

## ویژگی‌های روان‌سنجی امتحانات نهایی سال سوم متوسطه و قابلیت آن‌ها در گزینش داوطلبان ورود به دوره‌های کارشناسی

محمد رضا فلسفی نژاد<sup>۱</sup>، نورعلی فرخی<sup>۲</sup>، لیلا بهرامی<sup>۳</sup>

تاریخ دریافت: ۹۴/۱۱/۱۳

تاریخ پذیرش: ۹۵/۰۷/۱۵

### چکیده

امتحانات نهایی مقطع متوسطه یکی از سرنوشت‌سازترین ابزارهای سنجش علمی دانش‌آموزان است. با توجه به اهمیت این امتحانات، انجام تحقیقات نظام‌مند در مورد کیفیت و قابلیت سؤالات آن‌ها در تفکیک داوطلبان لازم و ضروری است. هدف از پژوهش حاضر تعیین ویژگی‌های روان‌سنجی و قابلیت سؤالات امتحانات نهایی رشته‌ی تجربی در دو درس زیست‌شناسی و ادبیات فارسی در گزینش داوطلبان ورود به دوره‌های کارشناسی بود. روش: جامعه آماری، تمام سؤالات امتحانات نهایی رشته‌ی علوم تجربی سال سوم مقطع متوسطه در خردادماه ۱۳۹۰ بوده است که برای تعیین ویژگی‌های روان‌سنجی آن‌ها از عملکرد ۶۰۰ دانش‌آموز ناحیه یک شهرستان خرم‌آباد در دو درس مذکور که به صورت تصادفی انتخاب شده بودند، استفاده شد. ضریب اعتبار آزمون‌های زیست‌شناسی و ادبیات فارسی با استفاده از آلفای کرونباخ (۰/۹۷)، تعیین گردید. بر اساس CTT، میانگین ضرایب دشواری و تمیز سؤالات در آزمون‌های ادبیات فارسی (۰/۹۶ و ۰/۶۵) و زیست‌شناسی (۰/۵۷ و ۰/۶۵) به دست آمد. در تحلیل IRT؛ مدل دو پارامتری برازش بیشتری با داده‌ها نشان داد. همچنین، میانگین دشواری و قدرت تشخیص سؤالات در مدل‌های IRT در

۱. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی

۲. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی

۳. کارشناس ارشد سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی (نویسنده مسئول)

bahrami.leala@yahoo.com

دروس ادبیات فارسی (۰/۶۹، ۱/۰۳) و زیست‌شناسی (۰/۰۹، ۰/۹۶) به دست آمد. بیشترین میزان آگاهی - دهندگی دو آزمون ادبیات فارسی و زیست‌شناسی، به ترتیب متعلق به سطوح توانایی (۰/۷، ۰/۱) است و میزان توافق CTT و IRT از نظر قدرت تشخیص در هر دو درس، به ترتیب (۰/۹۸/۳۶، ۰/۹۳/۵۹) بود. با توجه به خطیر بودن تصمیمات مبتنی بر آزمون‌های سرنوشت‌ساز، تلویحات استفاده از امتحانات نهایی در گزینش داوطلبان مورد بحث قرار گرفت.

**واژگان کلیدی:** امتحانات نهایی، پارامتر تشخیص، پارامتر دشواری، نظریه پرسش پاسخ، نظریه کلاسیک آزمون

## مقدمه

گزینش دانشجو و شناسایی ملاک‌های دقیق و معتبر آن، از چالش‌های جدی نظام آموزش عالی در جوامع است. با اینکه فرآیند گزینش دانشجو در جوامع مختلف، به شکل‌های متفاوتی صورت می‌پذیرد، اما در این میان سعی همه جوامع بر این است که امکانات آموزش عالی به شایسته‌ترین افراد تخصیص یابد. نظام پذیرش دانشجو در کشورمان از طریق کنکور صورت می‌گیرد که در بین داوطلبان ورود به نظام آموزش عالی به معنای «سرنوشت» تعبیر می‌شود و در کشور باعث بحران شده است.

