Quarterly Journal of Educational Psychology
Vol. 17, No.60, Summer 2021
Jep.atu.ac.ir
DOI: 10.22054/JEP.2022.63068.3450

ATU PRESS

Original Research
Accepted: 2021/10/13
Received: 2021/08/25
eISSN: 2538-3191
ISSN: 2538-3183

The effect of geographical clustering on parameter estimation and ranking of individuals in large scale assessments

Jalil Younesi *

Associate Professor, Psychology Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran

Seyed Amin Mousavi

Associate Professor, Psychometrics, Classroom Assessment, and Measurement, University of Saskatchewan, Saskatoon, Canada

Abstract

The aim of this study was to investigate the effect of geographical clustering on the estimation of ability parameter (before ranking) as well as the ranking of examinees in large-scale tests (such as national university entrance examination). The design of this study can be considered as survey and because of analyzing the data previously collected by the National Organization Educational Testing (NOET), this project can also be considered as a secondary data analysis. The statistical population of this project includes all the candidates of the Mathematics group who have participated in the national exam of 2013-2014. The sample used in this study includes data on 3,000 examinees in the mathematics group from regions (quota) one, two and three, that have been provided to researchers by the NOET. The main data used in this study are the examinees' scored responses to test items (1 for a correct answer and 0 for an incorrect answer) and the code related to the selected quota. In summary, it can be concluded that differences between the three regions, or any type of clustering, is considerable when the level of analysis is the overall sample and beyond clusters. In such a situation, it is possible to observe a significant difference between rankings using different methods, even when the value of intra-class correlation (ICC) is very low (like this study).

Keywords: Quota, Ability Parameter Estimation, Clustering, Large Scale Assessment..

* Corresponding Author: younesi@atu.ac.ir

How to Cite: Younesi, J, Mousavi, S. A. (2021). The effect of geographical clustering on parameter estimation and ranking of individuals in large scale assessments, *Journal of Educational Psychology*, 17(60), 1-40.

مقاله پژوهشی

تاریخ ارسال: ۰۰/۰۶/۰۳

تاریخ پذیرش: ۰۰/۰۷/۲۱

ISSN: 2538-3183

eISSN: 2538-3191

فصلنامه روان شناسی تربیتی

سال هفدهم، شماره ۶۰، تابستان ۱۴۰۰، ص ۱ تا ۴۰

jep.atu.ac.ir

DOI: 10.22054/JEP.2022.63068.3450

تأثیر منطقه بندی بر برآورد پارامتر توانایی و رتبه‌بندی آزمودنی‌ها در آزمون‌های بزرگ‌مقیاس

جلیل یونسی * دانشیار، سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران

سیدامین موسوی دانشیار، روان‌سنجی و سنجش و اندازه‌گیری کلاسی، دانشگاه ساسکاچوان، ساسکاتون، کانادا

چکیده

این پژوهش با هدف بررسی تأثیر ادغام سهمیه‌بندی در برآورد توانایی (پیش از رتبه‌بندی) و نیز رتبه‌بندی آزمودنی‌ها در آزمون‌های بزرگ مقیاس (مثل کنکور سراسری) انجام شد. طرح این پژوهش در زمرهٔ طرح‌های پیمایشی است. به دلیل این‌که هدف این پژوهش تحلیل داده‌هایی است که قبلاً طی یک طرح پیشین توسط سازمان سنجش جمع‌آوری شده‌اند، این طرح را می‌توان در زمرهٔ تحلیل ثانویه داده‌ها نیز در نظر گرفت. جامعه آماری این طرح شامل تمام داوطلبان گروه ریاضی است که در آزمون سراسری سال ۱۳۹۴-۱۳۹۳ شرکت نموده‌اند. نمونه مورداستفاده در این پژوهش شامل داده‌های مربوط به ۳۰۰۰ آزمودنی گروه ریاضی از مناطق (سهمیه) یک، دو و سه است که از طرف سازمان سنجش در اختیار پژوهشگران قرار گرفته است. داده‌های اصلی مورداستفاده در این پژوهش عبارت است از پاسخ‌های نمره‌گذاری شده آزمودنی‌ها به سؤالات آزمون (به‌صورت ۱ برای پاسخ صحیح و ۰ برای پاسخ غلط) و کد مربوط به سهمیه انتخابی. به‌طور خلاصه، می‌توان چنین نتیجه‌گیری کرد که تفاوت بین مناطق سه‌گانه و یا هر نوع خوشه‌بندی در داده‌ها هنگامی در نتایج تأثیر دارد که تحلیل در سطح کل نمونه و ورای خوشه‌ها صورت می‌گیرد. در چنین وضعیتی، امکان مشاهده تفاوت معنادار بین رتبه‌دهی با استفاده از روش‌های متفاوت، حتی زمانی که مقدار ICC بسیار پایین باشد (مانند تحلیل حاضر)، وجود دارد.

کلیدواژه‌ها: آزمون بزرگ مقیاس، پارامتر توانایی، سهمیه‌بندی، منطقه‌بندی.

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

* نویسنده مسئول: younesi@atu.ac.ir

مقدمه

محققان بر این نکته توافق دارند که اساس موفقیت در جامعه شایسته‌سالار، نظام آموزشی است (کرخوف[1]، ۱۹۹۵؛ گنزبوم، تریمن و اولتی[2]، ۱۹۹۱) و برقراری نظام آموزشی مطلوب و کارا و آموزش مثمر ثمر فرایندی دشوار و پرزحمت است. هر کشوری از رویکردهای متفاوتی برای آموزش استفاده می‌کند ولی اینکه آموزش، مؤثر و موفق باشد مستلزم برنامه‌ریزی صحیح و تلاش بسیار، برنامه درسی روزآمد و مهم‌تر از آن فراهم کردن فرصت‌های برابر آموزشی برای یادگیری منابع آموزشی است (شارع‌پور، ۱۳۸۶؛ فلدمن[3]، ۲۰۰۱؛ اشمیت و مایر[4]، ۲۰۰۹؛ اسلامی هرندی، کریمی و نادی، ۱۳۹۸)؛ اما در مورد اینکه چگونه می‌توان موفقیت آموزشی را تقویت، عدالت آموزشی را برقرار و تصمیم‌های عادلانه‌تری اتخاذ کرد اختلاف‌نظر زیادی وجود دارد. آمارهای منتشرشده حاکی از آن است که در نظام آموزشی ایران علیرغم تمام تلاش‌های صورت گرفته در نظام آموزش‌وپرورش تا دستیابی به عدالت آموزشی فاصله زیادی وجود دارد و این عدم برابری آموزشی می‌تواند آینده تحصیلی و شغلی دانش‌آموزان را در ورود به دانشگاه‌ها، محیط شغلی آینده و زندگی اجتماعی دستخوش تغییرات جدی کند (یونسی، دلاور، اسکندری، فلسفی‌نژاد و فرخی، ۱۳۹۳).

ازجمله سنجش‌های بزرگ مقیاس در ایران، آزمون یا کنکور سراسری است که مهم‌ترین آزمون سرنوشت‌ساز ایران نیز محسوب می‌شود. در بحث ضوابط پذیرش دانشجو ایده‌آل‌ترین سیستم گزینش علمی این است که شما هیچ قیدوبندی نداشته باشید یعنی، قیودی مانند بومی، سهمیه و نظایر آن از میان برداشته شود و ارزیابی کاملاً بر اساس نمره به‌دست‌آمده داوطلب باشد (خدایی، ۱۳۹۳).

به دلیل شرایط نابرابر تحصیلی در نقاط مختلف کشور و توزیع ناعادلانه امکانات

1. Kerckhoff
2. Ganzeboom, Treiman And Ultee
3. Feldman
4. Schmidt & Maier

تحصیلی و شرایط فراگیری یکسان و برابر برای تمامی داوطلبان و به‌منظور ایجاد شرایط رقابت یکسان برای آنان، با اعمال سهمیه‌بندی تلاش شده تا حدودی ضعف‌های موجود جبران شود. بر اساس مصوبات مراجع ذی‌صلاح، هر یک از داوطلبان در آزمون سراسری دارای سهمیه مشخصی هستند. از سال ۱۳۶۸ به بعد، مبنای سهمیه‌بندی مناطق، تقسیم‌بندی وزارت کشور بوده که بر اساس شاخص‌های آموزشی، اقتصادی، فرهنگی و اجتماعی، کشور به مناطق یک، دو و سه طبقه‌بندی‌شده و سهمیه خاصی برای آن‌ها منظور شده است. در حال حاضر با توجه به مصوبه مجلس شورای اسلامی این سهمیه شامل سهمیه مناطق (منطقه یک، دو و سه)، خانواده شهدا، ایثارگران و رزمندگان است.

نمره کل افراد با توجه به زیرگروه انتخابی توسط آزمودنی و تعیین ضرایب دروس، میانگین وزنی نمرات تراز به‌دست‌آمده محاسبه می‌شود (پیک سنجش، ۱۳۹۳ الف، ۱۳۹۳ ب). سپس نمره کل به‌دست‌آمده جهت تعیین رتبه آزمودنی در سهمیه و بدون سهمیه مورداستفاده قرار می‌گیرد. در فرآیند مورداشاره هیچ کجا تفاوت‌های درون و بین سهمیه‌ای در نظر گرفته نمی‌شود. این امر می‌تواند تأثیر بسزایی در سرنوشت آزمودنی داشته باشد.

دراین‌بین آنچه قابل تأکید است آن است که سنجش‌هایی که در مقیاس وسیع انجام می‌شوند ازنظر اهداف، حوزه‌های محتوایی، سؤالات آزمون و جامعه دانش‌آموزان متفاوت هستند ولی ویژگی‌های معینی وجود دارد که برای تمامی این مطالعات ضروری و اساسی هستند. مهم‌ترین و بنیادی‌ترین ویژگی روان‌سنجی هر آزمونی، «روایی» است. مهم‌ترین عاملی که در ارزشیابی هر آزمون و تک‌تک سؤالات آزمون باید بدان توجه داشت روایی نمرات حاصل از اجرای سؤالات است (تیلور، ۲۰۱۳، ترجمه یونسی، ۱۳۹۸). مقصود از روایی، مناسب بودن[1]، با معنا بودن[2] و مفید بودن[3] استنباط‌های خاصی است که از روی نمره‌های حاصل از اجرای آزمون می‌توان به عمل آورد. در تفسیر یا استفاده صحیح و دقیق

1. Appropriateness
2. Meaningfulness
3. Usefulness

از نمرات آزمون بایستی شواهد متعددی گردآوری شود تا بتوان مطمئن شد که نمرهٔ حاصل دارای روایی است (مسیک[1]، ۱۹۸۹، در لین[2]، ۱۹۸۹). یکی از عواملی که می‌تواند در روایی نمرات برآورد شده تأثیر داشته باشد، ساختار داده‌های گردآوری‌شده و انتخاب مدل آماری مناسب برای تحلیل داده‌هاست. کاربرد نامناسب روش‌های آماری برای تحلیل این داده‌ها مشکلی معمول در تحقیقات کاربردی است (راتکوسکی، گنزالن، جونکاس و ون داویر[3]، ۲۰۱۰). در سنجش‌های وسیع معمولاً پاسخ‌دهندگان خوشه‌بندی می‌شوند و جامعه موردنظر شامل چند زیرجامعه است. مشاهدات درون هر خوشه با هم همبستگی دارند و گویای این هستند که خوشه‌ها از برخی جهات با هم تفاوت دارند. ازاین‌رو، داده‌های حاصل از سنجش‌های وسیع جمعیتی اغلب دارای ساختار سلسله‌مراتبی هستند، بدین نحو که آزمودنی‌ها در گروه‌ها و یا طبقات مختلف آشیان می‌شوند. در برخی موارد این ساختار سلسله مراتبی به‌صورت طبیعی وجود دارد. مثلاً در بافت نظام آموزشی مدارس، دانش‌آموزان در کلاس‌ها آشیان شده، کلاس‌ها در مدارس آشیان شده و مدارس نیز در ناحیه‌های شهری آشیان شده‌اند. گاه ساختار سلسله مراتبی داده‌ها منتج از تصمیمات و سیاستگذاری‌هاست. مثلاً در آزمون سراسری دانشگاه‌ها، کلیه استان‌ها به مناطق چندگانه تقسیم‌بندی شده‌اند. درنظر نگرفتن ساختار سلسله مراتبی داده‌ها می‌تواند منجر به برآوردهای اریب[4] شده و روایی آماری نتایج تحلیل را خدشه‌دار نماید. یکی از راه‌های افزایش روایی نتایج آماری، تجزیه‌وتحلیل‌های درست و وارد کردن سطوح در تحلیل‌ها است (نه اینکه صرفاً به ادغام سطوح بپردازیم و آن‌ها را یکی کنیم)، به عبارتی تحلیل‌های چندسطحی یکی از راه‌های کنترل آماری است و موجب افزایش روایی نتایج آماری و روایی پژوهش می‌شود (مسیک، ۱۹۸۹؛ کین[5]، ۲۰۰۶؛ تیلور، ۲۰۱۳، ترجمه یونسی، ۱۳۹۸).

