Educational Measurement and Evaluation Studies

National Organization for Educational Testing

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

The Application of IRT Polytomous Models in Scoring High-Stakes Tests (Case of Study: Lawyer's License Test)

Reza Payravi [1] , Mohammad Reza Falsafinejad [2] , Asghar Minaei [3] , Ali Delavar [4] , Noorali Farrokhi [5]

1.Ph.D student, Faculty of Psychology and Education, Allameh Tabataba'i University, Tehran, Iran, Email: Reza.Payravi@yahoo.com
2. Associate Professor, Department of Educational Measurement, Allameh Tabataba'i University, Tehran, Iran; (Corresponding Author), Email. falsafinejad@yahoo.co.uk
3. Associate Professor, Department of Educational Measurement, Allameh Tabataba'i University, Tehran, Iran, Email: asghar.minaei@yahoo.com
4. Professor, Department of Educational Measurement, Allameh Tabataba'i University, Tehran, Iran, Email: delavar@atu.ac.ir
5. Professor, Department of Educational Measurement, Allameh Tabataba'i University, Tehran, Iran, Email: farrokhi@atu.ac.ir

Article Info

Article Type:
Research Article

Received:2023.02.01
Received in revised form: 2023.05.05
Accepted: 2023.05.30
Published online: **2023.06.25**

ABSTRACT

Objective: The aim of this study was to compare the accuracy and measurement error of dichotomous and Polytomous IRT models in scoring high-stakes, large-scale ability tests.

Methods: The statistical population of this study was included all the participants of the lawyer's license external tests in 2016 and 2018, from which 5000 persons and 5000 persons respectively were selected by random sampling. In addition, data collection was done using the responses of the participants of the above exam. Accordingly, the research method is experimental.

Results: The analysis of the findings showed that among the dichotomous IRT logistic models, the 3-parameter model, and among the nominal Polytomous models studied, the 3-parameter model are a better fits and information compared with other models on the data under study.

Conclusion: Considering the more favorable fit and the level of information of the 3-parameter dichotomous model and the 3-parameter Polytomous model compared with other models, the use of these models in scoring can increase the accuracy of measurement and reduce the error. In addition, the use of these models also helps the fairness of the selection process of the applicants for the lawyer's license exam.

Keywords: *IRT scoring, Dichotomous models, IRT nominal Polytomous models, Fairness of assessment*

Cite this article: Payravi, Reza; Falsafinejad, Mohammad Reza; Minaei, Asghar; Delavar, Ali; Farrokhi, Noorali (2023). The Application of IRT Polytomous Models in Scoring High-Stakes Tests (Case of Study: Lawyer's License Test). *Educational Measurement and Evaluation Studies*, 13 (42):72-87 Pages.
DOI:10.22034/EMES.2023.563268.2426

© The Author(s).
Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی
سازمان سنجش آموزش کشور
شاپا چاپی: ۲۴۷۶-۲۸۶۵ شاپا الکترونیکی: ۲۷۸۳-۰۹۴۲

کاربرد مدل‌های چند ارزشی IRT در نمره‌گذاری آزمون‌های سرنوشت‌ساز (مورد مطالعه: آزمون پروانه وکالت)

رضا پیروی[۱]، محمدرضا فلسفی‌نژاد[۲]، اصغر مینایی[۳]، علی دلاور[۴]، نورعلی فرخی[۵]

۱. دکتری سنجش و اندازه‌گیری ، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: Reza.Payravi@yahoo.com
۲. دانشیار، گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم‌تربیتی، دانشگاه علامه طباطبائی، تهران، ایران؛ (نویسنده مسئول)، رایانامه: falsafinejad@yahoo.co.uk
۳. دانشیار، گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: asghar.minaei@yahoo.com
۴. استاد ممتاز، گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: delavar@atu.ac.ir
۵. استاد، گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: farrokhi@atu.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت: ۱۴۰۱/۱۱/۱۲
اصلاح: ۱۴۰۲/۰۲/۱۵
پذیرش: ۱۴۰۲/۰۳/۰۹
انتشار : ۱۴۰۲/۰۴/۰۴

چکیده

هدف: هدف مطالعه حاضر، مقایسه میزان دقت و خطای اندازه‌گیری مدل‌های دوارزشی و چند ارزشی IRT در نمره‌گذاری آزمون‌های توانایی سرنوشت‌ساز بود.

روش پژوهش: جامعه پژوهش شامل تمامی شرکت‌کنندگان آزمون سراسری پروانه وکالت سال‌های ۱۳۹۶ و ۱۳۹۸ بوده که از میان آن‌ها تعداد ۵۰۰۰ نفر از سال ۱۳۹۶ و تعداد ۵۰۰۰ نفر از سال ۱۳۹۸ با روش نمونه‌گیری تصادفی ساده انتخاب شدند. همچنین، گردآوری داده‌ها با استفاده از پاسخ‌های شرکت‌کنندگان آزمون انجام یافت. متغیر مستقل این پژوهش، شیوه و مدل نمره‌گذاری و متغیر وابسته، میزان برازش و آگاهی (دقت) مدل محسوب می‌شود. بر این اساس، روش پژوهش آزمایشی است.

یافته‌ها: تجزیه و تحلیل یافته‌ها نشان داد که از میان مدل‌های لجستیک دوارزشی IRT، مدل ۳ پارامتری، و از میان مدل‌های چندارزشی اسمی مورد مطالعه نیز، مدل ۳ پارامتری در مقایسه با سایر مدل‌ها، برازش و نیز آگاهی‌دهندگی بیشتر و مطلوب‌تری بر روی داده‌های مورد مطالعه داشتند.

نتیجه‌گیری: با توجه به برازش و میزان آگاهی مطلوب‌تر مدل ۳ پارامتری دو ارزشی و مدل ۳ پارامتری چندارزشی در مقایسه با سایر مدل‌ها، استفاده از این مدل‌ها در نمره‌گذاری می‌تواند به افزایش دقت اندازه‌گیری و کاهش خطا، و نیز به منصفانه بودن فرآیند گزینش متقاضیان آزمون پروانه وکالت کمک نماید.

واژه‌های کلیدی: نمره‌گذاری IRT ، مدل‌های دوارزشی IRT ، مدل‌های چندارزشی اسمی IRT ، منصفانه بودن سنجش

استناد: پیروی، رضا؛ فلسفی‌نژاد، محمد؛ مینائی، اصغر؛ دلاور، علی؛ فرخی، نورعلی (۱۴۰۲). کاربرد مدل‌های چند ارزشی IRT در نمره‌گذاری آزمون‌های سرنوشت‌ساز. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱۳ (شماره ۴۲)، ۷۲-۸۷ صفحه
DOI: 10.22034/EMES.2023.563268.2426
ناشر: سازمان سنجش آموزش کشور
حق‌مؤلف © نویسندگان.

مقدمه

در سال‌های اخیر به منظور افزایش کارایی و سودمندی سنجش شایستگی[۱] و توانایی[۲]، تمایل به مطالعه و کاربرد روش‌ها و راهبردهای متدولوژیک با هدف استخراج اطلاعات بیشتر از داده‌های پاسخ سوال، گسترش یافته است (بولت[۳]، وولاک[۴] و سو[۵] ، ۲۰۱۲ ؛ باک[۶] و گیبونز[۷]، ۲۰۲۱). مدل‌های اندازه‌گیری که چارچوبی برای پایش، تفسیر نمرات و تصمیم‌گیری فرآهم می‌کنند، همواره ضرورت دارد واجد شاخص‌ها و ویژگی‌های مطلوبی باشند که از جمله این شاخص‌ها، دقت در اندازه‌گیری است (تامپسون[۸] ، ۲۰۲۱). اهمیت دقت اندازه‌گیری[۹] به این دلیل است که "هدف غایی و نهایی اندازه‌گیری، همواره تولید اندازه یا نمره‌ای است که توسط آن افراد بتوانند ارزیابی و تفکیک شوند" (براون[۱۰] و گرودس[۱۱]، ۲۰۱۵، ص۳۰۷). بنا به اعتقاد پرایس[۱۲] (۲۰۱۷)، "برای این‌که یک مدل اندازه‌گیری موثر در دست داشته باشیم، دقت اندازه‌گیری و عینیت[۱۳] از عناصر اساسی آن محسوب می‌شود" (ص۱۴۶).