از تحولات سال‌های اخیر در زمینه نظام پذیرش دانشجو، مطرح کردن جایگزینی سوابق تحصیلی بجای کنکور است که در سال ۱۳۸۶ با عنوان قانون حذف کنکور به تصویب مجلس شورای اسلامی رسید و می‌بایست از سال ۹۰ اجرایی می‌شد که با توجه به مشکلات علمی، فنی و اجرایی به سرانجام نرسید. تا اینکه در مردادماه ۱۳۹۲ کلیات طرح جدید سنجش و پذیرش دانشجو به تصویب نمایندگان مجلس شورای اسلامی رسید. بر اساس آن، تأثیر سوابق تحصیلی (دور دوم دوره متوسطه) از سال ۹۳ به صورت پلکانی از تأثیر ۲۵ درصد شروع شده تا در نهایت به ۸۵ درصد طی ۵ سال برسد. با توجه به طرح جدید سنجش و پذیرش دانشجو، از امتحانات نهایی مقطع متوسطه می‌توان به‌عنوان یکی از مهم‌ترین و سرنوشت‌سازترین ابزار سنجش علمی دانش آموزان نام برد.

امتحانات نهایی، ماهیتی ملاک مرجع دارند که با هدف میزان حصول اهداف آموزشی دوره متوسطه تنظیم شده‌اند. سؤالات این امتحانات، معمولاً به گونه‌ای طراحی می‌شود که بیشترین میزان آگاهی را در محدوده‌ی نمره‌ی ملاک داشته باشند. با اینکه این امتحانات انواع گوناگون سؤال را شامل می‌شوند، ولی درصد بیشتری از سؤالات آن‌ها از نوع تشریحی و کوتاه پاسخ است. در بسیاری از تحقیقات، آزمون‌های عینی مورد تجزیه و تحلیل قرار گرفته‌اند و این خود گامی برای بهبود کیفیت سؤالات این آزمون‌هاست؛ اما به آزمون‌های ذهنی و روش‌های تحلیل آن‌ها کمتر توجه شده است و این در شرایطی است که با جایگزینی سوابق تحصیلی که در چند سال اخیر مطرح شده، بیش‌ازپیش بر اهمیت امتحانات نهایی مقطع متوسطه افزوده شده است.

همان‌طور که حسنی و کیامنش (۱۳۸۸) بیان کرده‌اند؛ تحلیل آیین‌نامه‌های امتحانات نشان می‌دهد در طول هشتاد و اندی سال که از تدوین اولین آیین‌نامه‌های امتحانات در شورای عالی معارف می‌گذرد، تدوین‌کنندگان آیین‌نامه‌ها کمتر به صحت و دقت سؤالات و نتایج ارزشیابی‌ها توجه داشته‌اند و بیشترین دغدغه آن‌ها صحت اجرا و ثبت نمره بوده است و سازوکار امتحانات نهایی، علیرغم تشکیلات گسترده و پرهزینه آن برای ارزشیابی آموزشی، داده‌های مناسبی فراهم نمی‌کند، چون در خصوص اعتبار و روایی سؤالات این امتحانات جای تردید وجود دارد. لذا مستندسازی کیفیت سؤالات امتحانات نهایی و تعیین قابلیت آن‌ها در شناسایی و تمایز گذاری میان داوطلبان کنکور یک ضرورت مهم به شمار می‌آید. از طرفی معمولاً در فرآیند بررسی کیفیت سؤالات این امتحانات، از مدل‌های سنتی سنجش استفاده می‌شود که مشکلات فنی و نظری شناخته‌شده‌ای دارند. در قسمت زیر، به مواردی از تحقیقات انجام‌شده بر اساس مدل کلاسیک، اشاره می‌شود.

رحیمی (۱۳۸۷) در مطالعه‌ای به بررسی شاخص‌های روان‌سنجی سؤالات امتحانات نهایی پایه سوم راهنمایی خردادماه ۱۳۸۶ در استان خوزستان پرداخت و به این نتیجه رسید

---

#### 1. reliability & validity

که در همه‌ی دروس، اکثر سؤالات در حد مطلوب و متوسط طرح شده‌اند. همچنین، ضریب تمیز سؤالات مطلوب بوده است.