1. Messick
2. Linn
3. Rutkowski, Gonzalen, Joncas, & Von Davier
4. Bias
5. Kane

این ویژگی‌ها نشان‌دهندهٔ چالش‌های سنجش‌های بزرگ مقیاس هستند. اول اینکه، سنجش‌های وسیع روی نمونهٔ وسیعی از دانش‌آموزان از مدارس و ناحیه‌های متفاوت و گاهی اوقات استان‌ها یا مناطق کشور اجرا می‌شود. پاسخ‌های دانش‌آموزان از یک مدرسه (یا یک ناحیه) به خاطر عدم‌همگنی غیرقابل مشاهده[1] در سطح مدرسه از سایر پاسخ‌ها مستقل نیستند (رادنبوش و بریک، ۲۰۰۲). از این نظر، داده‌ها چندسطحی هستند و لذا مدل‌های آماری سنتی با مشکل مواجه خواهند بود. نوعاً مشاهدات پاسخِ درون هر خوشه در مقابل (مشاهدات غیرآشیانه‌ای) از خوشه‌های مختلف، دارای توزیع مستقلی نیستند. این آشیانه‌بندی منجر به ساختارهای وابسته[2] پیچیده با منابع تغییر در سطوح مختلف سلسله مراتبی می‌شود (گلدشتاین، ۱۹۹۹، ۲۰۰۳). دوم اینکه، متغیرهای اصلی موردنظر، توانایی‌های دانش‌آموزان در حوزه‌های محتوایی متفاوت به‌طور مستقیم قابل مشاهده نیستند. این متغیرها از طریق تعدادی از سؤالات آزمون که به‌صورت دوارزشی یا چند ارزشی نمره‌گذاری می‌شود، به‌صورت غیرمستقیم اندازه‌گیری می‌شود. ناهمگنی آزمودنی‌ها یک منبع معمول تغییر در داده‌های پاسخ است که باید با مدل‌های پاسخ آماری تبیین شود. معمولاً تفاوت‌های آزمودنی‌ها را می‌توان از طریق یک توزیع احتمالاتی[3] شناخته‌شده به‌عنوان یک توزیع جامعه آزمودنی‌ها مدل‌سازی کرد و شواهد درباره پاسخ‌دهندگان همیشه با توجه به توزیع جامعه[4] اتخاذ می‌شود. در نظریه سؤال- پاسخ (IRT)، سطح توانایی فرد با توجه به پاسخ‌های وی به سؤال‌های آزمون برآورد می‌شود. مدل سؤال- پاسخ مشخص می‌کند که سطح توانایی و ویژگی‌های سؤال چگونه به یکدیگر مرتبط می‌شوند. سطح توانایی فرد در بافت یا زمینه مدل برآورد می‌شود و بر همین اساس، IRT، اندازه‌گیری بر پایه مدل[5] است (لرد، ۱۹۸۰، ترجمه دلاور و یونسی، ۱۳۹۱). به‌طورمعمول، منطق نظریه اندازه‌گیری شامل یک مدل رفتاری است (امبریتسون و

1. Unobserved Heterogeneity
2. Dependency Structures
3. Probability Distribution
4. Population Distribution
5. Model- Based

رایس، ۲۰۰۰). در نظریه سؤال- پاسخ (IRT) باور بر این است که اگر چگونگی ارتباط هر سؤال آزمون با فرد موردمطالعه درک شود، می‌توان نمرهٔ واقعی یا صفت واقعی موردنظر را به‌صورت مستقیم برآورد کرد (برنان، ۲۰۰۶؛ نانالی[1] و برنشتاین[2]، ۱۹۹۴؛ کلاین[3]، ۱۹۹۸؛ وندرلیندن[4] و همبلتون، ۱۹۹۷). به‌طورکلی، در چارچوب متغیرهای مکنون، مدل‌های ساختاری متفاوتی وجود دارد که این مدل‌ها بر اساس سطح تحلیل (فردی یا سلسله مراتبی)، نوع داده‌ها (طبقه‌ای یا پیوسته) و شیوهٔ جمع‌آوری داده‌ها (آشیانه‌ای یا غیر آشیانه‌ای) متفاوت‌اند؛ بنابراین، در برآورد توانایی آزمودنی باید هر دو موضوع ساختار سلسله مراتبی داده‌ها و مکنون بودن متغیر توانایی مدنظر قرار بگیرد. ترکیب یک مدل سؤال- پاسخ برای اندازه‌گیری توانایی‌های افراد با یک مدل چندسطحی ساختاری که تفاوت‌ها را در سطوح مختلف توانایی تبیین می‌کند با استفاده از مدل‌های سؤال-پاسخ سلسله‌مراتبی[5] امکان‌پذیر است (موسوی، ۲۰۱۳). تکنیک‌های مدل‌سازی چندسطحی زمانی که برای تحلیل‌های روان‌سنجی استفاده می‌شود، « مدل‌سازی اندازه‌گیری چندسطحی[6]» (برتواس و کاماتا[7]، ۲۰۰۵؛ کاماتا، باوئر و میازاکی[8]، ۲۰۰۸) نامیده می‌شود. زمانی که در مدل‌سازی اندازه‌گیری چندسطحی (MMM) با نشانگرهای اندازه‌گیری طبقه‌ای روبرو باشیم، مثل سؤالاتی که به‌صورت دوارزشی یا چندارزشی نمره‌گذاری می‌شوند، به این نوع مدل‌سازی، « مدل‌سازی نظریه سؤال- پاسخ چندسطحی[9]» گفته‌می‌شود. معمولاً در مدل‌های سنتی IRT به ساختار آشیانه‌ای داده‌ها، مثلاً دانش‌آموزان در درون مدارس آشیانه می‌گیرند، توجهی نمی‌شود. در این چارچوب، مدل اندازه‌گیری رابطهٔ بین توانایی و داده‌های حاصل از پاسخ آزمودنی را مدنظر قرار می‌دهد و مدل

1. Nunnally
2. Bernstein
3. Kline
4. Van Der Linden
5. Multilevel Item Response Theory (MLIRT)
6. Multilevel Measurement Modeling (MMM)
7. Beretvas, & Kamata
8. Kamata, Bauer, & Miyazaki
9. Multilevel Item- Response Theory Modeling (MIRT)

چندسطحی ساختاری هم ساختار سلسله مراتبی افراد در جامعه موردپژوهش را در تحلیل وارد می‌کند (فاکس[1]، ۲۰۱۰؛ یونسی و همکاران، ۱۳۹۳).

در حال حاضر، توانایی یا نمره کل داوطلبان آزمون سراسری بدون توجه به دو مورد ذکرشده در بالا یعنی ساختار سلسله مراتبی داده‌ها و مکنون بودن متغیر توانایی صورت می‌گیرد. روند کلی بدین‌صورت است که ابتدا نمرهٔ کل آزمودنی برآورد شده، سپس وضعیت طبقه‌بندی (منطقه، سهمیه،...) در نظر گرفته می‌شود. ازاین‌رو، مسئله اصلی پژوهش حاضر این بوده که با عنایت به اینکه سازمان سنجش آموزش کشور بر اساس قانون مصوب مجلس شورای اسلامی مکلف است سهمیه‌های مورداشاره را قبل از اعلام رتبه‌های دانش‌آموزان در کنکور سراسری اعمال کند و سهمیه‌ها نقش چشمگیری در رتبه‌های کنکور دارند و در برخی از رشته‌های پرطرفدار و محبوب جامعه در زیرگروه‌های مختلف (مثلاً رشته‌های پزشکی، دندانپزشکی، مهندسی برق و مکانیک و نظایر این‌ها) نقش حساس‌تری را ایفا می‌کند، در نظر گرفتن و یا نگرفتن ساختار سلسله مراتبی داده‌ها و نیز تفاوت‌های درونی ساختار سلسله مراتبی داده‌ها، بر برآورد توانایی و رتبه‌بندی آزمودنی‌ها در آزمون‌های بزرگ مقیاس (نظیر کنکور سراسری) چه تأثیری دارد.

روش پژوهش

این مطالعه در دو فاز انجام شد. در فاز نخست، با استفاده از داده‌های واقعی و در فاز دوم که شبیه‌سازی صورت گرفت به کمک داده‌های تولیدشده سؤالات پژوهش موردبررسی قرار گرفتند. با توجه به اینکه در این پژوهش محقق در پی مقایسه رویکردهای مختلف در تحلیل داده‌های با ساختار آشیانه‌ای و سلسله مراتبی، بررسی برازش مدل‌های مختلف اندازه‌گیری، به‌دست آوردن برآوردهای باثبات‌تر پارامترها (اعم از پارامترهای سؤال و توانایی آزمودنی‌ها) و افزایش اعتبار تصمیم‌ها و دقت اندازه‌گیری بوده است، لذا روش این پژوهش به‌طور عام جزو پژوهش‌های توصیفی است. ازاین‌رو، طرح این پژوهش را می‌توان

1. Fox

در زمرهٔ طرح‌های پیمایشی در نظر گرفت، چراکه منبع و تمرکز اصلی طرح بر یک پیمایش آموزشی بزرگ در ایران است. در چنین طرح‌هایی داده‌های وسیعی که از افراد مختلف به‌وسیلهٔ آزمون گردآوری‌شده، مورد تجزیه‌وتحلیل قرار می‌گیرد. از طرف دیگر، به دلیل این‌که هدف این پژوهش تحلیل داده‌هایی است که قبلاً طی یک طرح پیشین جمع‌آوری شده‌اند، این طرح را می‌توان در زمرهٔ تحلیل ثانویه داده‌ها[1] نیز در نظر گرفت (وارتانیان[2]، ۲۰۱۰).

جامعه آماری و گروه نمونه

جامعه آماری این پژوهش شامل تمام داوطلبان گروه ریاضی است که در آزمون سراسری سال ۱۳۹۴-۱۳۹۳ شرکت کرده‌اند. داده‌های موردنظر قبلاً توسط سازمان سنجش آموزش کشور جمع‌آوری شده است. نمونه مورداستفاده در این پژوهش شامل داده‌های مربوط به ۳۰۰۰ آزمودنی گروه ریاضی از مناطق (سهمیه) یک، دو و سه است که از طرف سازمان سنجش در اختیار پژوهشگران قرار گرفته است.

ابزار گردآوری داده و شیوهٔ اجرا

ابزار گردآوری داده در این پژوهش، آزمون سراسری سال ۱۳۹۴-۱۳۹۳ سازمان سنجش آموزش کشور بوده که روایی و پایایی آن توسط کارشناسان سازمان موردبررسی و تأیید قرار گرفته است. این آزمون از دو بخش سؤالات عمومی و تخصصی تشکیل گردیده است. آزمون عمومی دربرگیرندهٔ دروس زبان و ادبیات فارسی، زبان عربی، فرهنگ و معارف اسلامی و نیز زبان انگلیسی مشتمل بر ۲۵ سؤال برای هر درس (۱۰۰ سؤال درمجموع) می‌باشد. آزمون تخصصی گروه ریاضی شامل دروس ریاضیات (با ۵۵ سؤال)، فیزیک (با ۴۵ سؤال) و شیمی (با ۳۵ سؤال) با مجموع ۱۳۵ سؤال می‌باشد. داده‌های اصلی مورداستفاده در این پژوهش عبارت است از پاسخ‌های نمره‌گذاری شده (به‌صورت ۱ برای

1. Secondary Data Analysis
2. Vartanian

پاسخ صحیح و ۰ برای پاسخ غلط) آزمودنی‌ها به سؤالات آزمون و کد مربوط به سهمیه انتخابی. علاوه بر این داده‌های مربوط به جنسیت، معدل دیپلم و سال تولد آزمودنی‌ها به همراه نمرات خام و تراز شده پیشرفت تحصیلی، کنکور، نمره کل و رتبه با/بدون سهمیه نیز در فایل ارسالی از طرف سازمان سنجش قرار داشت.

شیوه تحلیل داده‌ها

به‌منظور پاسخ به سؤالات پژوهش علاوه بر بررسی مفروضه‌های اساسی IRT و مفروضه‌های اختصاصی داده‌های سلسله مراتبی، از تکنیک‌های تحلیل سلسله‌مراتبی خطی و نیز مدل‌های سؤال-پاسخ سلسله مراتبی[1] استفاده شد. محاسبات و تحلیل‌ها به‌وسیله برنامه‌نویسی در نرم‌افزار آماری R (با استفاده از بسته‌های نرم‌افزاری lme4, mirt و mlirt) صورت پذیرفت. تحلیل داده‌های آزمون سراسری و سؤالات آن طبق IRT، از طریق نرم‌افزارهای BILOG-MG3 (زیموسکی، موراکی، میسلوی و باک[2]، ۱۹۹۶) و Mplus (موتن و موتن[3] ۲۰۱۲) انجام شد. به‌منظور پاسخ به سؤالات تحقیق، ویژگی‌های توزیع برآورد توانایی، خطای استاندارد برآورد پارامتر و تفاوت رتبه‌ها در سهمیه و بدون سهمیه در دو حالت ذکرشده بررسی شده است.