اساسا دو شیوه برای نمره‌گذاری آزمون‌های چندگزینه‌ای توانایی، موجود است که عبارتند از شیوه مبتنی بر رویکرد کلاسیک (CTT) و شیوه مبتنی بر نظریه پرسش پاسخ (IRT). مطابق ادعای برخی متون اندازه‌گیری، اعتبار[۱۴] نمرات حاصل از نمره‌گذاری IRT بیشتر از نمره‌گذاری CTT است (دی مارس[۱۵]، ۲۰۱۰؛ ون‌در لیندن[۱۶]، ۲۰۱۶). در حال حاضر و در شرایط موجود نمره‌گذاری آزمون‌های ملی، اگرچه مهمترین مزیت مدل کلاسیک، سهولت و سادگی فرآیند نمره‌گذاری و فهم ساده شیوه نمره‌گذاری برای کاربران نمره (مانند فراگیر، معلم، والدین و مدیران آموزشی) محسوب می‌شود، اما این مدل از محدودیت‌هایی نیز در نمره‌گذاری برخوردار است (تامپسون، ۲۰۲۱). نخستین و مهمترین محدودیت مدل کلاسیک، این است که در این مدل، ارزش تمام سوالات یک آزمون برابر و یکسان فرض می‌شود. این فرض درحالی است که می‌دانیم شاخص‌های آماری و ویژگی‌های روان‌سنجی سوال‌های یک آزمون، همانند شاخص دشواری و ضریب تمیز، غالبا با یکدیگر متفاوت هستند. روشن است، *ارزش ارائه پاسخ درست به یک سوال دشوار، بسیار بیشتر از ارزش ارائه پاسخ درست به یک سوال ساده است* و یا این‌که یک سوال با مقدار تمیز زیاد در مقایسه با یک سوال با مقدار تمیز پایین، آگاهی[۱۷] بیش‌تری فرآهم می‌کند. لذا منطقی است انتظار داشته باشیم سوال‌های دشوارتر و یا ممیزتر نمره بیش‌تری کسب کنند (لاکورلی[۱۸]، سن‌مارتین[۱۹]، سیلوا[۲۰] و همکاران، ۲۰۱۸).

در شیوه نمره‌گذاری کلاسیک، به پاسخ درست، نمره یک و به پاسخ نادرست، نمره صفر اختصاص داده می‌شود و از این‌رو، این روش را گاهی نمره‌گذاری دو مقوله‌ای یا دو وجهی[۲۱] نیز می‌نامند. در این شیوه، هیچگونه شاخص دیگری همچون مقدار سختی و دشواری یا تمیز سوال برای تعیین نمره سوال و نمره‌گذاری در نظر گرفته نمی‌شود و به نظر می‌رسد کمتر عادلانه[۲۲] است (میسکوسکی[۲۳] و استورم[۲۴]، ۲۰۱۸). این

1. Efficiency
2. Ability
3. Bolt
4. Wollack
5. Suh
6. Bock
7. Gibbons
8. Thompson
9. Measurement precision
10. Brown
11. Croudace
12. Price
13. Objectivity
14. Reliability
15. DeMars
16. Van der Linden
17. Information
18. Lacourly
19. Sanmartin
20. Silva
21. Binary Scoring
22. Fair
23. Myszkowski
24. Storme

محدودیتCTT، در نمره‌گذاریIRT (مدل‌های دوارزشی[1] و مدل‌های چندارزشی[2]) مرتفع شده و در واقع، برای سوالات یک آزمون با توجه به مقادیر مختلف شاخص‌های سوال، ارزش‌های متفاوتی برآورد و تعیین می‌شود و به این ترتیب، "دقت نمره توانایی افزایش می‌یابد" (سامجیما[3] ، ۱۹۹۶، ص۱۲).

دشواری و محدودیت دوم نمره‌گذاری کلاسیک، عدم توجه به گزینه‌های انحرافی (انتخاب شده توسط آزمودنی) در ساخت نمره است. بنا به گفته دی‌آیالا[4]، پاسخ‌های انحرافی آزمون‌های توانایی حاوی اطلاعاتی است که از این اطلاعات می‌توان برای افزایش دقت نمره استفاده نمود (۲۰۰۹). درCTT و همچنین مدل‌های دو ارزشیIRT، پاسخ‌های انحرافی نقشی در ساخت نمره سوال و آزمون ندارند. در مدل‌های دو ارزشیIRT ، اگرچه ارزش سوال‌ها نابرابر است اما نمره فرد همانندCTT بر اساس تعداد پاسخ درست (تعداد گزینه‌های درست انتخاب شده) تعیین می‌شود. در مقابل، در مدل‌های چند ارزشیIRT علاوه بر این‌که برای سوالات مختلف، ارزش‌های متفاوتی تعیین می‌شود، به پاسخ‌های انحرافی نیز بر اساس میزان درستی[5] یا میزان جذابیت[6]، نمره تعلق می‌گیرد و به این ترتیب، از پاسخ‌های انحرافی نیز اطلاعاتی به منظور برآورد دقیق‌تر پارامتر توانایی استخراج می‌شود (کیم[7]، ۲۰۰۶). به این ترتیب، یکی از پیشرفت‌های اساسی در مدل‌سازی پاسخ سوال‌های چندگزینه‌ای عبارت است از امکان بازیابی اطلاعات از پاسخ‌های انحرافی سوال، که این امکان به واسطه وجود مدل‌های چند ارزشی انجام می‌شود (باک، ۱۹۷۲).

همانطور که پنفیلد و لاتوره (۲۰۰۸) نیز بیان می‌کنند، رویکردهای موجود در اندازه‌گیری که به منظور مدل سازی پا سخ[8] سوالات چندگزینه‌ای به‌کار می‌روند، در دو طبقه قرار می‌گیرند:

الف) رویکردی که پاسخ درست سوال را در یک طبقه و تمام پاسخ‌های انحرافی را در یک طبقه دیگر قرار می‌دهد. مدل‌های مبتنی بر این رویکرد را مدل‌های دو مقوله‌ای یا دوارزشی می‌نامند. نوع رفتار و برخورد این مدل‌ها با *پاسخ* سوال، به صورت قبول- مردود (دوارزشی) است. قبول به معنای انتخاب پا سخ (گزینه) در ست و مردود به معنای عدم انتخاب پا سخ در ست. نقطه کانونی این مدل‌ها این ا ست که در آن‌ها، فقط احتمال انتخاب *"پاسخ درست"* به عنوان تابعی از توانایی، مدل‌سازی می‌شود و برای گزینه‌های انحرافی، هویت مستقلی در نظر گرفته نمی‌شود. به این ترتیب، مدل‌های دوارزشیIRT در این طبقه جای می‌گیرند.

ب) رویکردی که میان تمام پاسخ‌های یک سوال، تمایز قایل می‌شود بطوری‌که احتمال انتخاب هر پاسخ (خواه درست و خواه نادرست) به‌صورت جداگانه بررسی شده و هر پاسخ بطور جدا مدل‌سازی می‌شود. مدل‌هایی که در این طبقه قرار می‌گیرند را مدل‌های چندارزشی می‌نامند. بنابراین می‌توان گفت، تفاوت نمره‌گذاری دوارز شی با نمره‌گذاری چندارز شی در نوع نگاه و رویکرد آن‌ها به پا سخ‌های انحرافی ا ست. در نمره‌گذاری دو ارز شی، یک پا سخ بعنوان پا سخ در ست و مابقی پا سخ‌ها بعنوان پا سخ‌های نادر ست در نظر گرفته می شود و *تفاوتی ندارد* آزمودنی کدام پا سخ انحرافی را انتخاب می‌کند. در این مدل‌ها فرض می شود همه پا سخ‌های انحرافی دارای نادر ستی یک‌سان و برابر هستند. در مقابل، در مدل‌های چندارزشی، پاسخ‌های انحرافی با نادرستی نابرابر در نظر گرفته می‌شوند و احتمال انتخاب هر پاسخ بطور جداگانه بررسی و مدل‌سازی می‌شود.