سلیمانی (۱۳۸۴) به بررسی شاخص‌های سؤال و رعایت اصول علمی طراحی سؤالات امتحانی دروس ریاضی در پایه‌های اول، دوم و سوم مقطع متوسطه نظری مراکز آموزشی استان کرمانشاه در امتحانات پایانی خردادماه ۸۳ پرداخت. به این نتیجه رسید که هرچند سؤالات، ضریب دشواری متوسطی دارند ولی از ضریب تشخیص پایینی برخوردار می‌باشند. حشمتی، رافع و جمشیدی نژاد (۱۳۸۳) در مطالعه‌ای با هدف تعیین میزان اعتبار آزمون‌ها، ضریب تشخیص و سطح دشواری سؤالات به تحلیل نتایج امتحانات نهایی سال سوم متوسطه نظری خرداد ۸۱ در تمام استان‌ها پرداختند. به این نتایج رسیدند که در مجموع، اعتبار (محاسبه با استفاده از آلفای کرونباخ) آزمون‌های مورد مطالعه قابل توجه بوده است. همچنین، این امتحانات در تفکیک دانش‌آموزان قوی و ضعیف از توانایی قابل توجهی برخوردار نبوده و ضریب دشواری اکثر سؤالات در دامنه بهینه قرار دارد.

با توجه به ادبیات تحقیق، تحقیقاتی که بر روی کیفیت امتحانات نهایی بر اساس روش‌های جدید اندازه‌گیری صورت گرفته باشد، وجود نداشته و یا خیلی اندک است. از سوی دیگر، ساخت آزمون و تحلیل سؤالات آن‌ها بر پایه مدل‌های جدید اندازه‌گیری، از جمله فعالیت‌های اساسی در آموزش و پرورش است که متأسفانه این مهم برای خیلی از متخصصان این عرصه ناشناخته است. این در حالی است که شرکت‌های آزمون‌سازی در خارج از کشور، به منظور ارتقای کیفی آزمون‌های خود، به‌طور پیوسته آن‌ها را مورد تحلیل روان‌سنجی قرار می‌دهند. نظریه‌ی کلاسیک آزمون (CTT)<sup>۱</sup> و نظریه‌ی پرسش پاسخ (IRT)<sup>۲</sup> اغلب برای ارزیابی ویژگی‌های روان‌سنجی آزمون‌های روانی و تربیتی استفاده می‌شوند که در این میان، همان‌طور که سوئن و لی<sup>۳</sup> (۲۰۰۷) مطرح کرده‌اند، IRT به دلیل کاربردهای بسیار برجسته آن و برنامه‌های سنجش در مقیاس گسترده، اخیراً مرکز توجه

- 
1. classical test theory
  2. item response theory
  3. Suen & Lei

بسیاری از روان‌سنجان و متخصصان آزمون بوده است. علیرغم توجهات اخیر که به این نظریه شده است، امروزه CTT برای اکثر آزمون‌های شغلی و دانشگاهی و ابزارهای اندازه‌گیری، به‌خصوص در برنامه‌های سنجش در مقیاس محدود همچنان پر استفاده برجا مانده است که علت آن را می‌توان در مزیت‌هایی که ون‌درالست<sup>۱</sup> (۲۰۱۳) برای مدل‌های کلاسیک برمی‌شمارد جستجو کرد؛ برای بیشتر محققان شناخته شده‌اند، مفروضات نظری نسبتاً ضعیفی را دارند، روش‌های ریاضی ساده‌ای را به کار می‌برند و آن‌ها به نمونه‌ی نسبتاً کوچکی از آزمودنی‌ها جهت برآورد پارامترهای سؤال نیاز دارند. برنان<sup>۲</sup> (۲۰۱۱) متذکر می‌شود؛ مشخصه‌های آماری تحلیل کلاسیک سؤال، سودمندی طولانی مدت ثابت شده‌ای برای توسعه آزمون دارند، اما به این معنی نیست که آن‌ها توسط CTT به خوبی مدل‌سازی شده‌اند. همبلتون و ون‌درلیندن<sup>۳</sup> (۱۹۸۲) مدل کلاسیک آزمون را مدل ضعیفی می‌دانند و از جمله معایبی که برای آن عنوان می‌کنند؛ نمرات آزمون آزمودنی وابسته به آزمون‌اند، پارامترهای سؤال و آزمون به نمونه همبسته‌اند و دسترسی به آزمون‌های موازی ضرورتی است که در عمل به دشواری به دست می‌آید.