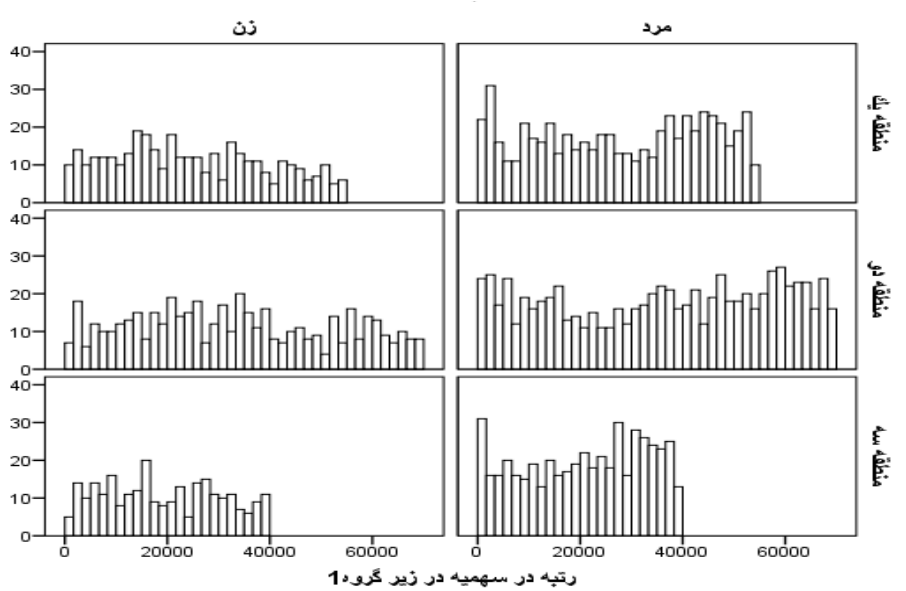
یافته‌های فاز اول پژوهش

الف) نتایج توصیفی

نتایج توصیفی نشان می‌دهد که اکثریت داوطلبان مرد (۶۲٫۴٪) و از منطقه دو (۴۲٫۸٪) بودند. شایان ذکر است که داده‌های موجود حاوی داوطلبانی از سهمیه خانواده شهدا (۱ نفر)، رزمندگان (۱ نفر) و ایثارگران (۵۹ نفر) بود که به دلیل پایین بودن حجم نمونه از تحلیل‌ها حذف شدند و تمرکز بر روی سهمیه مناطق سه‌گانه قرار گرفت. با حذف این

1. Multilevel Item Response Theory (MLIRT)
2. Zimowski, Muraki, Mislevy And Bock
3. Muthen & Muthen

افراد و داوطلبانی که به هیچ سؤالی پاسخ نداده بودند، حجم نمونه نهایی ۲۹۲۱ به‌دست آمد.



شکل ۱. توزیع رتبه در سهمیه

شکل ۱ تصویر خوبی از توزیع رتبه در سهمیه براساس جنسیت و سهمیه نهایی ارائه می‌دهد. در منطقه سه، حداکثر رتبه برای هر دو جنس کمتر از ۴۰۰۰۰ است، درحالی‌که برای منطقه دو حدود ۶۹۰۰۰ و برای منطقه ۱ حدود ۵۴۰۰۰ است. این یافته می‌تواند گویای همگنی بیشتر در بین داوطلبین منطقه سه نسبت به دو منطقه دیگر باشد.

ب) تحلیل آزمون از طریق نظریه کلاسیک اندازه‌گیری

داده‌های مربوط به ۲۳۵ سؤال آزمون از دو بخش سؤالات عمومی و تحلیل شد. جدول ۱ حاوی خلاصه آماره‌های کلاسیک سؤالات آزمون است.

جدول ۱. خلاصه نتایج تحلیل کلاسیک سؤالات آزمون

	درصد پاسخ صحیح	درصد پاسخ صحیح با احتساب بدون پاسخ‌ها	ضریب همبستگی پیرسون	ضریب همبستگی دورشته‌ای نقطه‌ای
حداقل	۱/۲۰	۱۰/۳۰	-۰/۰۲	-۰/۰۳
چارک اول	۶/۴۰	۴۰/۵۰	۰/۳۰	۰/۵۱
میانه	۱۱/۵۰	۵۳/۵۰	۰/۴۱	۰/۶۵
چارک سوم	۲۳/۸۰	۶۳/۱۰	۰/۴۸	۰/۸۰
حداکثر	۶۸/۱۰	۹۰/۹۰	۰/۶۳	۱/۰۵
میانگین	۱۶/۶۰	۵۲/۴۴	۰/۳۹	۰/۶۴
انحراف معیار	۱۴/۱۷	۱۶/۷۷	۰/۱۳	۰/۲۰

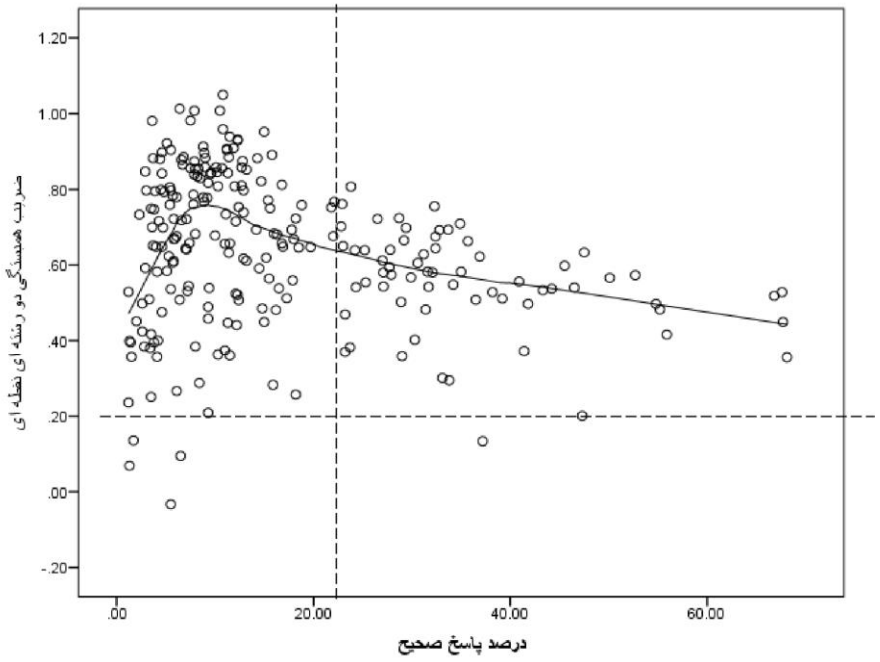
در جدول فوق، درصد پاسخ صحیح به سؤالات به دو شیوه موردمحاسبه قرار گرفته است: براساس کل حجم نمونه و با احتساب بدون پاسخ‌ها. به‌عنوان‌مثال، اگر ۱۰۰۰ نفر به سؤالی پاسخ صحیح داده باشند، درصد پاسخ صحیح برابر خواهد بود با (۱۰۰۰ تقسیم بر ۲۹۲۱) ×۱۰۰ یعنی تقریباً ۳۴٪، اما اگر تعداد ۵۰۰ نفر این سؤال را بدون پاسخ گذاشته باشند، درصد پاسخ صحیح برابر خواهد بود با (۱۰۰۰ تقسیم بر [۵۰۰-۲۹۲۱])×۱۰۰ یعنی تقریباً ۴۱٪. نرم‌افزارهای سنجش و اندازه‌گیری (مثل BILOG-MG) مقدار اول را گزارش می‌کنند اما نرم‌افزارهای آماری (مانند SPSS) عموماً مقدار دوم را به‌عنوان مقدار معتبر گزارش می‌کنند. بر این اساس، تفاوت نسبتاً چشمگیر بین ستون اول و دوم جدول ۳ حاکی از تعداد بسیار بالای داده‌های بدون پاسخ است. با توجه به ستون اول جدول فوق، ۵۰٪ سؤالات و کمتر توسط حدوداً ۱۲٪ کل داوطلبان به‌درستی پاسخ شده‌اند. این مقدار با احتساب داده‌های بدون پاسخ برابر با حدوداً ۵۴٪ داوطلبان است. متوسط دشواری سؤال برای نمونه حدود ۱۷٪ و با احتساب بدون پاسخ‌ها حدود ۵۳٪ است.

ضریب همبستگی دورشته‌ای نقطه‌ای شاخص ضریب تمیز سؤالات است و هرچه مقدار آن بیشتر باشد مطلوب‌تر است. حداقل مقدار قابل‌قبول برای ضریب تمیز، ۰/۲ است (مگنوسون، ۱۹۶۶). براساس مقادیر جدول ۱، ۵۰٪ سؤالات و کمتر دارای ضریب تمیز ۰/۶۵ و کمتر هستند. این نتیجه‌ای مطلوب است و نشان‌دهندهٔ کارکرد مناسب سؤالات

است. البته یک سؤال در داده‌ها وجود دارد که دارای ضریب تمیز منفی است (سؤال شماره ۱۲۷، درس ریاضی).

شکل ۲ نمودار پراکندگی بین درصد پاسخ صحیح سؤالات و ضریب همبستگی دورشته‌ای نقطه‌ای را نشان می‌دهد. اکثر سؤالات دارای ضریب تمیز بالای ۰/۲ (خط‌چین افقی) و دشواری بالا هستند (خط‌چین عمودی). سؤالاتی که در یک‌چهارم فوقانی و چپ قرار دارند سؤالاتی مشکل (حداکثر ۲۰٪ داوطلبان قادر به پاسخگویی بودند) و با ضریب تمیز بالا هستند.

کل سؤالات آزمون دارای آلفای کرونباخ برابر با ۰/۸۳۴ هستند که حاکی از پایایی بالای آزمون است. این امر با توجه به تعداد زیاد سؤالات با ضریب تمیز بالا قابل پیش‌بینی بود. درمجموع، فارغ از موضوع شیوهٔ محاسبه درصد پاسخ صحیح، سؤالات و آزمون از ویژگی‌های نسبتاً قابل‌قبولی به لحاظ نظریه کلاسیک اندازه‌گیری برخوردارند. موضوعی که ممکن است نیاز به بازنگری داشته باشد، تعداد سؤالات زیاد با درجه دشواری بالاست.



شکل ۲. رابطه درصد پاسخ صحیح با همبستگی دو رشته‌ای نقطه‌ای

تحلیل آزمون از طریق IRT

- **بررسی بعدّیت**

روش‌های متعددی (نظیر تحلیل عاملی خطی، تحلیل عامل غیرخطی و تحلیل عاملی با اطلاعات کامل) برای بررسی مفروضه تک‌بعدی بودن مجموعه سؤالات آزمون به‌طور کامل پیشنهاد شده است. با توجه به آنکه نتایج پژوهش‌ها گویای آن است که تحلیل عاملی خطی برای سنجش ابعاد داده‌های سؤالات چندگزینه‌ای مناسب نیست و از بین سایر پژوهش‌ها، یافته‌های برخی از پژوهش‌ها نشان داده است که در مطالعات بازیابی[1] ابعاد، NOHARM در مقایسه با TESTFACT بهتر عمل می‌کند، در این پژوهش از روش NOHARM استفاده شد. برنامه NOHARM ریشه دوم میانگین مجذورات باقیمانده‌ها (RMSR) را محاسبه و به‌عنوان شاخصی برای برازش مدل ارائه می‌دهد؛ بنابراین، مقادیر کوچک RMSR حاکی از برازش مدل با داده‌ها است. یک ملاک برای تفسیر RMSR این است که آن را با چهار برابر معکوس ریشه دوم حجم نمونه، (یعنی خطاهای استاندارد پس‌مانده‌ها) مقایسه کرد. شاخص دیگر برای بررسی برازش مدل، شاخص نیکویی برازش تاناکا (۱۹۹۳) است. مک دونالد (۱۹۹۷) پیشنهاد می‌کند مقدار ۰/۹۰ برای این شاخص حاکی از برازش قابل‌قبول و مقدار ۰/۹۵ بیانگر «برازش خوب» مدل با داده‌ها است. اگر شاخص تاناکا برابر یک باشد بیانگر برازش کامل است. تحلیل تعداد ابعاد آزمون در سه حالت تک‌بعدی، دوبعدی (عمومی و تخصصی) و هفت بعدی (تمام دروس آزمون) موردبررسی قرار گرفت. برای حالت هفت بعدی، نرم‌افزار موفق به تکمیل تحلیل نشد (به‌دلیل عدم همگرایی). نتایج این تحلیل در جدول ۲ ارائه شده است.

جدول ۲. نتایج تحلیل اکتشافی بعدّیت سؤالات

تعداد ابعاد	تعداد سؤالات	تعداد آزمودنی	مجموع مجذور مانده‌ها	ریشه میانگین مجذور مانده‌ها	شاخص برازش تاناکا
تک‌بعدی	۲۳۵	۲۹۲۱	۰/۸۰۵۸	۰/۰۰۵۴	۰/۹۵۳۱
دوبعدی			۰/۶۷۵۷	۰/۰۰۴۲	۰/۹۶۸۱
هفت‌بعدی			عدم همگرایی مدل		

1. Recovery

با توجه به شاخص‌های ذکرشده و با در نظر گرفتن ماتریس باقیمانده‌ها، به نظر می‌آید شواهد کافی برای ردّ فرضیه تک‌بعدی بودن آزمون وجود ندارد و بنابراین می‌توان فرض کرد که داده‌ها تک‌بعدی هستند. به‌منظور بررسی برازش داده‌ها با مدل IRT مطلوب، داده‌های آزمون توسط بسته نرم‌افزاری mirt در نرم‌افزار R موردبررسی قرار گرفت. نتایج این تحلیل در جدول ۳ ارائه شده است.

جدول ۳. نتایج تحلیل تأییدی بعدّیت سؤالات

تعداد ابعاد	تعداد سؤالات	تعداد آزمودنی	لگاریتم درست‌نمایی	AIC	BIC
تک‌پارامتری	۲۳۵	۲۹۲۱	-۱۱۲۳۸۱	۲۲۵۲۳۴/۰	۲۲۶۶۴۵/۲
دوپارامتری			-۱۱۰۱۳۵	۲۲۱۲۱۰/۱	۲۲۴۰۲۰/۵
سه‌پارامتری			-۱۱۰۰۵۷,۲	۲۲۱۱۲۴/۴	۲۲۴۰۱۲/۱

مقایسه مقادیر لگاریتم درستنمایی، شاخص AIC و شاخص BIC همگی حاکی از این هستند که مدل‌های دو و سه پارامتری برازش بهتری با داده‌ها دارند. از سویی دیگر، با توجه به تفاوت ناچیز بین شاخص‌های برازش مدل‌های دو و سه پارامتری می‌توان مدل دوپارامتری را به‌عنوان مدل تحلیل برگزید. مزیت مدل دوپارامتری به مدل سه‌پارامتری، ثابت فرض کردن پارامتر حدس است که فرآیند برآورد پارامترها را سهل‌تر و منجر به برآوردهای باثبات‌تری می‌شود (فاکس، ۲۰۰۱).