دلیل منطقی و شهودی پذیرش فرض نابرابری پاسخ‌های انحرافی را این طور می‌توان بیان کرد:

"وقتی یک آزمودنی، سوالی را بطور نادرست پاسخ می‌دهد، چه بسا یک مقدار دانش جزیی از پاسخ درست داشته باشد" (دی‌آیالا، ۱۹۸۹، ص۷۸۹). به قول میسکوسکی و استورم، " انتخاب برخی از پاسخ‌های انحرافی در مقایسه با انتخاب برخی پاسخ‌های دیگر، نشان دهنده توانایی بیشتری است" (۲۰۱۸، ص ۱۱۴). بنابراین می‌توان اظهار داشت، *هدف* مدل‌های چند ارزشی، افزایش دقت پارامتر توانایی از طریق استفاده از اطلاعات موجود در پاسخ‌های انحرافی سوال است.

1. Dichotomous model
2. Polytomous or Polychotomous model
3. Samejima
4. De Ayala
5. Correctness
6. Attractive
7. Kim
8. Response modeling

مبانی نظری و پیشینه پژوهش

همه ساله متقاضیان گسترده‌ای به طور منظم و سراسری در آزمون پروانه وکالت که یک برنامه سنجش گسترده مقیاس و یک آزمون سراسری توانایی سرنوشت‌ساز محسوب می‌شود، شرکت می‌کنند. *سنجش گسترده مقیاس*[۱] (LSA) اساسا به برنامه‌ای از سنجش اطلاق می‌شود که بر روی تعداد زیادی از افراد برای مقاصد گوناگونی (از قبیل رصد و مقایسه برنامه‌های آموزشی بین‌المللی، ارزشیابی برنامه‌های تحصیلی ملی، پذیرش ورود به دانشگاه، اعطای بورس تحصیلی و یا اعطای گواهینامه شغلی) اجرا می‌شوند (سیمون[۲]، ارکیکان[۳] و روسو[۴]، ۲۰۱۳). آزمون سرنوشت‌ساز[۵] (HST) نیز به آزمونی گفته می‌شود که بر اساس نتایج و نمرات آن‌ها تصمیماتی پراهمیت در باره آینده تحصیلی، شغلی و یا اجتماعی آزمودنی‌ها اتخاذ می‌گردد به طوری‌که این تصمیمات دارای تبعات و پیامدهای مهم برای فرد و جامعه هستند (ریت[۶]، ۲۰۱۶). بی‌گمان، یکی از مهم‌ترین عوامل موثر و تهدید کننده دقت اندازه‌گیری آزمون‌ها و نیز "یکی از ملاحظات اساسی در اندازه‌گیری آموزشی، شیوه نمره‌گذاری سوالات آزمون است" (دانینگ[۷] و هالادینا[۸] ، ۲۰۱۱، ص ۴۵۱).

در آزمون‌های گسترده مقیاس با شیوه نمره‌گذاری دو وجهی، غالبا مواقعی پیش می‌آید که افراد بسیاری به ویژه در میانه توزیع نمرات و اطراف نقطه برش، نمرات یکسان و یا کم و بیش یکسانی کسب می‌کنند. این وضعیت، موجب می‌شود در دامنه‌ای از توزیع نمرات، تراکم سنگینی ایجاد شود. انباشت نمرات برابر (و یا کم و بیش یکسان) سبب بروز دشواری‌هایی از جمله ایجاد گره در توزیع نمرات آزمون می شود. وجود گره[۹] در توزیع نمرات باعث می‌شود امکان تمایز و رتبه‌بندی[۱۰] مطلوب و دقیق آزمودنی‌ها از دست برود. در این شرایط، مدل‌های IRT بویژه مدل‌های چند ارزشی اسمی می‌توانند با بکارگیری اطلاعات از پاسخ‌های انحرافی و نمره‌گذاری آن‌ها، موجب پراکندگی بیشتر در توزیع نمرات شده و از بروز گره در توزیع جلوگیری نمایند. مطابق متون IRT ، اساسا دقت در سطوح مختلف نمره، متفاوت است و نباید یک مقدار ثابت و مشترک (همان‌طور که در CTT مطرح است) به عنوان شاخص دقت برای تمام نمرات یک آزمون در نظر گرفته شود. در این رویکرد، تابع آگاهی $I(\theta)$ ، به عنوان شاخص دقت شرطی اندازه‌گیری محسوب می‌شود (بیکر[۱۱] و هوکیم[۱۲]، ۲۰۱۷). از این‌رو، "مدل‌های چندارزشی مانند مدل پاسخ اسمی باک[۱۳] به منظور مطالعه کارکرد پاسخ‌های انحرافی سوال‌های چندگزینه‌ای آزمون‌های توانایی طراحی گردید" (پرستون[۱۴]، رایس، کای[۱۵] و همکاران، ۲۰۱۱ ، ص۵۲۶). پیش‌فرض کاربرد مدل‌های چند ارزشی اسمی برای آزمون‌های چندگزینه‌ای توانایی، عبارت است از این که این آزمون‌ها از سوال‌های چند ارزشی تشکیل شده باشد. سوال چند ارزشی[۱۶] به سوالی اطلاق می‌شود که پاسخ‌های انحرافی آن از یک *درستی نسبی یا جذابیت نسبی* برخوردار باشند به طوری‌که بتوان پاسخ‌ها را بر اساس درستی یا جذابیت در ۳ طبقه (یا بیش از ۳ طبقه) قرار داد و نمره‌گذاری نمود و به این ترتیب، دقت نمره را افزایش داد (پنفیلد، ۲۰۱۴).

در تاریخ اندازه‌گیری آموزشی، مدل پاسخ اسمی باک (NRM) ، نخستین مدل چند ارزشی اسمی است که در سال ۱۹۷۲ توسط دارل باک ارائه شده است (باک، ۱۹۹۷؛ ون‌در لیندن، ۲۰۱۶). این مدل یک مدل دو پارامتری است و اگرچه تاکنون رایجترین و پرکاربردترین مدل چند ارزشی اسمی بوده و پژوهش‌ها نشان داده برازش خیره‌کننده‌ای با داده‌ها دارد، اما محدودیت‌هایی نیز به همراه دارد (دراسگو[۱۷]، لوین[۱۸] و تسین[۱۹]،

1. Large Scale Assessment
2. Simon
3. Ercikan
4. Rousseau
5. High Stakes Test
6. Ritt
7. Downing
8. Haladyna
9. Knot
10. Ranking
11. Baker
12. Ho Kim
13. Nominal Response Model (NRM)
14. Pretson
15. Cai
16. Polytomous item
17. Drasgow
18. Levine
19. Tsien

۱۹۹۵). مهمترین محدودیت این مدل عبارت است از عدم دخالت پارامتر حدس در برآورد پارامترهای سوال و توانایی. مدل باک، فقط به برآورد پارامترهای د شواری و تمیز می‌پردازد و مقدار پارامتر حدس را برای تمام سوال‌ها برابر با صفر در نظر می‌گیرد (ایباد[1]، اوله‌آ[2]، پونسودا[3]، ۲۰۰۹). علاوه بر این، به گفته پنفیلد[4] و لاتوره[5] (۲۰۰۸)، مدل دو پارامتری باک اگرچه یک مکانیسم انعطاف‌پذیر برای مدل سازی داده‌های پاسخ اسمی فراهم می‌سازد، ولی یک محدودیت مهم نظری دارد. این پژوهشگران بیان می‌کنند:

در یک سوال فرضی، گزینه‌ای که *بیشترین* مقدار ضریب تمیز منفی را با یک خط اثر[6] کاهشی یکنواخت در اختیار دارد، دارای یک مجانب پایینی[7] کمتر از صفر و مجانب بالایی[8] بزرگتر از یک می‌شود در حالی‌که مجانب پایینی، بایستی صفر یا نزدیک به صفر و مجانب بالایی برابر با یک یا نزدیک به یک باشد. " نتیجه این‌که، مدل باک ممکن است در برخی از سوالات برازش ضعیفی داشته باشد" (پنفیلد و لاتوره، ۲۰۰۸، ص ۵).