استینی و نرینگ<sup>۴</sup> (۲۰۰۶) نیز، مزیت اصلی IRT را چنین مطرح می‌کنند؛ دشواری سؤال و سطح صفت شخص (توانایی) بر روی مقیاس یکسانی قرار دارند؛ بنابراین وقتی توانایی شخص بر پیوستار توانایی بالاتر از دشواری سؤال قرار گیرد، به احتمال قوی شخص به سؤال پاسخ صحیح می‌دهد و بالعکس. با وجود مزیت‌ها و کاربردهای بسیار مدل‌های IRT (در زمینه‌هایی از جمله توسعه‌ی آزمون‌ها، همترازسازی نمرات به دست آمده از آزمون‌های غیر موازی، بررسی سوگیری سؤال، ...) معایبی نیز برای آن بر شمرده‌اند. از جمله اینکه برای به دست آوردن برآورد دقیق پارامترهای مدل، نیاز به حجم نمونه بزرگ دارند. این در حالی است که در بسیاری از زمینه‌های آموزشی، اندازه نمونه بزرگ غیرقابل دسترس و نامطلوب

---

4. Van der Elst

1. Brennan

2. Hambleton & Vander Linden

3. Ostini & Nering

است (فولی، ۲۰۱۰). برنان (۲۰۱۱) متذکر می‌شود که تفاوت‌های میان مدل‌ها بنیادی و مهم هستند، اما هریک از این مدل‌ها قابل دفاع و سودمند هستند، و هیچ‌یک از آن‌ها جایگزینی برای دیگری نیست.

در راستای مباحث مذکور، این تحقیق می‌کوشد ویژگی‌های روان‌سنجی سؤالات امتحانات نهایی را بر اساس مدل کلاسیک آزمون در کنار مدل‌های پرسش پاسخ موردبررسی قرار دهد. همچنین، قابلیت این سؤالات را در گزینش داوطلبان ورود به دوره‌های کارشناسی مؤسسات و مراکز آموزش عالی کشور تعیین نماید. تحلیل سؤالات با استفاده از IRT می‌تواند به افزایش دانش متخصصان تعلیم و تربیت در حیطه سنجش کمک نماید و کیفیت سؤالات را ارتقاء بخشد. همچنین، می‌توان زیرمجموعه‌ای از سؤالات را برای اهداف و نمونه‌های خاصی از آزمودنی‌ها، برای آزمون‌های آینده انتخاب نمود که این امر به ایجاد بانک سؤالات استاندارد ادبیات فارسی و زیست‌شناسی کمک خواهد نمود.

در راستای نیل به اهداف، این تحقیق درصدد است تا به این سؤالات پاسخ دهد:

- آیا سؤالات امتحانات نهایی بر اساس CTT از ویژگی‌های روان‌سنجی مطلوب برخوردارند؟
- آیا سؤالات امتحانات نهایی بر اساس IRT از ویژگی‌های روان‌سنجی مطلوب برخوردارند؟
- میزان توافق CTT و IRT در تعیین سؤالات مناسب و نامناسب از نظر قدرت تشخیص به چه صورت است؟

## روش

تحقیق حاضر از نوع توصیفی است. جامعه‌ی آماری، تمام سؤالات امتحانات نهایی رشته‌ی علوم تجربی سال سوم مقطع متوسطه در خردادماه ۱۳۹۰ بوده است که جهت بررسی ویژگی‌های روان‌سنجی این سؤالات، عملکرد دانش‌آموزان ناحیه یک شهرستان خرم‌آباد

مورد تجزیه و تحلیل قرار گرفت. تعداد دانش‌آموزان سال سوم مقطع متوسطه شاخه نظری در سال تحصیلی ۹۰-۸۹ در ناحیه یک شهرستان خرم‌آباد ۳۳۰۶ نفر بود که از بین آن‌ها ۱۴۸۸ نفر معادل ۴۵ درصد، در رشته‌ی علوم تجربی مشغول به تحصیل بوده‌اند.

از آنجا که در میان دروس عمومی و اختصاصی رشته‌ی علوم تجربی، دو درس زیست‌شناسی و ادبیات فارسی ضرایب بالایی دارند، در بین سایر دروس انتخاب شدند. با توجه به عملی بودن بررسی یک‌یک سؤالات دو درس مذکور، ضرورتی برای نمونه‌گیری از سؤالات وجود نداشت و کلیه سؤالات این دو درس مورد تحلیل قرار گرفت. بدین منظور از بسته‌های اوراق امتحانی دروس ادبیات فارسی و زیست‌شناسی موجود در انبار امتحانات که متعلق به تمام دبیرستان‌های ناحیه یک شهرستان خرم‌آباد؛ اعم از استعداد‌های درخشان، عادی - دولتی، نمونه دولتی، غیرانتفاعی و بزرگ‌سالان بود، استفاده گردید. با توجه به اینکه نظریه‌های روان‌سنجی، نظریه‌های نمونه‌های بزرگ هستند و با در نظر گرفتن ملاحظات نظری و عملی این نظریه‌ها و برای تحقق برآورد باثبات پارامترها، نمونه‌ای با حجم ۶۰۰ ورقه‌ی امتحانی در دو درس مذکور به‌طور تصادفی انتخاب شد.