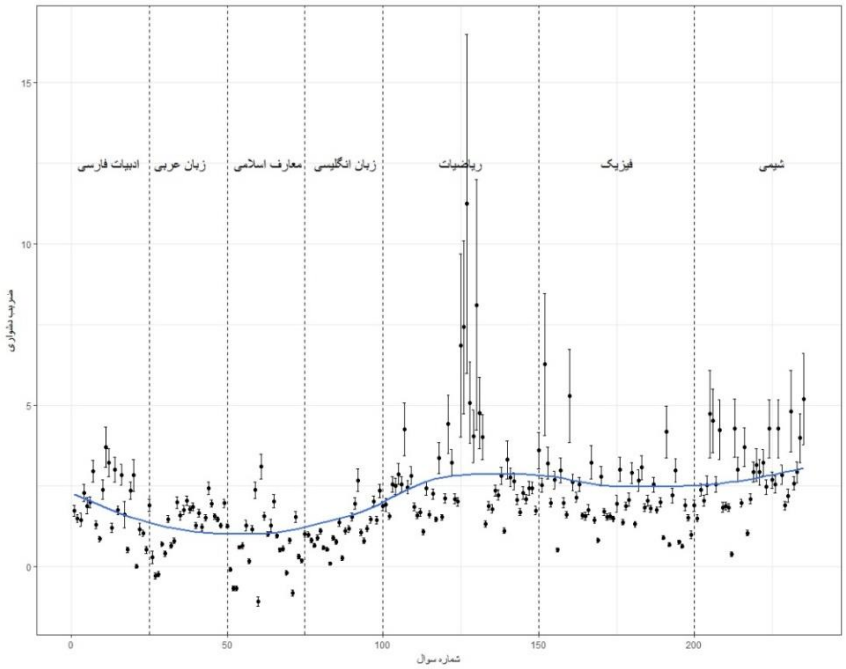
- **پارامترهای سؤال**

این بخش به بررسی پارامترهای دشواری و تمیز سؤالات ۲۳۵ گانه از جنبه‌های مختلف پرداخته است. تمامی برآورد پارامترهای سؤال و توانایی با استفاده از نرم‌افزار BILOG-MG صورت گرفته است. ابتدا، جدول ۴ خلاصه‌ای از آماره‌های توصیفی ضرایب تمیز و دشواری سؤالات را نشان می‌دهد.

جدول ٤. خلاصه نتایج تحلیل IRT سؤالات آزمون

	کمترین	چارک اول	میانه	چارک سوم	بیشترین	میانگین	انحراف معیار
ضریب تمیز سؤال	۰/۱۴۹۸	۰/۶۰۹۳	۰/۷۹۸۷	۰/۹۹۵۷	۱/۶۳۵۲	۰/۸۰۴۹	۰/۲۶۷۴
ضریب دشواری سؤال	-۱/۰۹۱۴	۱/۲۰۲۳	۱/۸۷۵۱	۲/۶۱۰۵	۱۱/۲۴۵۲	۲/۰۴۵۵	۱/۴۵۸۸

با توجه به مقادیر جدول ۴، می‌توان گفت که ۵۰٪ درصد سؤالات دارای ضریب تمیز حدود ۰/۸ و کمتر هستند. میانگین حدود ۰/۸ حاکی از توزیع نسبتاً نرمال ضرایب تمیز سؤالات دارد. تنها ۲۵٪ سؤالات دارای ضریب تمیز بین ۱ و ۱/۶۴ هستند. برخلاف ضریب همبستگی دورشته‌ای نقطه‌ای، مقدار کمینه‌ای برای ضریب تمیز سؤال در IRT تعیین نشده است. ملاک مطلوبیت بستگی به هدف آزمون دارد. این امر که ۵۰٪ سؤالات دارای ضریب تمیز برابر و یا کمتر از ۰/۸ هستند بر میزان حداکثر تابع آگاهی آزمون تأثیر می‌گذارد و آن را کمتر از مقدار بهینه (با توجه به تعداد سؤالات) می‌کند.



شکل ۳. توزیع ضریب دشواری سؤالات براساس مفاد آزمون

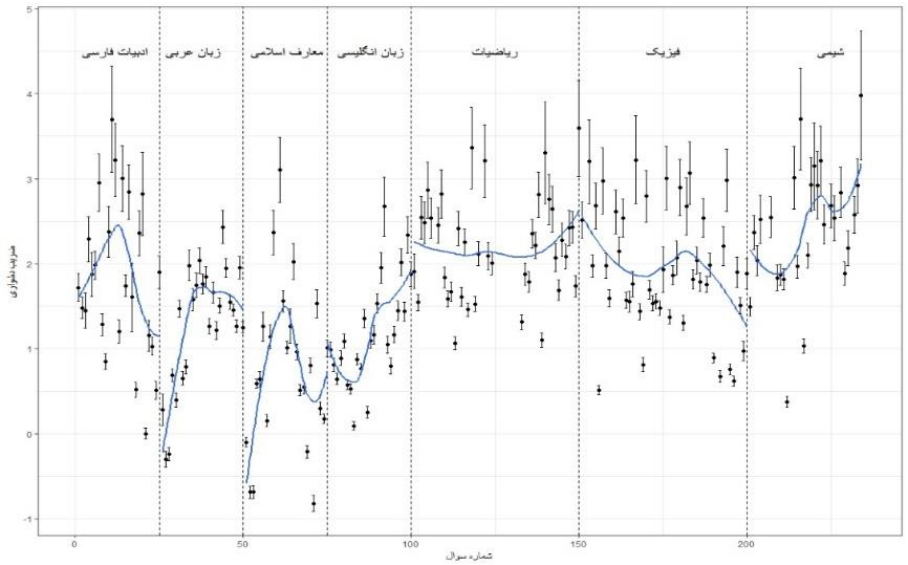
شکل ۳ توزیع ضرایب دشواری سؤالات براساس مفاد آزمون را نشان می‌دهد. شکل ۳ به‌خوبی نشان می‌دهد که سؤالات دارای دشواری نامعقول عموماً در دروس تخصصی، به‌ویژه ریاضی و شیمی، قرار دارند. منحنی درون شکل ۳ نشانگر روند دشواری سؤالات در سطح آزمون است. این روند تحت تأثیر وجود سؤالات با دشواری بسیار بالا قرار گرفته است. بررسی روند دشواری سؤالات در آزمون از این لحاظ اهمیت دارد که ترتیب دشواری سؤالات در آزمون بر عملکرد آزمودنی تأثیر مستقیم دارد. مطلوب این است که آزمون ابتدا با سؤالات آسان شروع‌شده و سپس به‌دشواری سؤالات افزوده شود. به‌منظور بررسی دقیق‌تر سؤالاتی که ضریب دشواری بسیار بالا دارند، مواردی را که ضریب دشواری بزرگ‌تر یا مساوی ۴ هستند در جدول ۷ لیست شده‌اند. درمجموع ۲۱ سؤال شناسایی شد که ۱۰ سؤال مربوط به درس ریاضی، ۳ سؤال درس فیزیک و ۸ سؤال مربوط به درس شیمی بود.

جدول ۵. ویژگی‌های روان‌سنجی سؤالات

آزمون	سؤال	ضریب تمیز	خطای برآورد	ضریب دشواری	خطای برآورد	درصد پاسخ صحیح	ضریب همبستگی دورشته‌ای نقطه‌ای
ریاضی (۱۰ سؤال)	۱۰۷	۰/۴۹۱	۰/۰۵۶	۴/۲۶۲	۰/۴۱۹	۳/۸	۰/۳۹۵
	۱۲۱	۰/۵۳۹	۰/۰۶۶	۴/۴۱۲	۰/۴۵۹	۲/۶	۰/۴۲۴
	۱۲۵	۰/۳۶۲	۰/۰۸۱	۶/۸۵۷	۱/۴۴۶	۱/۷	۰/۱۳۶
	۱۲۶	۰/۲۱۶	۰/۰۴۱	۷/۴۲۳	۱/۳۶۹	۶/۵	۰/۰۹۵
	۱۲۷	۰/۱۵۰	۰/۰۳۶	۱۱/۲۴۵	۲/۶۸۶	۵/۵	-۰/۰۳۳
	۱۲۸	۰/۲۷۴	۰/۰۳۶	۵/۰۸۳	۰/۶۴۲	۹/۳	۰/۲۰۹
	۱۲۹	۰/۳۷۱	۰/۰۴۳	۴/۰۴۱	۰/۴۲۱	۸/۴	۰/۲۸۸
	۱۳۰	۰/۳۲۳	۰/۰۸۲	۸/۱۱۴	۱/۹۸۸	۱/۳	۰/۰۶۹
	۱۳۱	۰/۳۵۸	۰/۰۴۶	۴/۷۶۷	۰/۵۶۴	۶/۱	۰/۲۶۷
	۱۳۲	۰/۶۰۶	۰/۰۶۴	۴/۰۱۱	۰/۳۵۳	۲/۶	۰/۴۹۸
فیزیک (۳ سؤال)	۱۵۲	۰/۴۴۱	۰/۰۸۶	۶/۲۶۳	۱/۱۲۸	۱/۲	۰/۲۳۶
	۱۶۰	۰/۳۹۲	۰/۰۶۰	۵/۲۸۳	۰/۷۳۴	۳/۵	۰/۲۵۱
	۱۹۱	۰/۵۱۸	۰/۰۵۹	۴/۱۷۱	۰/۴۰۸	۳/۵	۰/۴۱۷

آزمون	سؤال	ضریب تمیز	خطای برآورد	ضریب دشواری	خطای برآورد	درصد پاسخ صحیح	ضریب همبستگی دورشته‌ای نقطه‌ای
شیمی (۸ سؤال)	۲۰۵	۰/۵۸۸	۰/۱۰۵	۴/۷۲۵	۰/۶۹۶	۱/۴	۰/۳۹۵
	۲۰۶	۰/۵۰۶	۰/۰۶۵	۴/۵۲۳	۰/۵۰۵	۲/۸	۰/۳۸۴
	۲۰۸	۰/۴۸۱	۰/۰۶۲	۴/۲۳۶	۰/۴۷۰	۴/۱	۰/۳۵۷
	۲۱۳	۰/۶۰۴	۰/۰۷۹	۴/۲۸۰	۰/۴۵۹	۲	۰/۴۵۱
	۲۲۴	۰/۷۰۲	۰/۰۹۶	۴/۲۶۷	۰/۴۵۷	۱/۲	۰/۵۲۹
	۲۲۷	۰/۵۰۸	۰/۰۶۴	۴/۲۷۶	۰/۴۶۰	۳/۴	۰/۳۸۰
	۲۳۱	۰/۵۹۵	۰/۰۹۵	۴/۸۱۵	۰/۶۴۸	۱/۳	۰/۳۹۹
	۲۳۵	۰/۵۲۱	۰/۰۸۳	۵/۱۹۰	۰/۷۲۹	۱/۵	۰/۳۵۷

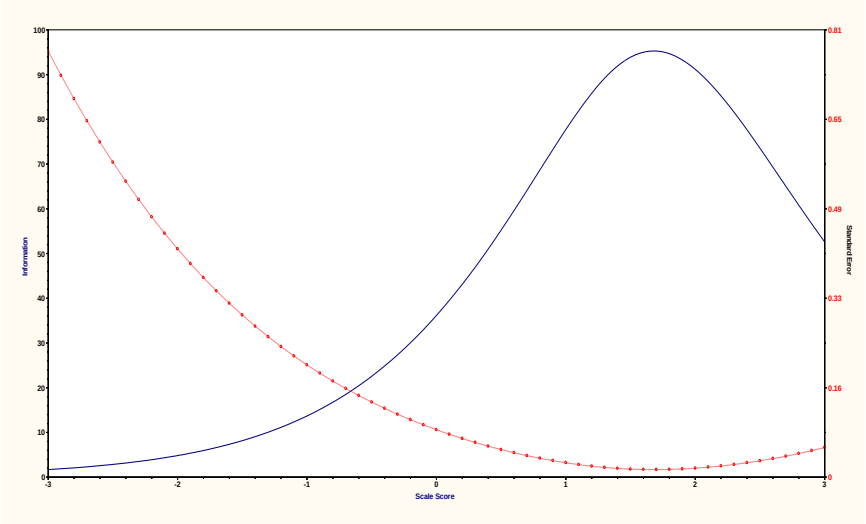
مرور جدول ۵ به‌خوبی نشان می‌دهد که سؤالات با ضریب دشواری نامعقول اکثراً دارای ضریب تمیز پایین هستند. مثلاً سؤال ۱۲۷ در درس ریاضی با ضریب دشواری ۱۱/۲۴۵ دارای ضریب تمیز برابر با ۰/۱۵ است، تنها ۵,۵٪ از کل آزمودنی‌ها پاسخ صحیح به این سؤال داده‌اند و ضریب همبستگی دو رشته‌ای نقطه‌ای آن‌هم منفی است. بررسی خطای استاندارد برآورد پارامتر به وضوح نشان می‌دهد که پارامتر برآورد شده این‌گونه سؤالات با داشتن مقدار بالای خطا، قابل‌اطمینان نیستند؛ بنابراین، به‌منظور داشتن تصویری منطقی‌تر از وضعیت دشواری سؤالات در آزمون، سؤالات مشخص‌شده در جدول ۵ از تحلیل حذف‌شده و شکل ۳ مجدداً تولید گردید. نمودار نهایی به همراه روند تغییر دشواری سؤالات در هر درس در شکل ۴ نشان داده شده است.