امروزه، مدل‌های چندارزشی پیشرفته‌تری به نام مدل‌های لوجیت آشیانه‌ای[9] (NLM) در دسترس قرار دارد که به منظور مدل‌سازی پاسخ آزمون‌های توانایی چندگزینه‌ای طراحی شده است (سو و بولت، ۲۰۱۰). در واقع، NLM خانواده‌ای از مدل‌های چندارزشی اسمی[10] است. این مدل‌ها که جدید بوده و دارای پشتوانه مستحکم پژوهشی هستند، عبارتند از:

الف) مدل دو پارامتری چندگزینه‌ای (2PL-NLM)

ب) مدل سه پارامتری چندگزینه‌ای (3PL-NLM)

ج) مدل چهار پارامتری چندگزینه‌ای (4PL-NLM)

در اینجا مدل ۳ پارامتری چندارزشی آشیانه‌ای سو و بولت را با جزییات بیشتری به عنوان نمونه و یک مدل رایج‌تر بررسی می‌کنیم.

فرض کنید یک آزمون چندگزینه‌ای از n سوال تشکیل شده و هر سوال دارای یک گزینه درست و m_i گزینه انحرافی است. به عبارت دیگر، مجموعا $m_i + 1$ مقوله پاسخ. همچنین فرض کنید U_{ij}، پاسخ آزمودنی j به سوال i است $(j = 1, \ldots, N)$ که برای پاسخ درست، ارزش برابر یک و برای پاسخ نادرست ارزش برابر *صفر* منظور می‌شود.

علاوه بر این فرض کنید، D_{ijv} نشان دهنده پاسخ آزمودنی j به گزینه انحرافی v سوال i است. مطابق مدل مذکور، احتمال این‌که یک آزمودنی با توانایی معین θ، گزینه درست سوال i را انتخاب کند به صورت معادله شماره ۱، مدل‌سازی می‌شود :

معادله شماره ۱

$$P(U_{ij} = 1 | \theta_j) = \gamma i + (1 - \gamma i) \frac{1}{1 + \exp^{-(\beta i + \alpha i \theta j)}}$$

در معادله بالا، پارامترها به شکل زیر نام‌گذاری می‌گردد :

βi (پارامتر عرض از مبدا)[11]

α_i (پارامتر شیب)[12]

γ_i (پارامتر پایینی یا پارامتر حدس)[13]

1. Abad
2. Olea
3. Ponsoda
4. Penfield
5. La Torre
6. Trace line
7. Lower asymptote
8. Upper asymptote
9. Nested Logit Models (NLM)
10. Nominal Polytomous Models
11. Intercept parameter
12. Slope parameter
13. Pseudo-guessing parameter

همچنین مطابق این مدل، احتمال این‌که یک آزمودنی با توانایی معین θ ، گزینه نادرست سوال i را انتخاب کند به صورت معادله شماره ۲ مدل‌سازی می‌شود :

معادله شماره ۲

$$P(U_{ij}=0, D_{ijv}=1|\theta_j) = \{1-[\gamma_i+(1-\gamma_i)1/1+\exp^{-(\beta_i+\alpha_i\theta_j)}]\}[\exp^{z_{iv}(\theta j)}/\Sigma \exp^{z_{ik}(\theta j)}]$$

از نقطه نظر فنی و تکنیکی، تفاوت مدل‌های چندارزشی NLM با مدل‌های دوارزشی IRT و مدل باک، عبارت است از این که " فرآیند درجه‌بندی در مدل‌های NLM ، طولانی‌تر و تعداد پارامترهایی که برآورد می‌شود بیشتر است" (میسکوسکی و استورم، ۲۰۱۸، ص۱۱۴). همچنین، پارامتر پردازی در مدل‌های آشیانه‌ای با دشواری و به‌سختی انجام می‌پذیرد و ثبات پارامترها نسبت به حجم نمونه بسیار حساس است. مدل‌های NLM که به مدل‌های ترکیبی[1] نیز شناخته می‌شوند از ترکیب مدل‌های دوارزشی IRT با مدل چندارزشی باک حاصل شده‌اند (سو و بولت، ۲۰۱۰)؛ به این معنی که در این مدل‌ها برای گزینه کلید سوال، بنا به شرایط، از یکی از مدل‌های دوارزشی لجستیک IRT و برای پاسخ‌های انحرافی، از مدل باک به منظور مدل سازی استفاده می شود. لذا می‌توان گفت، مدل‌های آشیانه‌ای در دو سطح به مدل سازی پاسخ‌ها می‌پردازد:

سطح ۱ یا سطح بالاتر، که عبارت است از توصیف احتمال انتخاب پاسخ درست (تمایزگذاری میان پاسخ درست با پاسخ‌های انحرافی).

سطح ۲ یا سطح پایین‌تر، که عبارت است از توصیف و تبیین احتمال انتخاب هر یک از پاسخ‌های انحرافی (تمایزگذاری میان پاسخ‌های انحرافی).

بنابراین و به بیان دیگر می‌توان گفت، در مدل‌های آشیانه‌ای گزینه‌های درست و انحرافی از یکدیگر جدا شده و درون ۲ طبقه آشیانه[2] می‌کنند: " یک طبقه منحصرا شامل پاسخ درست و طبقه دوم شامل کلیه پاسخ‌های انحرافی" (سو و بولت،۲۰۱۰، ص۴۵۷) .

مطالعه ادبیات پژوهشی نشان می‌دهد، استورم، میسکوسکی، بارون[3] و همکاران (۲۰۱۹) در یک بررسی بر روی داده‌های حاصل از ۲۹۴۹ شرکت‌کننده یک آزمون که شامل سوال‌های چندگزینه‌ای توانایی عمومی شناختی[4] بود، به مقایسه عملکرد مدل‌های لجستیک دو ارزشی (1PL-P2L-3PL-4PL)، با مدل‌های چند ارزشی اسمی شامل مدل باک و مدل‌های چند ارزشی آشیانه‌ای سو و بولت که شامل مدل‌های 2PNL, 3PNL,4PNL می‌شود، مبادرت کرده و به یافته‌های زیر دست یافتند:

۱- همه مدل‌های دو ارزشی، برازش رضایتبخشی با داده‌ها دارند.

۲- مدل پاسخ اسمی باک، برازش قابل قبولی با داده‌ها دارد.

۳- همه مدل‌های NLM در مقایسه با مدل NRM ، برازش بهتری دارند.

۴- همه مدل‌های NLM ، برازش رضایتبخشی با داده‌ها دارند. بطور کلی، یافته‌های این پژوهش نشان داد مدل‌های NLM جایگزین مناسبی برای مدل NRM و مدل‌های دو ارزشی IRT هستند.

در پژوهشی دیگر، میسکوسکی و استورم (۲۰۱۸) به مطالعه پاسخ‌های ۴۹۹ آزمودنی به سوالات چندگزینه‌ای آزمون ماتریس‌های پیشرونده ریون[5] پرداختند. هدف این مطالعه عبارت بود از مقایسه مدل‌های دوارزشی IRT با مدل‌های چندارزشی اسمی. یافته‌های این بررسی نشان داد:

الف) از میان مدل‌های لجستیک دو ارزشی، مدل‌های سه و چهار پارامتری برازش رضایت‌بخشی با داده‌ها دارند.

ب) در میان مدل‌های چند ارزشی اسمی مورد مطالعه، مدل ۳ و ۴ پارامتری آشیانه‌ای بهترین برازش را بر روی داده‌ها داشتند.