ابزار جمع‌آوری داده‌ها، امتحانات نهایی ادبیات فارسی و زیست‌شناسی رشته‌ی علوم تجربی سال سوم مقطع متوسطه خرداد ۹۰ بود که به‌طور همزمان در سراسر کشور برگزار شد و تعداد سؤالات آن‌ها به ترتیب ۵۳ و ۳۰ سؤال بوده است که با احتساب ریز سؤالات، این تعداد به ۶۱ و ۷۸ سؤال رسید. به‌عنوان نمونه، سؤال ۴۷ آزمون ادبیات فارسی دارای سه قسمت الف، ب، پ و سؤال ۱۵ آزمون زیست‌شناسی شامل ۴ قسمت الف، ب، ج، د است که با توجه به اینکه هر کدام از آن‌ها مفاهیم مستقلی را می‌سنجند، به‌عنوان سؤال جداگانه در نظر گرفته شده‌اند. همچنین در تعداد کمی از سؤالات، هر بخش از سؤال که به آن ریز بارم تعلق گرفته بود، به‌عنوان سؤال جداگانه در نظر گرفته شد.

جمع‌آوری داده‌ها پس از کسب مجوزهای لازم در انبار حوزه تصحیح امتحانات آموزش و پرورش ناحیه یک شهرستان خرم‌آباد صورت گرفت. برای هر دانش‌آموز و در هر امتحان، نمره‌ی تک‌تک سؤالات و همچنین ریزبارم هر سؤال (در صورت موجود بودن) که توسط مصححان مختلف داده شده بود به‌عنوان داده‌های تحقیق در فرم‌های مخصوص محقق

ساخته ثبت شد و سپس در نرم‌افزار SPSS وارد گردید. برای آماده‌سازی داده‌ها جهت انجام تحلیل کلاسیک، میانگین نمراتی که مصححان مختلف به هر سؤال داده بودند، محاسبه و جایگزین نمره‌ی فرد در هر سؤال شد. همچنین، برای آماده‌سازی داده‌ها جهت انجام تحلیل IRT، بعد از محاسبه میانگین نمرات داده‌شده توسط مصححان به هر سؤال، نمره‌ی هر سؤال بر اساس این قرارداد به ۰ و ۱ تبدیل گردید: به سؤالی که نمره‌ی آن برابر با نصف بارم و یا بیشتر از آن بود، یک و به سؤالی که نمره‌ی آن کمتر از نصف بارم بود، صفر تعلق گرفت. جهت محاسبه ویژگی‌های روان‌سنجی سؤالات بر اساس CTT؛ ابتدا ضریب اعتبار کل آزمون با استفاده از فرمول آلفای کرونباخ به دست آمد. سپس، میزان اعتبار آزمون بعد از حذف هر سؤال (روش لوپ) با کمک نرم‌افزار SPSS محاسبه شد. طبق بیان همبلتون و جونز<sup>۱</sup> (۱۹۹۳)، تکنیک‌های تحلیل استاندارد سؤال بر اساس CTT مستلزم یک ارزیابی از شاخص‌های دشواری و تمیز سؤال و گزینه‌های پیشنهادی سؤال است. بر این اساس، شاخص‌های آماری سؤالات (درجه دشواری و ضریب تمیز سؤال) بر اساس داده‌های خام با استفاده از روش نیتکو و با اکسل انجام گرفت.

محاسبه ویژگی‌های روان‌سنجی سؤالات بر اساس IRT به این شرح بود؛ ابتدا مفروضات اولیه و اساسی این نظریه موردبررسی قرار گرفتند. در IRT، معمولاً فرض می‌شود که سؤالات همگن هستند و به بیانی دیگر آن‌ها یک توانایی را اندازه می‌گیرند. طبق گفته‌ی لرد<sup>۲</sup> (۱۹۶۸)، مفروضه تک‌بعده بودن نمی‌تواند به‌طور کامل در بسیاری از آزمون‌ها برآورد گردد. باین‌حال ممکن است در بعضی موارد تقریب نسبتاً خوبی از آن فراهم شود. مفروضه‌ی دوم، استقلال موضعی است که اشاره دارد به اینکه پاسخ‌های آزمودنی به هر سؤال آزمون باید به لحاظ آماری به پاسخش به دیگر سؤالات آزمون نا همبسته باشد (همبلتون و تراوب<sup>۳</sup>، ۱۹۷۰). برای سنجش تک‌بعده بودن آزمون، از نمودار اسکری استفاده گردید. طبق گفته‌ی لرد (۱۹۸۰)، وقتی مفروضه تک‌بعده بودن برقرار است، استقلال