شکل ٤. توزیع ضریب دشواری سؤالات براساس مفاد آزمون (اصلاح‌شده)

براساس شکل ۴ به‌راحتی می‌توان دید که روند دشواری سؤالات در دروس مختلف از اصول مطلوب آزمون‌سازی پیروی نمی‌کند. مثلاً در درس فیزیک سؤالات از بسیار دشوار آغازشده و به آسان ختم می‌شود. درمجموع به نظر می‌رسد دروس عمومی از روند دشواری نسبتاً بهتری نسبت به دروس تخصصی برخوردارند.

درنهایت، شکل ۵ تابع آگاهی آزمون (منحنی آبی‌رنگ) به همراه خطای استاندارد برآورد توانایی (منحنی قرمزرنگ) را نشان می‌دهد. حداکثر آگاهی (و حداقل خطا) در این آزمون برای افرادی با توانایی حدود ۱/۸ قرار دارد و حداکثر آگاهی‌بخشی این آزمون ۱۰۰ است. این بدین معناست که از مجموع ۲۳۵ سؤال آزمون، حداکثر آگاهی ممکن از یک آزمون ۱۰۰ سؤالی قابل‌استخراج است. این امر ناشی از پراکندگی ضرایب تمیز سؤالات حول مقدار ۰/۸ ناشی می‌شود. در مدل دوپارامتری، آگاهی آزمون متناسب با مجذور ضریب تمیز است. اگر فرض کنیم تمام سؤالات دارای ضریب تمیز ۰/۸ هستند، در این‌صورت حداکثر آگاهی آزمون متناسب خواهد بود با (۰/۸^۲)×۲۳۵ یا ۱۵۰/۴ (با فرض ثابت بودن احتمال پاسخ صحیح برای تمام سؤالات).

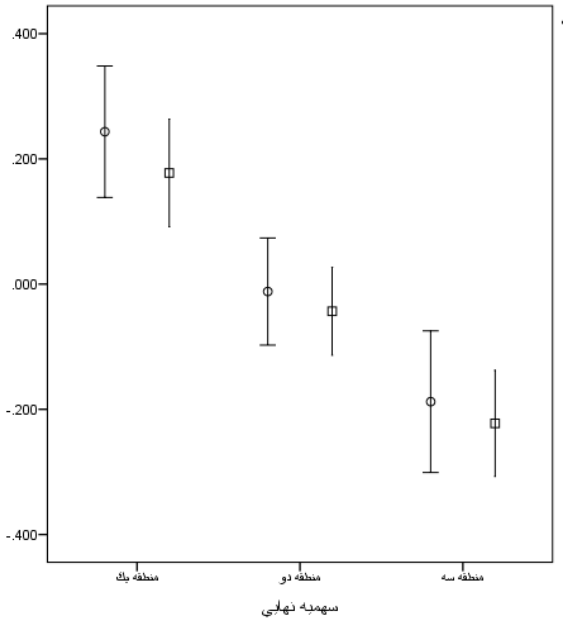


شکل ۵. تابع آگاهی آزمون

▪ توزیع توانایی

توانایی آزمودنی‌ها براساس مدل دو پارامتری IRT و با روش پسین مورد انتظار[1] (EAP) برآورد شد. شایان ذکر است که برآورد توانایی برای تمام آزمودنی‌ها به‌عنوان یک گروه واحد صورت گرفته و نه به‌صورت چندگروهی براساس سهمیه انتخابی. با توجه به میانگین‌ها می‌توان گفت که میانگین کل توانایی آزمودنی‌ها از منطقه یک به منطقه سه روندی کاهشی دارد. در هر سه منطقه، داوطلبان زن دارای میانگین توانایی بالاتر با توزیعی همگون‌تر هستند. بیشتر اختلاف در میانگین توانایی بین آزمودنی‌های زن و مرد در منطقه یک و کمترین اختلاف بین آزمودنی‌های منطقه دو وجود دارد. می‌توان گفت که توزیع توانایی آزمودنی‌های مناطق دو و سه، در مقایسه با منطقه یک، شباهت بیشتری به یکدیگر دارد. شکل ۶ میانگین توانایی آزمودنی‌ها براساس سهمیه و جنسیت به همراه فاصله اطمینان ۹۵٪ نشان می‌دهد.

1. Expected a Posteriori



شکل ٦. میانگین توانایی و فاصله اطمینان ۹۵٪ (دایره= زن، مربع= مرد)

○ یافته‌های فاز دوم: مطالعه شبیه‌سازی

ازآنجاکه هدف این تحقیق بررسی تأثیر ساختار سلسله مراتبی داده‌ها بر توزیع پارامتر توانایی و رتبه آزمودنی‌ها است، مقایسه بین حالتی که ساختار سلسله مراتبی در نظر گرفته می‌شود و حالتی که در نظر گرفته نمی‌شود نیازمند مبنایی برای مقایسه است. در عمل، هنگامی‌که داده‌ها مورد تحلیل قرار می‌گیرند هیچ‌گونه اطلاعاتی دربارهٔ وضعیت حقیقی داده‌ها (مثلاً مقدار حقیقی پارامترهای سؤال و توانایی) در دست نیست تا بتوان نتایج حاصل از تحلیل را با آن مقایسه نمود. لذا استفاده از مطالعات شبیه‌سازی[1] یا مونت کارلو[2] می‌تواند مبنایی برای مقایسه فراهم آورد. در این نوع مطالعات، داده‌ها توسط پژوهشگر و براساس مفروضه‌هایی معین تولیدشده و سپس نتایج تحلیل با آن‌ها مقایسه می‌شود (فاینبرگ و

1. Simulatuin
2. Monte Carlo

روبرایت[1]، ۲۰۱۶)؛ بنابراین، نتایج حاصل از تحلیل اولیه داده‌های آزمون به‌عنوان چارچوبی برای پاسخ دادن به سؤالات تحقیق بکار گرفته شد.

▪ طرح شبیه‌سازی

به‌منظور تولید داده‌هایی که تناظر بهتری با واقعیت داشته باشند، پارامترهای سؤال برآورد شده به‌عنوان پارامترهای حقیقی سؤال در مطالعه شبیه‌سازی بکار گرفته شدند. برای داده-های متناظر با داوطلبان سهمیه منطقه یک تعداد ۹۳۹ نمونه تصادفی از توزیع نرمال با میانگین ۰/۲۰۳ و انحراف استاندارد ۱/۰۳۴، سهمیه منطقه دو تعداد ۱۲۵۷ نمونه تصادفی از توزیع نرمال با میانگین ۰/۰۳۱- و انحراف استاندارد ۰/۹۷۷ و تعداد ۷۴۱ نمونه تصادفی از توزیع نرمال با میانگین ۰/۲۱۰- و انحراف استاندارد ۰/۹۳۶ براساس ساختار سلسله مراتبی (داوطلبان در سهمیه‌ها آشیان شده‌اند) به‌عنوان پارامتر توانایی تولید شد. سپس این پارامتر توانایی به همراه پارامترهای دشواری و تمیز سؤالات برای تولید ماتریس پاسخ سؤالات بکار رفت. بدین‌صورت که برای هر پارامتر توانایی و پارامترهای سؤال مورد تحلیل، احتمال پاسخ صحیح براساس مدل دو پارامتری IRT محاسبه شد. سپس این احتمال با یک عدد تصادفی تولیدشده در بازه ۰ تا ۱ مقایسه شد. چنانچه احتمال محاسبه‌شده از مقدار تصادفی بین ۰ و ۱ بیشتر بود، پاسخ سؤال مذکور صحیح (با کد ۱) و در غیراین‌صورت غلط (با کد ۰) کدگذاری گردید. این فرآیند برای تمام آزمودنی‌ها و تمام سؤالات انجام شد.

سپس داده‌های تولیدشده به دو روش مورد تحلیل قرار گرفت: ۱) بدون در نظر گرفتن ساختار سلسله مراتبی و ۲) با در نظر گرفتن ساختار سلسله مراتبی داده‌ها. درنهایت تفاوت توزیع پارامتر توانایی و رتبه افراد درون سهمیه و در کل نمونه با مقادیر حقیقی بررسی گردید.

1. Fienberg & Rubright

▪ نتایج شبیه‌سازی

جدول ۶ خلاصه نتایج تحلیل IRT سؤالات شبیه‌سازی‌شده را ارائه می‌دهد. تمامی مقادیر (به‌جز بیشترین مقدار ضریب دشواری) تقریب بسیار خوبی از توزیع پارامترهای اصلی آزمون را نشان می‌دهد. ضریب همبستگی بین مقادیر حقیقی و برآورد شده در مطالعه شبیه‌سازی برای ضرایب تمیز و دشواری سؤال، به ترتیب عبارت است از ۰/۹۷۷ و ۰/۹۷۴. هر دو ضریب همبستگی در سطح $\alpha \leq$ ۰,۰۰۱ معنادار بود.

جدول ٦. خلاصه نتایج تحلیل IRT سؤالات شبیه‌سازی‌شده

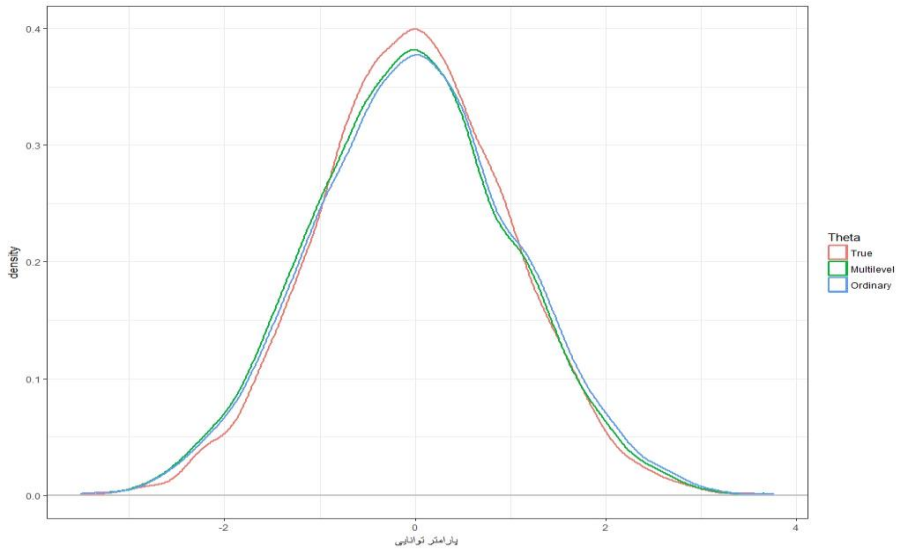
	کمترین	چارک اول	میانه	چارک سوم	بیشترین	میانگین	انحراف معیار
ضریب تمیز سؤال	۰/۱۷۹	۰/۵۸۵	۰/۷۹۱	۱/۰۱۶	۱/۶۳۱	۰/۸۰۷	۰/۲۸۰
ضریب دشواری سؤال	-۱/۱۵۹	۱/۲۳۶	۱/۸۲۳	۲/۶۶۸	۷/۹۴۵	۲/۰۴۷	۱/۳۹۷

جدول ۶ حاوی ویژگی‌های توزیع پارامتر توانایی براساس داده‌های حقیقی تولیدشده، برآورد توانایی با در نظر گرفتن ساختار سلسله مراتبی داده‌ها و برآورد توانایی بدون در نظر گرفتن ساختار سلسله مراتبی داده‌ها برای مناطق سه‌گانه و در کل نمونه است. در سطح کل نمونه، تحلیل چندسطحی میانگین حقیقی را کم برآورد (۰/۰۴۶-) و تحلیل غیر چندسطحی آن را بیش برآورد (۰/۰۰۳) کرده است. هر دو برآورد مقدار حقیقی انحراف استاندارد توزیع توانایی را بیش برآورد کرده‌اند، اما با اختلاف بسیار کمی، تحلیل چند سطحی برآورد نزدیک‌تری به‌دست داده است. روند مشابهی را می‌توان در نتایج مناطق یک و دو مشاهده کرد، با این تفاوت ناچیز که در تحلیل داده‌های منطقه سه روش چندسطحی مقدار میانگین را بیش برآورد (۰/۲۳۹-) و روش غیرچندسطحی میانگین را کم برآورد (۰/۲۴۹-) کرده است.

