آینده با کیفیت بهتری انجام دهند و مؤلفان کتب درسی و دبیران شیمی نیز می‌توانند با استفاده از یافته‌های تحلیل سؤال‌ها، نقاط ضعف در یادگیری دانش‌آموزان و همچنین ابهام‌های کتاب‌های درسی را مورد بازنگری قرار دهند و به اصلاح روش‌های تدریس و متن کتاب‌های درسی اقدام کنند.

تقدیر و تشکر

بدین‌وسیله از جناب آقای دکتر رضا محمدی، معاون محترم امور آزمون‌های سازمان سنجش آموزش کشور، که داده‌های اولیه پژوهش و پاســـخنامه‌های داوطلبان را در اختیار پژوهشگر قرار دادند صمیمانه سپاسگزاری می‌شود.

References

Ahmadi, A. (2008). *Scoring using classical theory techniques and its comparison with the models of Item-Response theory in the bachelor's entrance exam in mathematics*. Master's Thesis, Tehran. Allameh Tabatabaei University. [in Persian]

Ansarin, A. (1992). *Estimation of the Item characteristic curve and the ability of the subjects in Tehran-Stanford-Binet scale based on the two-parameter model of the Latent trait*. Master's Thesis, Islamic Azad University. [in Persian]

Asadi, K., Hooman, H. A., & Liaqat, R. (2012). Comparison of one and three-parameter models of Item-Response theory in assessment the ability of high school girls by Raven's progressive matrices. *Journal of Psychological Research*, *4*(13), 71-89. [in Persian]

Bakhshifard, S. (2002). *Application of IRT model in comparison between IQ test ability scores*

at Raven progressive matrices. Master's Thesis, Islamic Azad University, Central Tehran Branch. [in Persian]

Baqaei Moghadam, P., & Roshanzamir, M. (2010). A Study of the Divisibility of the Comprehension Structure in a Foreign Language Using the Multidimensional Item response theory. *Journal of Language & Translation Studies (JLTS)*, 3, 1- 20. [in Persian]

Delavar, A., Moghadamzadeh A., & Motiei Langroudi S. T. (2006). Comparison of classical measurement model and non-parametric Item response theory models in terms of question characteristics. *Quarterly Journal of Educational Innovations,* 18, 41-56. [in Persian]

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*(4), 283-298.

Falsafinejad, M. R., Delavar, A., Farrokhi, N. A., & Mohagheghi, M.A. (2013). Comparison of classic and latent trait models in the evaluation of specialized entrance exams for medical internships in Tehran. *Iranian Journal of Medical Education, 13*(3), 167-178. [in Persian]

Farahani, M. (1996). *Comparison of classic and and IRT models in terms of question parameters and ability estimating.* Master's Thesis, Allameh Tabatabaei University. [in Persian]

Hooman, H. A. (1989). Tehran-Stanford-Binet Individual Intelligence Test. *Quarterly Journal of* (*Educational Sciences, University of Tehran, Special Issue of Psychometrics,* New Volume, (1-4. [in Persian]

Izanloo, B., & Habibi, M. (2008). Introduction to new measurement approaches in the field of psychology and educational sciences. *Journal of Psychology & Education,* 8, 135-165. [in Persian]

Karami Gazafi, A., & Niknam Z. (2014). Evaluation of mobtakeran Book Questions by IRT Method in the Periodic Properties of Elements topics from the Second Year High School Chemistry Book, *Sixth National Conference on Education*, Shahid Rajaee Teacher Training University, Tehran, Iran. [in Persian]

Karami Gazafi, A., & Niknam Z. (2015). Evaluating the questions of ghalam-Chi book on light and light reflectiontopics in the first year physics high school textbook, *7th National Conference on Education,* Shahid Rajaee Teacher Training University,Tehran, Iran. [in Persian]

Karimi, B., Falsafinejad, M., & Dartaj, F. (2011). The effect of the number of question options on the psychometric properties of the test and the estimated ability in the the Item response theory and classical models. *Educational Measurement, 6*(2). [in Persian]

Linden, W.J., Van Der. (2010). Item Response Theory, *International Encyclopedia of Education*, 81-89.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Macayan, J., & Ofalia, B. (2011). Determining experts and novices in college algebra: A psychometric test development and analysis using the Rasch model (1PL-IRT). *Educational Measurement & Evaluation Review*, 2, 62-76.

Mahmoudian, H. (2011). *Setting up a question bank in the mathematics section in the university entrance exam at 2010 based on the Item response theory*. Master's Thesis, Faculty of Psychology and Educational Sciences, Allameh Tabatabai University. [in Persian]

Maleki, H. (2013*). Introduction to curriculum planning*. Tehran: The Organization for Researching and Composing University Textbooks in the Humanities (SAMT). [in Persian]

Mohammadzadeh Romiani, M. (1996). *The Question Selection Methods in the Classical and IRT Models*, Master's Thesis, Faculty of Psychology and Educational Sciences, Allameh Tabatabaei University. [in Persian]

Nafisi, G. (1997). *Assessment and Evaluation*. Tehran: Azad University, North Branch. [in Persian]

Naghizadeh, F. (2016). *Evaluation of several questions in chemical reactions, stoichiometry and chemical Thermodynamics topics using Item response theory in junior high school chemistry,* Master's Thesis, shahid Rajaee Teacher Training University, Tehran, Iran. [in Persian]

Rouhani, M. (2012). *The analysis of Content and psychometric properties of the Islamic Azad University entrance exam for humanities science at 2007-2009*, Master's Thesis, Faculty of Psychology and Educational Sciences, Allameh Tabatabaei University. [in Persian]

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing*, Springer, New York, 151-165.

Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemical Education*, *87*(11), 1268-1272.

Sepasi, H. (2003). Comparison of concepts and assumptions between classic and Item response theory in the construction of psychological and educational tests. *Journal of Educational Sciences & Psychology, 3*(3 , 4). [in Persian]

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, *23*(1), 35-57.

Stage, C. (1998). A Comparison between Item Analysis Based on Item Response theory. A Study of SWESAT subtest RCE. *Journal of Educational Measurement*, 31.

Valeh, M. (2013). *The Comprehensive evaluation and determination of psychometric properties of aptitude and academic readiness test for management major in the entrance exam for master's*

degree using classical theory and Item response theory (IRT) at 2010-2011. Master's Thesis, Faculty of Psychology and Educational Sciences. Allameh Tabatabai University. [in Persian]
Yen, A. (2002). *Introduction to measurement theory*. Long Grove IL: waveland press.
Younesi, J. (2007). *Study of Psychometric Properties of Comprehensive test Questions for Psychology major at Payame Noor University in 2006*. Master's Thesis, Faculty of Psychology and Educational Sciences, Allameh Tabatabai University. [in Persian]