- 
1. Jones
  2. lord
  3. Traub



موضوعی نیز حاصل می‌شود از این حیث این دو مفهوم با هم معادل‌اند (به نقل از همبلتون و همکاران، ۱۹۹۱، ترجمه‌ی فلسفی نژاد، ۱۳۸۹).

با توجه به اینکه سؤالات دو آزمون ادبیات فارسی و زیست‌شناسی بیشتر از نوع تشریحی (کوتاه پاسخ) بوده، لذا عامل حدس به‌عنوان یک راهبرد در پاسخ‌گویی به سؤالات نقشی نداشته است. بدین سبب، پس از اطمینان از برقراری مفروضات اصلی، تحلیل سؤالات در هر دو آزمون، با دو مدل یک پارامتری و دو پارامتری و با استفاده از نرم‌افزار BILOGMG3، انجام گرفت. لازم به ذکر است که با توجه به نامعلوم بودن پارامترهای سؤال و توانایی، جهت پرآورد همزمان آن‌ها از روش پیشینه درست‌نمایی حاشیه‌ای استفاده گردیده است. از شاخص RMS برای قضاوت در مورد برازش کلی مدل استفاده شد. همچنین، به‌منظور بررسی اینکه کدام مدل بهترین برازش را با داده‌ها دارد، از آزمون‌های دو استفاده گردید. در ادامه، میزان برازندگی تک‌تک سؤالات با مدل بررسی شد. همچنین، برای بررسی میزان توافق دو مدل کلاسیک و IRT در تشخیص سؤالات مناسب و نامناسب، از پارامتر تشخیص استفاده شد. بر این اساس، سؤالاتی که قدرت تشخیص کمتر از ۰/۲ در CTT و کمتر از ۰/۵ در IRT داشته‌اند به‌عنوان سؤالات نامناسب تعیین شدند.

#### یافته‌ها

##### تحلیل سؤالات امتحانات نهایی بر اساس مدل کلاسیک

درس ادبیات فارسی. ضریب اعتبار کل آزمون با استفاده از فرمول آلفای کرونباخ، (۰/۹۵۶)  $\alpha =$  به دست آمد. محاسبه میزان اعتبار آزمون بعد از حذف هر سؤال، ما را در شناسایی سؤالاتی که نقش مهمی در هماهنگی با کل آزمون ایفا می‌کنند و سؤالاتی که اثری بر اعتبار آزمون ندارند کمک می‌کند. نتایج حاصل از این تحلیل نشان داد با حذف سؤالات ۱، ۲، ۳، ۴، ۱۲، ۲۵، ۲۰، ۱۴، ۴۲، ۴۱، ۳۶ میزان اعتبار تغییر نکرد. با حذف باقی سؤالات، میزان اعتبار آزمون کاهش یافت که نشان می‌دهد این سؤالات با کل آزمون هماهنگی دارند و نقش مؤثری در اعتبار آزمون ایفا می‌کنند.

با توجه به جدول ۱، بیشترین و کمترین ضریب تمیز به ترتیب متعلق به سؤالات ۵۲ و ۴۶ (۰/۸۵، ۰/۰۷) است. سؤال ۲ با ضریب دشواری (۰/۹۱) و سؤال ۴۱ با ضریب دشواری (۰/۳۱) نیز، به ترتیب ساده ترین و دشوارترین سؤالات آزمون ادبیات فارسی هستند. توزیع فراوانی و درصد سؤالات آزمون برحسب ضرایب دشواری و تمیز به دست آمده از مدل کلاسیک اندازه گیری، در جدول ۲ نشان داده شده است. از نظر شاخص دشواری، ۳۷ سؤال معادل (۶۰/۶۶) درصد از سؤالات در دامنه دشواری ۰/۳ تا ۰/۷ قرار دارد و باقی سؤالات دارای دشواری بیشتر از ۰/۷ هستند. یک سؤال دارای ضریب تمیز کمتر از ۰/۲ و ۴ سؤال دارای ضریب تمیز بین ۰/۲ تا ۰/۳ است. باقی سؤالات (معادل ۹۱/۸۰ درصد سؤالات)، ضریب تمیز بیش از ۰/۳ دارند که از این تعداد ۴۴ سؤال (معادل ۷۲/۱۳ درصد سؤالات) دارای ضریب تمیز ۰/۵ و بالاتر از ۰/۵ هستند که بیانگر وضعیت بسیار مطلوبی است. سؤالات ۲، ۳، ۴، ۱۴، ۲۵ و ۴۶ دارای ضریب تمیز کمتر از ۰/۳ هستند و قادر به تفکیک آزمودنی های قوی و ضعیف نبوده و نامناسب اند.