ج) مدل 3PNL ، اطلاعات بیشتری از گزینه‌های انحرافی بویژه از سطوح پایین مقیاس توانایی به دست می‌دهد.

د) اگرچه همبستگی میان نمرات حاصل از مدل‌های مختلف بسیار بالاست، لیکن مدل‌های آشیانه‌ای برآورد متفاوتی برای بخش زیادی از مقیاس توانایی، بویژه نیمه پایینی مقیاس فراهم می‌کند. به طور کلی نتایج این پژوهش نشان داد در میان مدل‌های دو ارزشی، مدل ۳ پارامتری، و در بین مدل‌های چند ارزشی نیز، مدل ۳ پارامتری آشیانه‌ای برازش بهتری با داده‌ها دارند.

1. Combined models
2. Nest
3. Baron
4. General Mental Ability (GMA)
5. Raven's progressive matrices

به این ترتیب، در سنجش‌های گسترده مقیاس توانایی، یک سیستم نمره‌گذاری که قابلیت تفکیک بیشتر آزمودنی‌ها را داشته باشد، از اهمیت زیادی برخوردار بوده و شناسایی مدل‌های بهینه نمره‌گذاری که بالقوه قابلیت دستیابی به سنجش منصفانه را داشته و به برقراری بیشتر عدالت در سنجش کمک نماید، مقصود اصلی این پژوهش است. به طور کلی، *هدف* این پژوهش عبارت است از مقایسه دقت نمره‌گذاری مدل‌های دو ارزشی و مدل‌های چند ارزشی اسمی IRT و بررسی امکان کاربرد آن‌ها در نمره‌گذاری آزمون‌های انتخابی و گزینشی[1] است.

روش پژوهش

پژوهش حاضر با توجه به ماهیت مساله مورد مطالعه و در نظر گرفتن مدل‌های نمره‌گذاری آزمون (به عنوان متغیر مستقل و میزان برازش مدل‌ها به عنوان متغیر وابسته)، در زمره مطالعات آزمایشی و از لحاظ هدف در حوزه پژوهش‌های کاربردی قرار می‌گیرد. در این پژوهش، مدل‌های چندارزشی اسمی IRT که به آن‌ها مدل‌های چندگزینه‌ای[2] نیز گفته می‌شود، مورد مطالعه و با مدل‌های دو ارزشی IRT مورد مقایسه قرار گرفتند. برای تحلیل داده‌ها، ابتدا ضروری بود که سنجش ابعاد آزمون انجام گیرد. به منظور اطمینان از تعداد ابعاد (با هدف تنظیم و تهیه ابزارهای تجزیه و تحلیل داده‌ها)، تحلیل بر روی ابعاد آزمون با برنامه R انجام یافت و نتایج نشان داد داده‌های هر دو آزمون تک بعدی است. به منظور برآورد پارامترهای سوال و توانایی و مقایسه مدل‌ها با یکدیگر، برخی برنامه‌های کامپیوتری از قبیل نرم افزار R شامل بسته‌های mcIRT (ریف[3]، ۲۰۱۴)، mirt (چالمرز[4] ، ۲۰۱۲)، بسته psych (روله[5]، ۲۰۱۷) و نرم‌افزار SPSS مورد استفاده قرار گرفتند. همچنین، گردآوری داده‌ها با استفاده از پاسخ‌های شرکت‌کنندگان آزمون وکالت سال‌های ۱۳۹۶ و ۱۳۹۸ انجام شد.

درباره نوع برخورد با داده‌های گمشده[6] در تحلیل، باید اظهار داشت با توجه به این‌که آزمون وکالت دارای مدت زمان پاسخگویی کافی و مناسبی است (۱۲۰ سوال و با مدت زمان پاسخگویی ۱۲۰ دقیقه) و جزو آزمون‌های با محدودیت زمانی گنجانده نمی‌شود، لذا با سوال‌هایی که پاسخ داده نشده[7] (پاسخ سفید) به مثابه سؤال‌هایی رفتار شد که آزمودنی‌ها فرصت مطالعه آن‌ها را داشته اما توان ارائه پاسخ مناسب به آن‌ها را نداشته‌اند. از این‌رو، به این سؤال‌ها، امتیاز و نمره‌ای تعلق نمی‌گیرد.

همچنین، جامعه آماری پژوهش شامل تمامی شرکت‌کنندگان آزمون وکالت سال ۱۳۹۶ و سال ۱۳۹۸ کشور است و نمونه‌ها با روش نمونه‌گیری تصادفی ساده انتخاب گردید. تعداد نمونه پژوهشی آزمون ۱۳۹۶ برابر با ۵۰۰۰ نفر بوده که ۲۴۲۱ نفر از آن‌ها زن و ۲۵۷۹ نفر را مرد تشکیل می‌دهند. همچنین، نمونه آزمون سال ۱۳۹۸ از تعداد ۵۰۰۰ نفر تشکیل شده که ۲۳۵۶ نفر از آن‌ها زن و ۲۶۴۴ نفر نیز مرد می‌باشند.

یافته‌ها

به منظور دستیابی به اهداف پژوهش، با استفاده از مدل‌های دوارزشی IRT (مدل‌های دو، سه و چهار پارامتری) و نیز مدل‌های چند ارزشی اسمی IRT (شامل مدل پاسخ اسمی باک و مدل‌های دو پارامتری، سه پارامتری و چهار پارامتری آشیانه‌ای)، به مقایسه برازش و دقت اندازه‌گیری هر یک از این مدل‌ها پرداخته شد که نتایج در پی می‌آید.

در ابتدا، مفروضه‌های اساسی و اولیه کاربرد مدل‌های IRT، یعنی مفروضه‌های تک بعدی بودن و استقلال موضعی بررسی گردید و مطابق تحلیل انجام یافته، آزمون پروانه وکالت سال‌های ۱۳۹۶ و ۱۳۹۸ تک‌بعدی بوده و بدلیل وجود مفروضه تک بعدی بودن، مفروضه استقلال موضعی نیز برقرار است. سپس، مطلوبیت هر یک از مدل‌های دو ارزشی و چند ارزشی IRT در داده‌های آزمون سال ۱۳۹۶ و ۱۳۹۸، با استفاده از دو شاخص مطالعه گردید. این دو شاخص عبارتند از : الف) شاخص برازش مدل‌ها ، ب) شاخص دقت و آگاهی، که در ادامه به توضیح آن می‌پردازیم.

الف) شاخص برازش مدل‌ها :

به منظور مقایسه مدل‌ها، در ابتدا شاخص‌های برازش مدل‌ها در هر دو حالت (دوارزشی و چندارزشی)، محاسبه و مورد بررسی قرار گرفت تا تعیین شود کدام یک از مدل‌ها با داده‌های آزمون وکالت سال ۱۳۹۶ و ۱۳۹۸ برازش بهتری دارد. در واقع، "روایی[8] نتایج حاصل از کاربرد مدل،

1. Selective
2. Multiple choice
3. Reif
4. Chalmers
5. Revelle
6. Missing data
7. Omitted responses
8. Validity

به انطباق میان مدل و داده‌های آزمون وابسته است" (همبلتون و کوک، ۱۹۷۷، ص۸۴). لذا به منظور بررسی برازش، از شاخص آگاهی آکائیک[1] استفاده شد. " این شاخص، ملاک برازش مدل محسوب شده و برای مقایسه مدل‌ها مورد استفاده قرار می‌گیرد" (پک[2] و کول[3]، ۲۰۲۰، ص۴۴). در هنگام مقایسه مدل‌ها، مقادیر این شاخص هرچه کوچکتر باشد، نشان‌دهنده برازش مناسب‌تر مدل است. مدلی که توسط AIC به عنوان مدل مناسب تشخیص داده می‌شود، نه دارای بیش‌برازش[4] است و نه کم‌برازش[5] و می‌توان آن را مدلی با برازش مناسب در نظر گرفت و همچنین به کمک آن می‌توان یک رابطه ترتیبی[6] بین مدل‌ها، به منظور مقایسه و سنجش برتری بین آن‌ها به‌دست آورد (برنهام[7] و اندرسون[8]، ۲۰۰۲ و ۲۰۰۴). در جدول شماره ۱، مقادیر شاخص‌های برازش مدل‌های دوارزشی و مدل‌های چندارزشی آمده است.