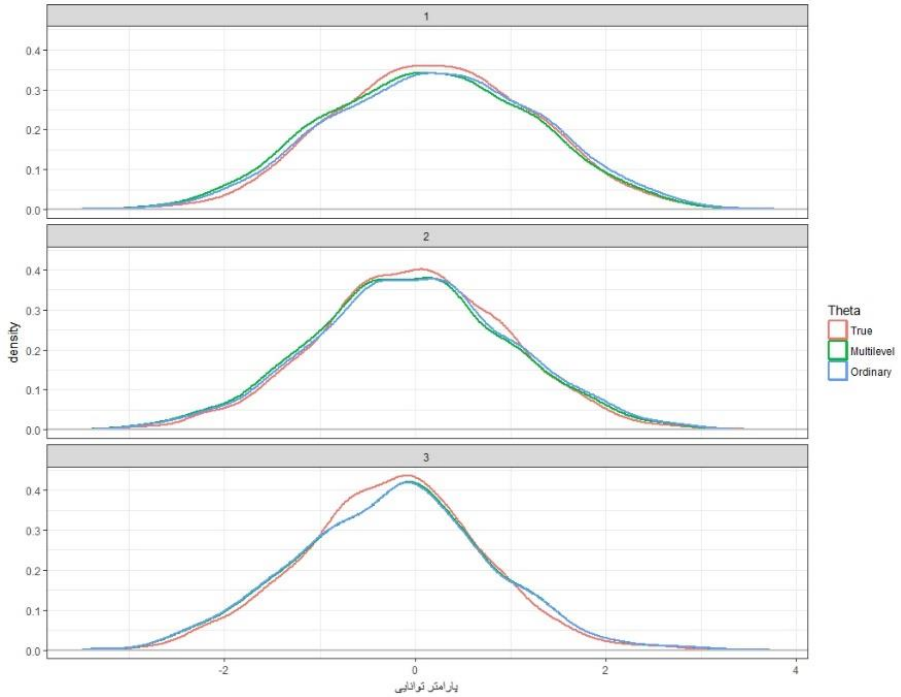
شکل‌های ۷ و ۸ تابع چگالی توزیع پارامتر توانایی به ترتیب برای کل نمونه و در سهمیه‌ها نشان می‌دهد. در این نمودارها، منحنی قرمز رنگ مربوط به مقادیر حقیقی تولیدشده برای پارامتر توانایی (True)، منحنی سبزرنگ مربوط به مقادیر برآورد شده برای

پارامتر توانایی در تحلیل چندسطحی (Multilevel) و منحنی آبی‌رنگ مربوط به مقادیر برآورد شده برای پارامتر توانایی در تحلیل غیرچندسطحی (Ordinary) است.

همان‌طور که در هر دو نمودار مشهود است، توزیع‌های حاصل از تحلیل چندسطحی و غیرچندسطحی بسیار به یکدیگر نزدیک هستند، به‌ویژه در داده‌های منطقه سه. به نظر می‌رسد که برآوردهای پارامتر توانایی در میانه توزیع عموماً کم‌برآورد و در حاشیه‌های توزیع عموماً بیش‌برآورد شده‌اند. علاوه بر این، مقایسه بین تحلیل چندسطحی و غیرچندسطحی نشان می‌دهد که برآورد چگالی توزیع (و متعاقباً فراوانی مشاهده‌شده) برای مقادیر توانایی کمتر از میانگین در تحلیل چندسطحی کمی بیشتر از تحلیل غیرچندسطحی است، درحالی‌که این روند برای مقادیر توانایی بالاتر از میانگین برعکس است؛ یعنی برای مقادیر بالاتر از میانگین، چگالی توزیع (و متعاقباً فراوانی مشاهده‌شده) در تحلیل چندسطحی کمی کمتر از تحلیل غیر چندسطحی است. این نتیجه می‌تواند بر رتبه افراد تأثیر بگذارد.

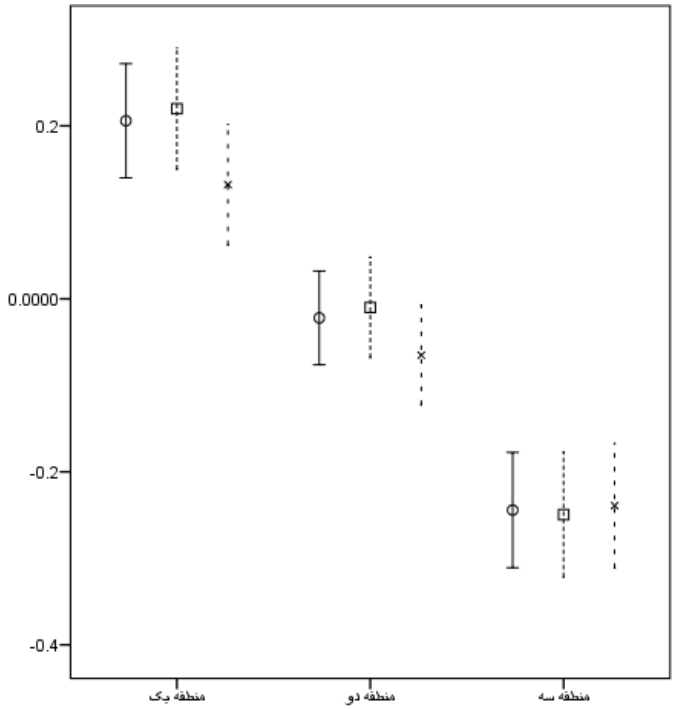


شکل ۷. توزیع پارامتر توانایی در کل نمونه (قرمز= حقیقی، سبز= چندسطحی، آبی= غیرچندسطحی)



شکل ۸. توزیع پارامتر توانایی در سهمیه (قرمز= حقیقی، سبز= چندسطحی، آبی= غیر چندسطحی)

شکل ۹ نیز نشان می‌دهد که فاصله اطمینان برآورد میانگین توانایی در تحلیل غیرچندسطحی اندکی بیشتر از تحلیل چندسطحی، اما قابل‌مقایسه است.



شکل ۹. میانگین توانایی و فاصله اطمینان ۹۵٪ (دایره= حقیقی، مربع= غیرسلسله مراتبی، ضربدر= سلسله مراتبی)

براساس مبانی نظری تحلیل‌های چندسطحی، چنین انتظار می‌رود که درصورتی‌که داده‌ها دارای ساختار سلسله‌مراتبی باشند، تحلیل چندسطحی باید نتایج بهتری نسبت به تحلیل غیرچندسطحی ارائه دهد (اسنایدر و بوسکر، ۲۰۰۰). این امر تا حدی زیادی بستگی به مقدار همبستگی درون طبقه‌ای[1] (یا خوشه‌ای یا رده‌ای) یا ICC دارد. این ضریب همبستگی نشانگر درصدی از واریانس متغیر وابسته است که ناشی از ساختار سلسله‌مراتبی و یا تفاوت بین طبقات است. مقدار ICC برای توانایی‌های برآورد شده از داده‌های آزمون حدود ۰/۰۴ بود. این بدین معناست که تنها حدود ۴٪ از واریانس مشاهده‌شده در برآورد توانایی ناشی از قرار گرفتن افراد در سهمیه‌های سه‌گانه است. ازآنجاکه طرح مطالعه شبیه‌سازی نیز

1. Intraclass Correlation

براساس داده‌های آزمون طرح‌ریزی‌شده بود، همین مقدار نیز در تولید داده‌ها بکار رفته است. این مقدار پایین برای ICC می‌تواند دلیل اصلی در شباهت بسیار زیاد بین نتایج تحلیل چندسطحی و غیر چندسطحی باشد. در چنین حالتی، استفاده از مدل پیچیده‌تری مانند مدل چندسطحی کمکی به بهبود نتایج نمی‌کند.

جدول ۷ توزیع رتبه آزمودنی‌ها در کل نمونه را به تفکیک منطقه و کل نشان می‌دهد. مقایسه میانگین و میانه رتبه نشان می‌دهد که محاسبه پارامتر توانایی و رتبه‌بندی آزمودنی‌ها با استفاده از تحلیل چندسطحی برآورد بهتری از مقادیر حقیقی، در مقایسه با روش غیر چندسطحی را به‌دست می‌دهد. به‌عنوان‌مثال، تحلیل غیر چندسطحی میانگین رتبه در مناطق یک و دو را کم برآورد و در منطقه سه، بیش برآورد کرده است. تفاوت مشهود دیگر، مربوط به کمترین و بیشترین رتبه است. علیرغم اینکه هر دو روش برآورد پارامتر مقادیر یکسانی برای کمترین رتبه مشاهده شده دارند، این برآوردها با مقادیر حقیقی تفاوت دارد. به‌ویژه در منطقه سه که حداقل رتبه حقیقی ۸ بوده اما مقدار آن ۲ برآورد شده است.

جدول ۷. ویژگی‌های توزیع رتبه

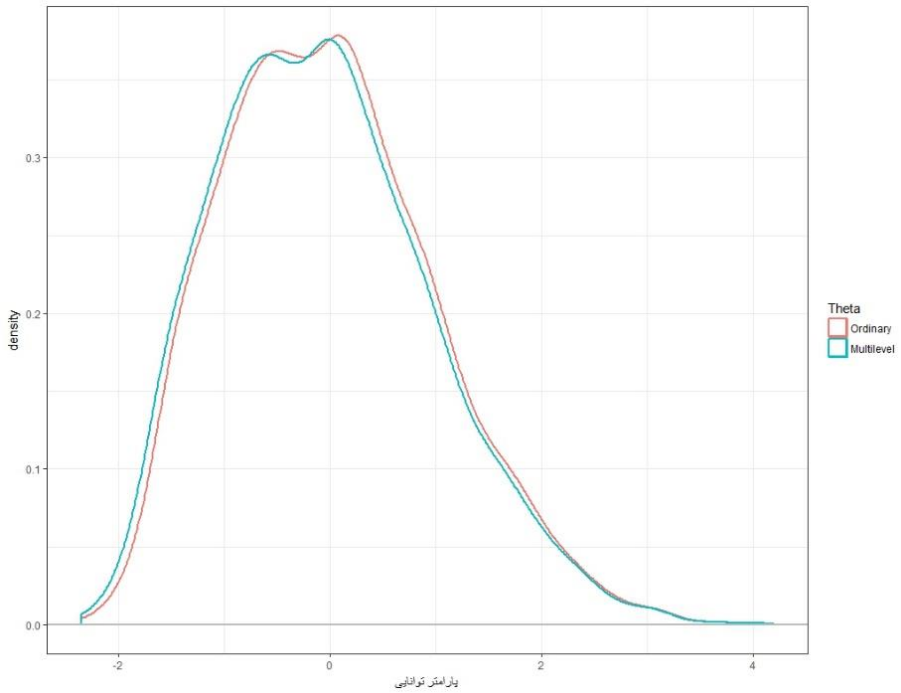
		سهمیه			
		منطقه یک	منطقه دو	منطقه سه	کل نمونه
پارامتر حقیقی توانایی	میانگین	۱۶۳۲	۱۴۵۱	۱۲۶۲	۱۴۶۱
	کمترین	۲	۱	۸	۱
	میانه	۱۷۳۱	۱۴۶۱	۱۲۲۸	۱۴۶۱
	بیشترین	۲۹۲۱	۲۹۲۰	۲۹۰۸	۲۹۲۱
پارامتر توانایی حاصل از تحلیل چندسطحی	میانگین	۱۶۲۸	۱۴۵۲	۱۲۶۵	۱۴۶۱
	کمترین	۱	۳	۲	۱
	میانه	۱۶۹۵	۱۴۵۲	۱۲۳۹	۱۴۶۱
	بیشترین	۲۹۲۰	۲۹۲۱	۲۹۱۷	۲۹۲۱
پارامتر توانایی حاصل از تحلیل غیر چندسطحی	میانگین	۱۵۹۹	۱۴۴۶	۱۳۱۱	۱۴۶۱
	کمترین	۱	۳	۲	۱
	میانه	۱۶۵۲	۱۴۴۱	۱۲۹۴	۱۴۶۱
	بیشترین	۲۹۲۰	۲۹۲۱	۲۹۱۷	۲۹۲۱

○ تحلیل داده‌ها براساس نتایج شبیه‌سازی

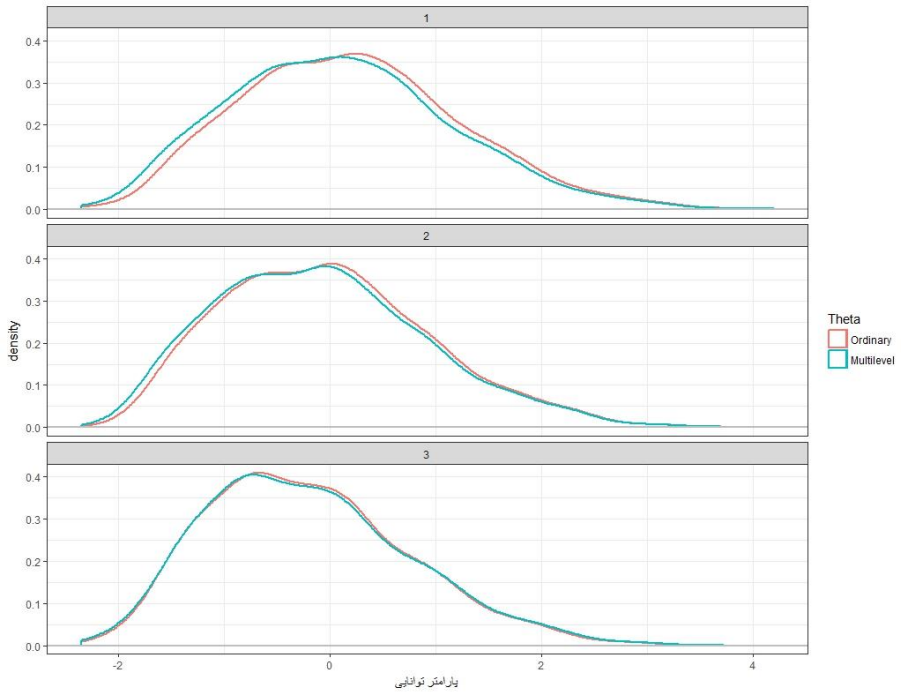
شایان ذکر است که به دلیل در اختیار قرار دادن تنها بخشی از داده‌های آزمون توسط سازمان سنجش، تمامی رتبه‌ها در فایل داده موجود نبود، لذا از رتبه مشاهده‌شده افراد در فایل داده‌ها برای این تحلیل استفاده شد. برای این منظور نمره کل آزمون موجود در فایل داده‌ها برای رتبه‌بندی آزمودنی در کل نمونه و در سهمیه بدین‌صورت بکار رفت که رتبه ۱ به بالاترین نمرهٔ کل تخصیص داده شد.