جدول ۱. تحلیل سؤالات آزمون ادبیات فارسی بر اساس مدل کلاسیک و

مدل پرسش پاسخ به همراه مقادیر آگاهی

سؤال	مدل کلاسیک		مدل دو پارامتری			مقادیر پارامتر آگاهی	
	ضریب دشواری	ضریب تمیز	تشخیص (شیب)	خی دو	سطح معناداری	بیشینه آگاهی	تنای ماکسیمم
۱	۰/۶۸	۰/۳۷	۰/۵۳	۱۲/۸	۰/۲۳	۰/۲۰	-۱
۲	۰/۹۱	۰/۲۲	۰/۹۲	۵/۳	۰/۵۱	۰/۶۱	-۲
۳	۰/۸۴	۰/۲۳	۰/۵۳	۱۱/۹	۰/۲۲	۰/۲۰	-۲/۱
۴	۰/۵۸	۰/۵۵	۰/۶۴	۵/۱	۰/۸۳	۰/۲۹	-۰/۴
۵	۰/۷۱	۰/۷۰	۱/۲۳	۷/۸	۰/۳۵	۱/۰۹	-۱
۶	۰/۶۵	۰/۶۵	۱/۱۸	۸/۹	۰/۲۶	۱	-۰/۷
۷	۰/۶۱	۰/۶۲	۱/۱۱	۱/۵	۰/۹۸	۰/۹۰	-۰/۹
۸	۰/۷۵	۰/۶۰	۱/۴۱	۶/۹	۰/۲۳	۱/۴۴	-۱
۹	۰/۸۰	۰/۳۷	۰/۸۷	۷/۶	۰/۴۸	۰/۵۵	-۱/۷
۱۰	۰/۶۵	۰/۸۰	۱/۶۸	۳/۳	۰/۵۰	۲/۰۴	-۰/۷