جدول ۱. مقادیر شاخص برازش مدل‌ها

آزمون	سطح تحلیل	مدل نمره‌گذاری	شاخص AIC
۱۳۹۶	دوارزشی	**3PLM**	۶۲۵۱۷۴/۱
		4PLM	۶۲۵۵۵۶/۹
	چندارزشی	NRM	۱۴۹۷۲۰۰
		2PL-NLM	۱۴۹۵۷۰۶
		3PL-NLM	۱۴۹۲۶۵۴
		4PL-NLM	۱۴۹۳۲۱۵
۱۳۹۸	دوارزشی	**3PLM**	۶۶۹۲۲۲/۴
		4PLM	۶۶۹۹۳۱/۵
	چندارزشی	NRM	۱۵۸۳۵۵۶
		2PL-NLM	۱۵۸۱۱۸۸
		3PL-NLM	۱۵۷۳۸۹۹
		4PL-NLM	۱۵۷۵۵۴۴

بطور خلاصه، مقادیر این جدول (با سطح معناداری ۰،۰۰۱) نشان می‌دهد:

الف) مدل‌های دوارزشی در مقایسه با مدل‌های چندارزشی، برازش بهتری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارند.

ب) در میان مدل‌های دو ارزشی، مدل ۳ پارامتری (با فاصله اندک از مدل ۴ پارامتری) برازش مطلوب‌تری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارد.

ج) همچنین، در میان مدل‌های چندارزشی، مدل ۳ پارامتری (با فاصله اندک از مدل ۴ پارامتری)، برازش بهتری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارد.

بنا بر آنچه گفته شد و بر اساس نتایج به دست آمده، در این پژوهش از میان مدل‌های دوارزشی و مدل‌های چندارزشی، مدل۳ پارامتری دو ارزشی و مدل۳ پارامتری چندارزشی به عنوان مدل‌هایی با برازندگی مناسب‌تر انتخاب گردید و به این ترتیب، میانگین پارامترهای برآورد شده مدل ۳ پارامتری دو ارزشی و مدل ۳ پارامتری چندارزشی به تفکیک آزمون سال ۱۳۹۶ و سال ۱۳۹۸ در جداول شماره ۲ تا ۴ آورده شده است. جدول شماره ۲، میانگین پارامترهای سوال‌های آزمون سال ۱۳۹۶ و ۱۳۹۸ را بر اساس مدل ۳ پارامتری دوارزشی نشان می‌دهد.

1. Akaike Information Criterion (AIC)
2. Paek
3. Cole
4. Overfitting
5. Underfitting
6. Ordered Relation
7. Burnham
8. Anderson

جدول شماره ۲. میانگین پارامترهای آزمون بر اساس مدل ۳ پارامتری دوارزشی – 3PLM

آزمون	۱۳۹۶			۱۳۹۸		
میانگین	شیب	دشواری	حدس	میانگین	شیب	دشواری	حدس
	۱/۵۲	-۱/۲۰	۰/۰۸		۲/٤٥	-۲/۰٥	۰/۱۷

در جدول شماره ۳ نیز میانگین پارامترهای سوال‌های آزمون سال ۱۳۹۶ به‌علاوه شیب و دشواری گزینه‌های سوال بر اساس مدل ۳ پارامتری چندارزشی ارائه شده است.

جدول ۳. میانگین پارامترهای آزمون ۱۳۹٦ (مدل ۳ پارامتری چندارزشی - 3PL-NLM)

میانگین	دشواری	حدس	شیب ۱	شیب ۲	شیب ۳	شیب ۴	دشواری ۱	دشواری ۲	دشواری ۳	دشواری ۴
	-۲/۲۵	۰/۱۵	۰/۸۱	۰/۸۳	۰/۷۹	۰/۸۶	-۰/۸۱	-۰/۹۰	-۰/۸۰	-۶/۷۰

همچنین، در جدول شماره ۴ میانگین پارامترهای سوالات آزمون سال ۱۳۹۸ به‌علاوه شیب و دشواری گزینه‌ها بر اساس مدل ۳ پارامتری چندارزشی آورده شده است.

جدول ٤. میانگین پارامترهای آزمون ۱۳۹۸ (مدل ۳ پارامتری چندارزشی - 3PL-NLM)

میانگین	دشواری	حدس	شیب ۱	شیب ۲	شیب ۳	شیب ۴	دشواری ۱	دشواری ۲	دشواری ۳	دشواری ۴
	-۲/۸۰	۰/۲۰	-۰/۹٥	-۱	۱	۰/۲۰	۰/۲٥	۰/۳٥	-۳	-۹

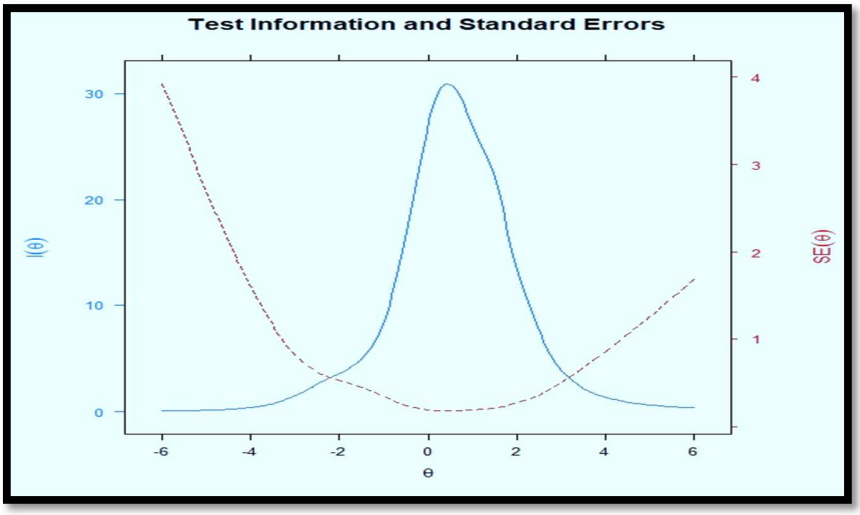
ب) شاخص دقت و آگاهی :

علاوه بر بررسی شاخص برازش، مدل‌های مورد مطالعه با استفاده از شاخص تابع آگاهی نیز مورد ارزیابی قرار گرفتند. بررسی آگاهی داده‌های آزمون سال ۱۳۹۶ در شکل‌های شماره‌های ۱ و ۲ و سال ۱۳۹۸ در شکل‌های شماره‌های ۳ و ۴ نشان می دهد:

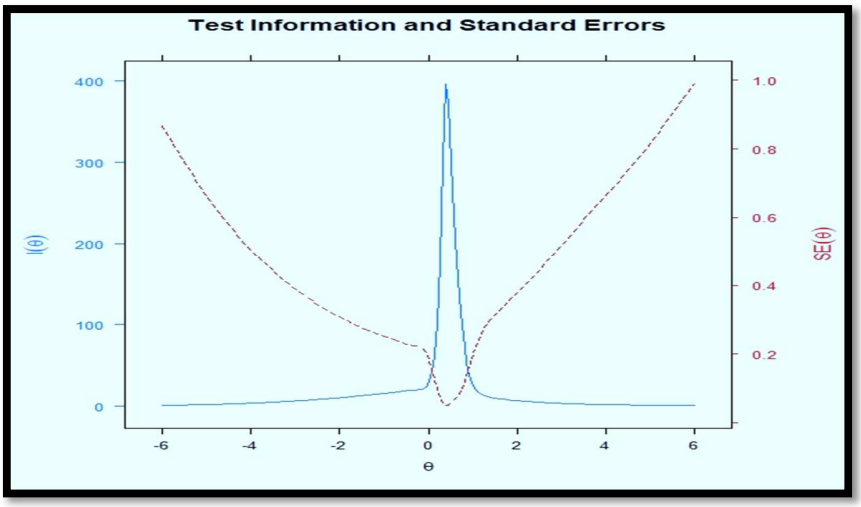
الف) مدل دوارزشی (در هر دو آزمون ۱۳۹۶ و ۱۳۹۸) در مقایسه با مدل چندارزشی آشیانه‌ای سو و بولت، در دامنه گسترده‌تری از مقیاس تتا دارای آگاهی‌دهندگی است. در نتیجه می‌توان گفت مدل دوارزشی نسبت به مدل چندارزشی، دقت بیشتری را در سطوح وسیع‌تری از مقیاس توانایی ایجاد می‌کند.