بر اساس نتایج، پارامتر توانایی برآورد شده با استفاده از تحلیل غیر چندسطحی دارای میانگینی کمی بیشتر از روش چندسطحی، با انحراف استاندارد تقریباً یکسان، در کل نمونه است. این نتیجه با توجه به مقدار ICC و مطالعه شبیه‌سازی قابل پیش‌بینی بود. روندی مشابه را می‌توان در داده‌های مربوط به مناطق یک و دو مشاهده کرد. نتایج حاصل از منطقه سه با استفاده از دو روش تقریباً یکسان هستند.

شکل ۱۰ نیز تابع چگالی مشاهده‌شده براساس پارامتر توانایی حاصل از تحلیل غیرچندسطحی (Ordinary) و تحلیل چندسطحی (Multilevel) را نشان می‌دهد. هر دو توزیع دارای کمی چولگی مثبت هستند. روندی نسبتاً مشابه را می‌توان در توزیع توانایی در مناطق سه‌گانه مشاهده کرد.



شکل ۱۰. توزیع پارامتر توانایی در کل نمونه (قرمز= غیرچندسطحی، آبی= چندسطحی)



شکل ۱۱. توزیع پارامتر توانایی در سهمیه (قرمز= غیر چندسطحی، آبی= چندسطحی)

به نظر می‌رسد که در کل نمونه و مناطق یک و دو برآوردهای حاصل روش غیرسلسله مراتبی برای مقادیر توانایی کمتر از صفر پایین‌تر و برای مقادیر بیشتر از صفر بیشتر از برآوردهای حاصل از روش چندسطحی باشد. این امر می‌تواند بر توزیع رتبه‌ها در کل نمونه و درون سهمیه‌ها اثرگذار باشد. جدول ۸ حاوی آمار توصیفی توزیع رتبه کل آزمودنی براساس نمرهٔ کل آزمون و پارامتر توانایی حاصل از تحلیل‌های چندسطحی و غیرچندسطحی است.

جدول ۸. ویژگی‌های توزیع رتبه‌های کل

		سهمیه		
		منطقه یک	منطقه دو	منطقه سه
نمره کل آزمون	میانگین	۱۲۸۱	۱۴۹۸	۱۶۲۷
	کمترین	۱	۲	۱۱
	میانه	۱۱۸۷	۱۵۰۳	۱۶۶۷
	بیشترین	۲۹۰۹	۲۹۲۱	۲۹۱۸
تحلیل چندسطحی	میانگین	۱۳۳۲	۱۴۸۴	۱۵۸۴
	کمترین	۱	۳	۱۰
	میانه	۱۲۵۷	۱۴۸۲	۱۶۴۳
	بیشترین	۲۹۲۱	۲۹۱۷	۲۹۲۰
تحلیل غیرچندسطحی	میانگین	۱۲۹۲	۱۴۸۴	۱۶۳۶
	کمترین	۱	۳	۱۵
	میانه	۱۲۰۳	۱۴۷۹	۱۶۸۱
	بیشترین	۲۹۲۰	۲۹۱۶	۲۹۲۱

بررسی داده‌های جدول فوق نشان می‌دهد که نتایج حاصل از نمره کل آزمودنی و تحلیل غیرچندسطحی نتایجی نسبتاً مشابه دارند. تحلیل چندسطحی میانگین و میانه رتبه بالاتری برای منطقه دو و میانگین و میانه رتبه پایین‌تری برای منطقه سه را در مقایسه با دو روش دیگر برآورد کرده است. بررسی کمترین و بیشترین مقدار رتبه نشان می‌دهد که همخوانی بیشتری بین تحلیل‌های چندسطحی و غیرچندسطحی وجود دارد. نکته قابل‌توجه تغییر روند مقادیر بیشترین رتبه آزمودنی براساس نمره کل آزمودنی در مقایسه با دو روش دیگر است. براساس نمره کل آزمودنی منطقه دو دارای بالاترین مقدار رتبه است، درحالی‌که در دو روش دیگر منطقه دو دارای کمترین مقدار برای حداکثر رتبه است.

به‌منظور بررسی این‌که آیا تفاوت معناداری بین رتبه‌ها براساس سه روش مذکور وجود دارد، از آزمون فریدمن استفاده شد. مقدار خی دو به‌دست‌آمده با دو درجه آزادی برابر بود با ۱۳/۲۳۴ که حاکی از معناداری نتایج در سطح $\alpha \leq$ ۰,۰۰۱ است. مقایسه چندگانه نشان داد که تفاوت معناداری بین رتبه‌دهی، به‌ویژه، براساس نمره کل آزمون و روش

چندسطحی و نیز بین روش‌های چندسطحی و غیرچندسطحی وجود دارد. همچنین نتایج نشان داد تفاوت چشمگیری بین رتبه‌دهی افراد درون سهمیه‌ها بین سه روش موردبررسی وجود ندارد. آزمون فریدمن نیز هیچ‌گونه تفاوت معناداری بین رتبه‌دهی‌ها درون سهمیه نشان نداد.

٤. بحث و نتیجه‌گیری

عدالت آموزشی ایجاب می‌کند که همه آحاد انسانی باید از آموزش‌وپرورش یکسان برخوردار گردند. یکی از شاخص‌های پیشرفت اجتماعی، آموزش‌وپرورش و چگونگی بهره‌مندی از آن می‌باشد. بر همین مبنا کشورها به‌خصوص کشورهای درحال‌توسعه درصدد ایجاد فرصت‌های برابر دسترسی همگان به آموزش‌وپرورش هستند (هرمان، کلاین و عابدی[1] (۲۰۰۰؛ جی[2]، ۲۰۰۸). برابری به مهیاسازی فرصت‌های برابر اشاره دارد، چراکه افراد در توانایی، دانش، مهارت و نیازهای آموزشی‌شان با هم متفاوت‌اند. منظور از فرصت برابر نیز، جلوگیری، حذف یا کاهش تبعیض بین افراد ازلحاظ جنسیت، نژاد، وضعیت جسمانی، سنی، زبانی، طبقهٔ اجتماعی است (بنت، بوث و یندل[3]، ۲۰۰۱).

روایی تصمیمات مهم آموزشی ازجمله پاسخگویی و یا دسته‌بندی افراد در طبقات مختلف مهارتی به‌طور مستقیم تحت تأثیر روایی نمره به‌دست‌آمده از آزمون است؛ بنابراین ارزیابی روایی تفسیرهای منتج از نمرات آزمون که می‌تواند تحت تأثیر عوامل مختلفی ازجمله مدل اندازه‌گیری، نوع سؤال، شیوهٔ نمره‌گذاری آزمون، نحوه برآورد پارامتر توانایی و غیره باشد، از اهمیت بسزایی برخوردار است. یکی از عواملی که می‌تواند در روایی نمرات برآورد شده تأثیر داشته باشد، ساختار داده‌های گردآوری‌شده و انتخاب مدل آماری مناسب برای تحلیل داده‌هاست (راتکوسکی، گنزالن، جونکاس و ون داویر[4]، ۲۰۱۰). در برخی موارد این ساختار سلسله‌مراتبی به‌صورت طبیعی وجود دارد. گاه ساختار

1. Herman, Klein, & Abedi
2. Gee
3. Bennett, Both & Yeandle
4. Rutkowski, Gonzalen, Joncas, & von Davier

سلسله‌مراتبی داده‌ها منتج از تصمیمات و سیاستگذاری‌هاست. مثلاً در آزمون سراسری دانشگاه‌ها، کلیه استان‌ها به مناطق چندگانه تقسیم‌بندی شده‌اند.

یکی از مشکلات قابل‌توجه در داده‌های بکار گرفته‌شده در این پژوهش، وجود تعداد بسیار بالای داده‌های بدون پاسخ بود. این امر نه‌تنها در تعیین مؤثر دشواری و ضریب تمیز سؤالات براساس نظریه کلاسیک و IRT تأثیر می‌گذارد، بلکه با توجه به تأثیر نمره منفی در محاسبه نمره خام، تأثیر آن در رتبه افراد نیز غیرقابل‌اغماض است. مسئله عمده‌ای که رخ می‌دهد عدم توانایی آزمون در تمیز افرادی است که دارای بخشی از توانایی موردسنجش و یا عدم آن هستند. چراکه داده‌های بدون پاسخ امکان چنین تمایزی را نمی‌دهد. این موضوع می‌تواند یکی از دلایل عمده همپوشانی بالا بین بازه رتبه افرادی باشد که براساس تعداد پاسخ دسته‌بندی‌شده‌اند. نکته دیگری که در داده‌ها مشاهده شد ولی توضیح و توجیهی منطقی برای آن یافت نشد، بازه رتبه افراد در سه منطقه و براساس جنسیت بود. برای مناطق ۱ و ۲ حداکثر رتبه بسیار بالاتر از رتبه آزمودنی‌های منطقه ۳ بود. این امر ممکن است ناشی از شیوهٔ نمونه‌گیری داده‌های مورداستفاده در این پژوهش از خزانه اصلی داده‌ها باشد. پژوهشگر هیچ‌گونه توضیحی و یا توصیفی دراین‌باره از سازمان سنجش دریافت نکرده است. این ساختار ویژه داده‌ها می‌تواند بر نتایج تحلیل مؤثر بوده باشد.

مروری بر برآوردهای پارامتر سؤال منتج از نظریه کلاسیک و IRT حاکی از این بود که اکثریت سؤالات از توان تمیز متوسط برخوردارند و این سؤالات عموماً دارای دشواری متوسط و بالا هستند. این توزیع پارامترهای سؤال منجر به آزمونی نسبتاً دشوار می‌شود. اصول آزمون‌سازی پیشنهاد می‌کند که سؤالات آزمون از آسان به دشوار مرتب شوند تا از خستگی آزمودنی، داده‌های بدون پاسخ در اثر خستگی و یا از دست دادن انگیزه در پاسخ به سؤالات جلوگیری شود (مگنوسون، ۱۹۶۶). بررسی نحوهٔ توزیع سؤالات در دروس آزمون نشانگر این است که تنها در مورد این روند نسبتاً مشاهده می‌شود و در سایر موارد روند دشواری سؤالات معکوس یعنی از دشوار به آسان، یکنواخت و یا آسان-دشوار-

آسان است. بررسی ویژگی‌های آزمون نشان می‌دهد که داده‌ها احتمالاً از کیفیت مطلوب نظری برخوردار نیستند، لذا در تفسیر نتایج باید دقت لازم را بکار برد.

بررسی تأثیر عوامل مختلف بر برآورد پارامتر در مدل‌های آماری نیازمند داشتن مبنایی برای مقایسه است. ازآنجاکه هدف این پژوهش بررسی تأثیر در نظر گرفتن و یا نگرفتن ساختار سلسله مراتبی داده‌ها بر توزیع پارامتر توانایی و متعاقباً، رتبه‌دهی آزمودنی‌ها بود و با توجه به این‌که تنها داده‌های مشاهده‌شده در دسترس هست، لذا یک مطالعه مقدماتی شبیه‌سازی انجام گرفت. هدف از مطالعه شبیه‌سازی ایجاد مبنایی برای مقایسه بود. در طراحی مطالعه شبیه‌سازی تلاش شد تا حد امکان خصایص مشابهی با داده‌های مشاهده‌شده بکار گرفته شود.

نتایج مطالعه شبیه‌سازی حاکی از این بود که هر دو تحلیل سلسله‌مراتبی و غیرسلسله‌مراتبی توزیع توانایی را در حاشیه‌های توزیع کم‌برآورد و در میانه توزیع بیش‌برآورد کردند. چنین روندی برای کل نمونه و مناطق سه‌گانه مشاهده شد. هر دو تحلیل سلسله‌مراتبی و غیرسلسله‌مراتبی نتایج نسبتاً مشابهی تولید کردند. این نتیجه در محاسبه رتبه‌ها نیز مشاهده شد. علیرغم این‌که تحلیل سلسله‌مراتبی برآوردهای بهتری از داده‌های تولید شده به‌دست داد، اما به لحاظ عملی تفاوت‌ها قابل‌اغماض بود. این امر ناشی از مقدار بسیار پایین همبستگی درون طبقه‌ای (ICC) است که در داده‌های مشاهده‌شده حدود ۴٪ بود. این بدان معناست که تنها ۴٪ واریانس مشاهده‌شده ناشی از تفاوت بین مناطق سه‌گانه است؛ به‌عبارت‌دیگر، در نظر نگرفتن ساختار سلسله مراتبی در تحلیل این داده‌ها منجر به به‌کارگیری ۹۶٪ واریانس کل به‌عنوان کل واریانس مشاهده‌شده می‌گردد. چنین انتخابی منتهی به برآورد اریب از خطای استاندارد برآورد[1] می‌گردد. در حالت کلی، خطای استاندارد برآورد پارامتر در سطح خوشه (مثلاً متوسط توانایی در سطح منطقه) متناسب است با (بلوم[2]، ۲۰۰۵):

1. Standard Error Of Estimate
2. Bloom

$$\sqrt{\frac{\tau^2}{j}+\frac{\sigma^2}{n_j}}$$

در رابطه فوق، τ^2 و σ^2 به ترتیب عبارت‌اند از واریانس متغیر وابسته در سطح دوم (مثلاً مناطق) و واریانس متغیر در سطح اول (مثلاً آزمودنی‌ها). j تعداد خوشه‌ها (مثلاً تعداد مناطق)و n_j حجم نمونه در خوشه j. خطای استاندارد برآورد پارامتر در سطح اول نیز متناسب است با (بلوم، ۲۰۰۵):

$$\sqrt{\frac{\tau^2}{n_j}+\frac{\sigma^2}{n_j}}$$

بنابراین در هر دو سطح تحلیل، خطای استاندارد برآورد پارامتر تحت تأثیر ساختار سلسله‌مراتبی داده‌هاست. مقایسه این دو رابطه نشان می‌دهد که تنها در حالتی که واریانس متغیر در سطح دوم برابر با صفر است، این روابط مقادیر یکسانی دارند. در غیر این‌صورت، مقدار رابطه اول بزرگ‌تر از رابطه دوم خواهد بود. از طرفی دیگر، مقدار ICC برابر است با واریانس متغیر وابسته در سطح دوم تقسیم بر مجموع دو واریانس. این بدین معناست که رابطه‌ای مستقیم بین مقدار ICC و خطای استاندارد برآورد وجود دارد، بدین‌صورت که با افزایش ICC مقدار خطای استاندارد برآورد افزایش پیدا می‌کند. در نظر نگرفتن ساختار سلسله مراتبی هنگامی‌که ICC مقدار قابل‌توجهی دارد تنها منجر به برآوردهای اریب و غیردقیق از پارامترها می‌شود (یونسی، اسکندری، دلاور، فلسفی‌نژاد و فرخی، ۱۳۹۳).