ب) مدل چندارزشی مورد مطالعه، در دامنه محدودتری از مقیاس تتا (میانه و پایین توزیع)، میزان دقت بالا و به تبع خطای اندازه‌گیری پایینی داشته و در سایر سطوح تتا، میزان آگاهی‌دهندگی پایین‌تری دارد.

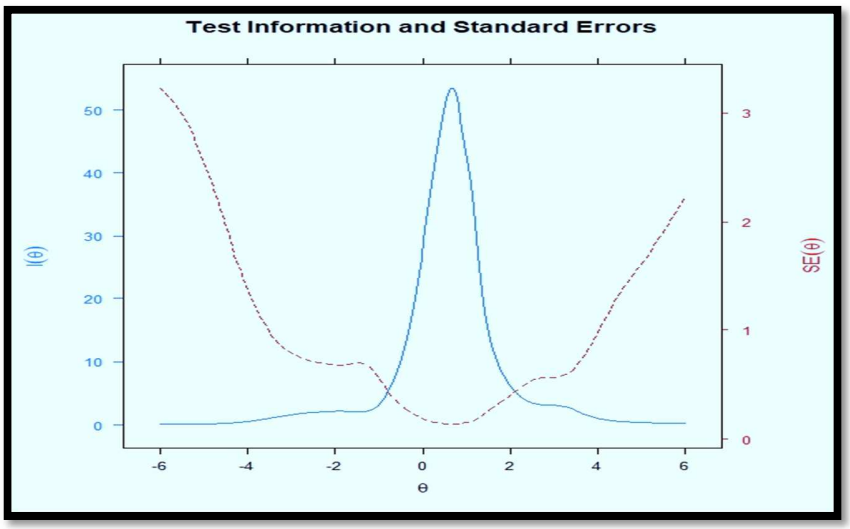
شکل شماره ۱. مقدار آگاهی و خطای آزمون ۱۳۹۶ (مدل ۳ پارامتری دوارزشی)



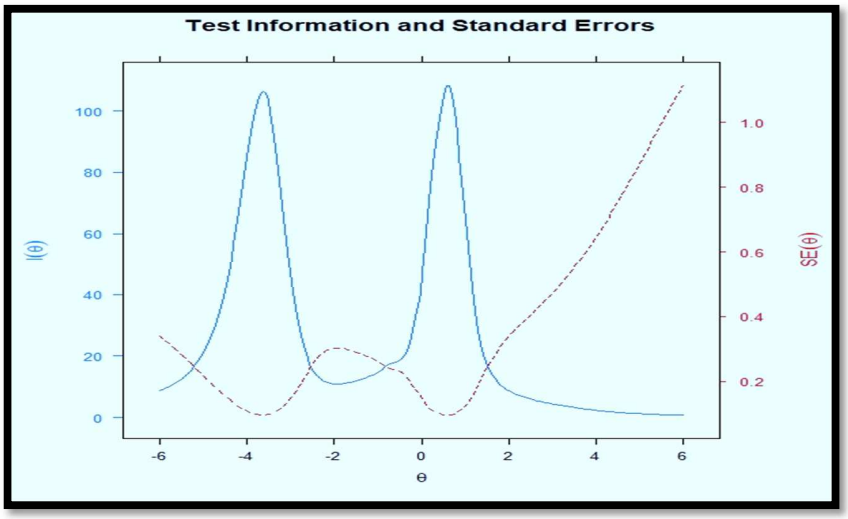
شکل شماره ۲. مقدار آگاهی و خطای آزمون ۱۳۹۶ (مدل ۳ پارامتری چندارزشی)



شکل شماره ۳ . مقدار آگاهی و خطای آزمون ۱۳۹۸ (مدل ۳ پارامتری دوارزشی)



شکل شماره ٤. مقدار آگاهی و خطای آزمون ۱۳۹۸ (مدل ۳ پارامتری چندارزشی)



بحث

در برنامه‌های سنجش گسترده مقیاس همانند آزمون (کنکور) سراسری[1] یا آزمون پروانه وکالت، سوال‌های چندگزینه‌ای همواره گونه رایج سوالات محسوب می شوند و "دقت اندازه‌گیری مربوط به تصمیمات *قبولی- مردودی* افراد در این نوع سنجش‌ها از اهمیت بسیاری برخوردار است" (لانر[2]، شوبر[3]، لاروالد[4] و همکاران، ۲۰۲۰، ص۲۲۰). در این راستا، مطالعه و مدل‌سازی پاسخ سوالات این سنجش‌ها در سال‌های اخیر از تحلیل‌های مبتنی بر نظریه کلاسیک آزمون فاصله گرفته و بر نظریه پرسش پاسخ متمرکز شده است (کارلسون[5] و وان‌داویر[6]، ۲۰۱۸).

1. External test (exam)
2. Lahner
3. Schauber
4. Lorwald
5. Carlson
6. Von Davier

بررسی ادبیات پژوهشی ازجمله مطالعات استورم و همکاران (۲۰۱۹) به‌طور کلی نشانگر آن است که مدل‌های چند ارزشی اسمی IRT جایگزین مناسبی برای مدل چندارزشی NRM و مدل‌های دوارزشی IRT هستند. همچنین، مطالعات دیگر از جمله میسکو سکی و استورم (۲۰۱۸) نشان داد، از میان مدل‌های لجستیک دوارزشی، مدل سه و چهار پارامتری، و از میان مدل‌های چندارزشی اسمی مورد مطالعه نیز، مدل سه و چهار پارامتری در مقایسه با مدل دو پارامتری، بهترین برازش را بر روی داده‌های چندارزشی اسمی دارد و این مدل‌ها، اطلاعات بیشتری از گزینه‌های انحرافی بویژه از سطوح پایین مقیاس توانایی به دست می‌دهند. همچنین، استورم و همکاران (۲۰۱۹) در یک بررسی بر روی داده‌های حاصل از ۲۹۴۹ شرکت‌کننده یک آزمون که شامل سوالات چندگزینه‌ای توانایی عمومی شناختی بود، به مقایسه عملکرد مدل‌های لجستیک دو ارزشی IRT (1PL-P2L-3PL-4PL)، با مدل‌های چند ارزشی اسمی شامل مدل پاسخ اسمی باک(NRM) و مدل‌های چند ارزشی سو و بولت (NLM) که شامل مدل‌های 4PNL, 3PNL و 2PNL می‌شود، مبادرت کردند. یافته‌های این پژوهش نشان داد مدل‌های NLM جایگزین مناسبی برای مدل NRM و مدل‌های دو ارزشی IRT هستند.

نتیجه‌گیری

بی تردید، دقت و مدل نمره‌گذاری آزمون‌های توانایی چندگزینه‌ای در فرآیند انتخاب متقاضیان آزمون‌های انتخابی سرنوشت‌ساز و گسترده مقیاس، سهم تاثیرگذاری دارد. این موضوع در برنامه‌های سنجش آموزش[1] اهمیت زیادی داشته و با چالش‌هایی روبرو است (لانر، شوبر، لاروالد و همکاران، ۲۰۲۰). از این‌رو، هدف مطالعه حاضر، مقایسه میزان دقت و خطای اندازه‌گیری مدل‌های دو ارزشی و چند ارزشی IRT در نمره‌گذاری بر اساس داده‌های آزمون پروانه وکالت سال ۱۳۹۶ و ۱۳۹۸ کشور بود.

مطلوبیت مدل‌های دو ارزشی و چند ارزشی IRT در داده‌های آزمون سال ۱۳۹۶ و ۱۳۹۸، با استفاده از شاخص برازش مدل و شاخص آگاهی (دقت) بررسی و مطالعه گردید. به‌طور خلاصه، تحلیل برازش مدل‌ها (در سطح معناداری ۰/۰۰۱) نشان داد:

الف) مدل‌های دوارزشی در مقایسه با مدل‌های چند ارزشی (مطابق جدول ۱)، برازش بهتری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارند.

ب) در میان مدل‌های دوارزشی، مدل ۳ پارامتری (با فاصله اندک از مدل ۴ پارامتری) برازش مطلوب‌تری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارد.

ج) همچنین، در میان مدل‌های چندارزشی، مدل ۳ پارامتری (با فاصله اندک از مدل ۴ پارامتری) برازش بهتری با داده‌های هر دو آزمون سال ۱۳۹۶ و ۱۳۹۸ دارد.

علاوه بر این، مطابق نمودارهای شماره ۱ تا ۴، مقایسه دقت نمره‌گذاری مدل‌های دو ارزشی و نیز مدل‌های چندارزشی اسمی آزمون سال ۱۳۹۶ نشان داد دقت مدل۳ پارامتری دوارزشی، بویژه در میانه توزیع توانایی حد فاصل مقادیر $\theta = \pm 2$ ، در دامنه وسیع‌تری در مقایسه با مدل۳ پارامتری چند ارزشی ($\theta = \pm 0/5$) بیشتر بوده و آگاهی دهندگی بالاتری دارد. تحلیل آزمون سال ۱۳۹۸ نیز نشان داد مدل ۳ پارامتری دوارزشی، در میانه توزیع توانایی حد فاصل مقادیر $\theta = \pm 2$ و در مقایسه با مدل ۳ پارامتری چندارزشی، آگاهی دهندگی بیشتری دارد.

در خصوص کاربرد مدل‌های چندارزشی باید به این نکته توجه داشت که پیش‌فرض کاربرد مدل‌های چندارزشی اسمی برای نمره‌گذاری آزمون‌های چندگزینه‌ای توانایی، بر این اساس استوار است که این آزمون‌ها از سوال‌های چندارزشی تشکیل شده باشد. همچنین باید توجه داشت از آنجا که فراوانی پاسخ‌های نادرست اغلب در سؤال‌های دشوارتر بیشتر مشاهده می‌شود و از طرف دیگر، افرادی که از سطح توانایی پایین‌تری برخوردارند تعداد پاسخ نادرست بیشتری انتخاب می‌کنند، لذا انتظار می‌رود که مدل‌های چندارزشی، دقت برآورد تتا را در سؤال‌های دشوارتر و در سطوح پایین‌تر توانایی، بیشتر ارتقاء دهند (تور، منگچنگ و تائو، ۲۰۱۷). بنابراین، دقت و آگاهی‌دهندگی بیشتر مدل چندارزشی نسبت به مدل دوارزشی در صورتی محقق خواهد شد که در گزینه‌های انحرافی سؤالات آزمون، اطلاعاتی متناسب با سطح توانایی پاسخ‌دهندگان موجود باشد در غیر این صورت نمی‌توان در مورد داده‌های تحت مطالعه، دقت بیشتر مدل‌های چندارزشی را انتظار داشت.

تقدیر و تشکر

بدین‌وسیله از همکاری سازمان سنجش آموزش کشور در اجرای پژوهش حاضر سپاسگزاری می‌شود.

1. Educational measurement

References

Abad, F.; Olea, J. & Ponsoda, V. (2009). The Multiple-Choice Model Some Solutions for Estimation of Parameters in the Presence of Omitted Responses**.** *Applied Psychological Measurement*, Vol. 33, No. 3, pp. 200-221.

Baker, F. B. & Ho Kim, S. (2017). The Basics of Item Response Theory Using R. Springer International Publishing.

Brown, A. & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of Item Response Theory Modeling*: *Applications to Typical Performance Assessment*. New York: Routledge/Taylor & Francis Group.

Bock, R. D. (1997). The nominal categories model. *In Handbook of modern item response theory.* New York: Springer.

Bock, R. D. & Gibbons, R D. (2021). Item response theory. John Wiley & Sons Ltd.

Bolt, D.; Wollack, J. & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika,* 77(2), 339–357.

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag.

Burnham, K. P., & Anderson, D. R. (2004). *Multimodel Inference: understanding AIC and BIC in Model Selection*, Amsterdam Workshop on Model Selection.

Carlson, J. E. & Von Davier, M. (2018). Item Response Theory. Available from https://ets.org

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three parameter logistic model in computerized adaptive testing. *Applied measurement in education,* 5, 17-34.

De Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. Guilford Publications, Inc.

Drasgow, F. & Levine, M.V. & Tsien, S. (1995). Fitting Polytomous Item Response Theory Models to Multiple-Choice Tests. *APPLIED PSYCHOLOGICAL MEASUREMENT.* Vol. 19, No. 2.

DeMars, C. (2010). Item response theory. Published by Oxford University Press, Inc.

Kim, Jee-Seon. (2006). Using the Distractor Categories of Multiple-Choice Items to Improve IRT Linking. *Journal of Educational Measurement*, Vol. 43, No. 3, pp. 193–213.

Lacourly, N; Sanmartin, J; Silva, M; & Uribe, P. (2018). IRT Scoring and the principle of consistent order. Available from https://arXiv.org

Lahner, F; Schauber, S; Lorwald, A; Kropf, R; Guttormsen, S; Fischer, M; & Huwendiek, S. (2020). Measurement precision at the cut score in medical multiple-choice exams: Theory matters. *Perspectives on Medical Education,* 9, 220-228.

Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and Polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence*, 68, 109–116.

Paek, I. & Cole, K. (2020). USING R FOR ITEM RESPONSE THEORY MODEL APPLICATIONS. Routledge

Penfield, R, & La Torre, J. (2008). A new response model for multiple-choice items. Paper presented at the 2008 annual meeting of the National Council on Measurement in Education, New York.

Penfield, R. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. *Educational Measurement: Issues and Practice*, Vol. 33, No. 1, pp. 36–48.

Preston, K; Reise, S; Cai, L; & Hays, R. (2011). Using the Nominal Response Model to Evaluate Response Category Discrimination in the PROMIS Emotional Distress Item Pools. *Educational and Psychological Measurement,* 7(3), 523-550.

Price, L. R. (2017). Psychometric Methods. Guilford Publications, Inc.

Reif, M. (2014). IRT models for multiple-choice items (mcIRT).
Available from https://github.com/manuelreif/mcIRT

Ritt, M. (2016). The impact of high-stakes testing on the learning environment. Master of social work clinical research papers. Paper 658.

Revelle, W. (2017). Psych: Procedures for Personality and Psychological Research. Available from https://personality-project.org/r/psych.

Samejima, F. (1996). Polychotomous responses and the test score. Available from https://eric.ed.gov

Simon, M; Ercikan, K; & Rousseau, M. (2013). Improving Large-Scale Assessment in education. New York: Routledge.

Storme, M; Myszkowski, N; Baron, S; & Bernard, D. (2019). Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Intelligence*, 7(3), 1-17.

Suh, Y., & Bolt, D. (2010). Nested logit models for multiple-choice item response data. *Psychometrika,* 75(3), 454–473.

Suh, Y. & Bolt, D. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement, 48*, 188-205.

Thompson, N. (2021). Classical Test Theory vs. Item Response Theory: What are some key differences, and how to choose? Available from https://assess.com

Tour, L; Mengcheng, W; & Tao, X. (2017). An investigation of enhancement of ability evaluation by using a nested logit model for multiple-choice items. *Annals of psychology,* 33(3), 530-537.

Van der Linden, W. J. (2016). Handbook of Item Response Theory. Taylor & Francis Group, LLC.