در تحلیل داده‌های مشاهده‌شده نیز، همان‌طور که انتظار می‌رفت، تفاوت چشمگیری در استفاده از تحلیل سلسله مراتبی در مقایسه با تحلیل غیرسلسله‌مراتبی دیده نشد. توزیع پارامتر توانایی تفاوت‌هایی برای مقادیر بیشتر و کمتر از صفر نشان داد. این تفاوت در چگالی توزیع باعث می‌شود که تعداد افراد متفاوتی در یک بازه مشخص توانایی قرار بگیرند و این تفاوت بر رتبه‌بندی آزمودنی‌ها تأثیر می‌گذارد. تأثیر این تفاوت را می‌توان به‌ویژه در حاشیه‌های توزیع که مقادیر کمترین و بیشترین رتبه را در بر می‌گیرد مشاهده کرد. مقایسه بین نتایج تحلیل‌ها نشان داد که تفاوت معناداری در رتبه‌بندی آزمودنی‌ها در

کل نمونه با استفاده از نمرهٔ کل آزمون، تحلیل چندسطحی و تحلیل غیر چندسحطی وجود دارد. بیشترین تفاوت معنادار بین رتبه‌بندی براساس نمرهٔ کل آزمون و روش چندسطحی مشاهده شد. نکته جالب‌توجه اینجاست که وقتی تفاوت رتبه‌دهی براساس نمرهٔ کل آزمون، تحلیل چندسطحی و تحلیل غیر چندسحطی درون سهمیه موردبررسی قرار گرفت، تفاوت معناداری بین سه روش مشاهده نشد. این بدین معناست که استفاده از روش‌های مختلف تحلیل بر رتبه کل آزمودنی تأثیر دارد و این تأثیر ناشی از تفاوت بین مناطق سه‌گانه یا خوشه‌هاست.

به‌طور خلاصه، می‌توان چنین نتیجه‌گیری کرد که تفاوت بین مناطق سه‌گانه و یا هر نوع خوشه‌بندی در داده‌ها هنگامی در نتایج تأثیر دارد که تحلیل در سطح کل نمونه و ورای خوشه‌ها صورت می‌گیرد. در چنین وضعیتی، امکان مشاهده تفاوت معنادار بین رتبه‌دهی با استفاده از روش‌های متفاوت، حتی زمانی که مقدار ICC بسیار پایین باشد (مانند تحلیل حاضر)، وجود دارد. تصمیم‌گیری در مورد این‌که آیا چنین تفاوت معناداری باید هنگام تحلیل داده‌ها و محاسبه رتبه آزمودنی‌ها مدنظر قرار گیرد، بستگی به عوامل مختلفی ازجمله تصمیم‌ها و سیاست‌گذاری‌ها دارد.

مهم‌ترین محدودیت این پژوهش ناظر به تحلیل بخش بسیار کوچکی از داده‌های آزمون است. استفاده از کل داده‌های موجود می‌تواند تصویری بهتر و دقیق‌تر از تأثیر در نظر گرفتن و یا نگرفتن ساختار سلسله مراتبی داده‌ها به‌دست دهد. محدودیت دیگر، ناظر به استفاده از پاسخ‌های مشاهده‌شده و نمره خام کل در تحلیل‌هاست؛ به‌عبارت‌دیگر ضریب دروس مختلف، سهم پاسخ غلط در نمره کل و سهم نمرهٔ پیشرفت تحصیلی در تحلیل‌های این پژوهش مدنظر قرار نگرفت. رتبه آزمودنی‌ها براساس نمونه موردمطالعه و نه براساس رتبه نهایی منتج از رتبه‌بندی کل نمونه آزمودنی‌ها صورت پذیرفت. ضمناً، در مطالعه شبیه‌سازی تنها یک‌بار داده‌های حقیقی تولید و مورد تحلیل قرار گرفت. اگرچه حجم نمونه نسبتاً مناسب بود، اما مطلوب این است که نتایج شبیه‌سازی مبتنی برای تولید و تحلیل

چندباره داده‌ها با استفاده از تکرار[1] باشد. این امر باعث لحاظ کردن خطای نمونه‌گیری خواهد شد.

با توجه به اثر تجمعی نمرات پیشرفت تحصیلی و کل آزمون در نمره کل نهایی آزمودنی، پیشنهاد می‌شود تا تحلیلی مشابه به‌طور مجزا برای این دو نمره و ترکیب آن‌ها صورت گیرد. نتایج چنین تحلیلی می‌تواند اطلاعاتی بیشتری درباره منشأ اثر احتمالی طبقه‌بندی به‌دست دهد. با توجه به توزیع متفاوت پارامترهای سؤال در دروس مختلف و نیز اثر تجمعی نمرات حاصل از دروس مختلف، بررسی تأثیر ساختار سلسله مراتبی بر اساس دروس مختلف می‌تواند منجر به نتایجی تحلیلی‌تر گردد. در این پژوهش تنها بر پارامتر توانایی تمرکز شد. پژوهش‌های متعددی نشان داده‌اند که ساختار سلسله مراتبی داده‌ها می‌تواند بر برآورد پارامتر سؤال تأثیر داشته باشد. چنانچه پارامترهای سؤال دارای مقادیر متفاوتی برای طبقات مختلف باشد، تفاوت توزیع توانایی بین طبقات قابل‌انتظار است. همچنین پیشنهاد می‌شود مطالعه شبیه‌سازی مجدداً و با کنترل عوامل مختلفی همچون حجم نمونه، ICC، توزیع‌های مختلف توانایی و پارامتر سؤال و نیز با تکرار انجام پذیرد.

ORCID
Jalil Younesi https://orcid.org/ 0000-0002-9619-1600
Seyed Amin Mousavi https://orcid.org/0000-0002-6920-2319

1. .replication

منابع

پیک سنجش (۱۳۹۳ الف). *راهنمای داوطلبان برای انتخاب رشته و ضوابط پذیرش دانشجو در آزمون سراسری ۹۳*، سال نوزدهم، شماره ۱۴.

پیک سنجش (۱۳۹۳ ب). *نمره کل آزمون چگونه ساخته می‌شود؟* سال نوزدهم، شماره ۱۶.

تیلور، کاترین، اس. (۲۰۱۳). روایی و رواسازی. ترجمه یونسی (۱۳۹۸). انتشارات دانشگاه علامه طباطبائی.

خدایی، ابراهیم (۱۳۹۳). *توضیحات رئیس سازمان سنجش درباره سهمیه‌ها*. https://www.tabnak.ir/fa/news/

شارع‌پور، محمود (۱۳۸۶). *جامعه‌شناسی آموزش‌وپرورش*. تهران: انتشارات سمت.

فاطمه اسلامی هرندی، فریبا کریمی، محمدعلی نادی (۱۳۹۸). شناسایی مؤلفه‌های عدالت آموزشی در آموزش‌وپرورش ایران، فصلنامه پژوهش در نظام‌های آموزشی، ۱۳(۴۷)، ۵۵-۷۵.

لرد، فردریک، ام. (۱۹۸۰). کاربردهای نظریه سؤال- پاسخ. ترجمه دلاور و یونسی (۱۳۹۱). انتشارات رشد.

مگنوسون، داوید. (۱۹۶۶). *مبانی نظری آزمون‌های روانی*. ترجمهٔ محمد نقی براهنی (۱۳۷۰)، انتشارات دانشگاه تهران.

یونسی، جلیل اسکندری، فرزاد، دلاور، علی، فلسفی‌نژاد، محمدرضا، فرخی، نورعلی (۱۳۹۳). مقایسه توانمندی رویکرد بیزی مدلIRT چندسطحی و مدل کلاسیک چندسطحی: تحلیل داده‌های آزمون فیزیک تیمز پیشرفته (۲۰۰۸)، فصلنامه اندازه‌گیری تربیتی، ۵ (۱۵)، ۱۶۶-۱۸۶.

یونسی، جلیل، دلاور، علی، اسکندری، فرزاد، فلسفی‌نژاد، محمدرضا، فرخی، نورعلی (۱۳۹۳). توانمندی رویکرد بیزی مدل چندسطحی: تحلیل داده‌های آزمون ریاضیات تیمز پیشرفته (۲۰۰۸)، فصلنامه پژوهش در نظام‌های آموزشی، ۸ (۲۴)، ۲۹۵.

Bennett, C. Both, C. & Yeadle, S. (2001). *Mainstreaming equality in the committees of the Scottish parliament*. University of strathclyde.

Beretvas, S. N. & Kamata, A. (2005). The multilevel measurement models: Introduction to the special issue. *Journal of Applied Measurement*, 6, 247-254.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*.

Brennan, R. L (2006). (Ed.). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Embereston, Susan E. Reise, Steven P. (2000). *Item Response Theory for psychologists*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey. London.

Feinberg, R. A. & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement*: Issues and Practice, 35(2), 36-49.

Feldman, S. (2001). Closing the Achievement Gap. *American Educator, 25*(3), 7-9.

Fox, J. P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Fox, J.P. (2001). Multilevel IRT: A Bayesian Perspective on Estimating Parameters and Testing Statistical Hypotheses. PhD dissertation, University of Twente, Faculty of Behavioural Sciences.

Ganzeboom, H.B.G. Treiman, D.J. and Ultee, W. (1991). Comparative intergenerational stratification research: Three generations and beyond, *Annual Review of Sociology*, 17: 277-302.

Gee, J. P. (2008). A sociocultural perspective on opportunity to learn. *Assessment, equity, and opportunity to learn*, 76-108.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. London: Hodder Arnold.

Golstein, H. (1999). *Multilevel Statistical Models*. London: Institute of Education, Multilevel Models Project.

Herman, J. L. Klein, D. C. & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice*, *19*(4), 16-24.

Kamata, A. Bauer, D.J. & Miyazaki, Y. (2008). Multilevel measurement modeling. In A.A. O'Connell & D.B. McCoach (Eds.) *Multilevel Modeling of Educational Data* (pp. 345-388). Charlotte, NC: Information Age Publishing.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th Ed.) (pp. 17-64). Westport, CT: Praeger.

Kerckhoff, A.C. (1995). Institutional arrangements and stratification processes in industrial societies, *Annual Review of Sociology*, 15: 323-347.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: Macmillan.

Mousavi, A. (2013, April). *Analyzing data from educational surveys: a comparison of HLM and Multilevel IRT*, Paper presented at the annual

conference of National council on Measurement in Education. San Francisco. USA.

Muthén, L. K. & Muthén, B. O. (2012). Mplus 7.1 for Windows. Los Angeles, CA: Muthén & Muthén.

New York, NY: Russell Sage Foundation.

Nunnally. J. C. & Burnstein (1994). Psychometric Theory. Mc Graw- Hill book Co.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, CA: Sage.

Rutkowski, L. Gonzalez, E. Joncas, M. & von Davier, M. (2010). International large-scale assessment data issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142-151.

Schmidt, W.H. & Maier, A. (2009). Opportunity to Learn. In Sykes, G. Schneider, B. & Plank, D.N (Eds.), *Handbook of Educational Policy* (p. 541-559). New York, NY: Routledge.

Snijders T. A. B. & Bosker R. J. (2000). *Multilevel analysis. An introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Vartanian, T. P. (2010). *Secondary data analysis*. Oxford University Press.

Zimowski, M. F. Muraki, E. Mislevy, R. J. and Bock, R. D. (1996). *BILOG MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items* [computer software]. Chicago, IL: Scientific Software International.

استناد به این مقاله: یونسی، جلیل.، موسوی، سیدامین. (۱۴۰۰). تأثیر منطقه بندی بر برآورد پارامتر توانایی و رتبه‌بندی آزمودنی‌ها در آزمون‌های بزرگ‌مقیاس، *فصلنامه روان‌شناسی تربیتی* ، ۱۷(۶۰)، ۱-۴۰.

DOI: 10.22054/JEP.2022.63068.3450