ویژگی‌های روان‌سنجی امتحانات نهایی سال سوم متوسطه و ... / ۵۵

-۱/۴	۲	۰/۹۴	۱/۳	۱/۶۷	-۱/۳۶	۰/۵۰	۰/۸۰	۱۱
-۱/۲	۰/۲۹	۰/۹۱	۴/۸	۰/۶۴	-۱/۱۶	۰/۴۴	۰/۷۳	۱۲
-۰/۷	۰/۶۹	۰/۳۳	۸/۱	۰/۹۸	-۰/۶۸	۰/۶۲	۰/۶۸	۱۳
-۲/۲	۰/۴۲	۰/۹۹	۱/۴	۰/۷۶	-۲/۱۶	۰/۲۲	۰/۹۰	۱۴
۰/۶	۰/۸۳	۰/۰۲	۱۷/۵	۱/۰۷	۰/۵۷	۰/۷۰	۰/۳۳	۱۵
-۰/۵	۰/۷۴	۰/۶۲	۵/۳	۱/۰۲	-۰/۵۴	۰/۶۶	۰/۶۴	۱۶
-۰/۷	۰/۶۶	۰/۷۶	۴/۲	۰/۹۶	-۰/۷۲	۰/۶۲	۰/۶۸	۱۷
-۰/۸	۰/۷۹	۰/۷۹	۴/۷	۱/۰۵	-۰/۸۱	۰/۵۶	۰/۷۲	۱۸
-۰/۸	۰/۵۴	۰/۱۳	۱۱/۳	۰/۸۷	-۰/۷۶	۰/۵۹	۰/۶۹	۱۹
-۱/۷	۰/۴۰	۰/۵۵	۷/۸	۰/۷۵	-۱/۶۹	۰/۳۳	۰/۸۴	۲۰
-۰/۷	۰/۵۶	۰/۹۰	۳/۵	۰/۸۸	-۰/۶۵	۰/۵۹	۰/۶۶	۲۱
-۰/۶	۰/۷۹	۰/۸۵	۴/۱	۱/۰۵	-۰/۶۴	۰/۶۶	۰/۶۷	۲۲
-۱	۰/۸۴	۰/۹۴	۲/۹	۱/۰۸	-۱/۰۵	۰/۵۲	۰/۷۸	۲۳
-۱/۶	۰/۵۵	۰/۴۵	۶/۸	۰/۸۷	-۱/۵۸	۰/۳۶	۰/۸۵	۲۴
-۱/۷	۰/۵۰	۰/۹۱	۳/۳	۰/۸۴	-۱/۷۴	۰/۲۸	۰/۸۷	۲۵
-۰/۵	۰/۷۶	۰/۹۹	۱/۲	۱/۰۲	-۰/۵۰	۰/۷۱	۰/۵۰	۲۶
-۱/۲	۰/۷۴	۰/۹۶	۲/۴	۱/۰۱	-۱/۲۳	۰/۴۷	۰/۷۸	۲۷
-۰/۵	۰/۷۲	۰/۱۹	۹/۹	۱	-۰/۵۳	۰/۶۴	۰/۶۴	۲۸a
-۰/۶	۱/۰۹	۰/۷۶	۳/۴	۱/۲۳	-۰/۵۷	۰/۶۸	۰/۶۶	۲۸b
-۱/۱	۰/۶۳	۰/۱۸	۱۱/۳	۰/۹۳	-۱/۰۷	۰/۵۴	۰/۷۷	۲۹
-۱/۲	۰/۴۵	۰/۷۱	۶/۳	۰/۷۹	-۱/۲۱	۰/۴۶	۰/۷۷	۳۰
-۰/۵	۰/۳۲	۰/۹۴	۳/۵	۰/۶۷	-۰/۴۸	۰/۵۵	۰/۶۰	۳۱
۰/۶	۰/۵۲	۰/۶۲	۶/۲	۰/۸۵	۰/۶۰	۰/۶۴	۰/۳۴	۳۲
۰/۲	۰/۷۵	۰/۸۲	۴/۴	۱/۰۲	۰/۲۲	۰/۷۵	۰/۴۳	۳۳
-۰/۷	۰/۶۳	۰/۳۷	۸/۷	۰/۹۳	-۰/۶۶	۰/۶۲	۰/۶۷	۳۴
-۰/۸	۰/۹۹	۰/۳۸	۷/۵	۱/۱۷	-۰/۷۵	۰/۶۱	۰/۷۱	۳۵
-۰/۳	۰/۱	۰/۹۲	۴/۵	۰/۴۰	-۰/۳۲	۰/۴۴	۰/۵۵	۳۶
۰/۳	۰/۵۵	۰/۴۰	۸/۳	۰/۸۷	۰/۳۱	۰/۶۶	۰/۴۱	۳۷a
۰/۳	۰/۵۳	۰/۸۹	۳/۶	۰/۸۶	۰/۳۴	۰/۶۹	۰/۴۰	۳۷b
-۰/۶	۰/۵۲	۰/۶۸	۴/۸	۰/۸۵	-۰/۵۶	۰/۶۰	۰/۶۴	۳۸ الف

۰/۲	۰/۳۵	۰/۷۷	۵/۶	۰/۷۰	-۰/۲۳	۰/۶۱	۰/۵۵	۳۸ ب
-۰/۳	۰/۵۰	۰/۷۳	۶/۱	۰/۸۳	-۰/۳۳	۰/۶۵	۰/۵۸	۳۹
-۰/۲	۰/۷۴	۰/۵۳	۸	۱/۰۱	-۰/۲۱	۰/۷۴	۰/۵۵	۴۰
۱	۰/۲۳	۰/۵۸	۷/۶	۰/۵۷	۰/۹۷	۰/۴۲	۰/۳۱	۴۱
-۱/۷	۰/۳۹	۰/۵۰	۸/۴	۰/۷۳	-۱/۶۶	۰/۳۸	۰/۷۶	۴۲
-۱/۲	۰/۶۴	۰/۲۲	۱۰/۷	۰/۹۴	-۱/۲۰	۰/۴۱	۰/۷۹	۴۳ الف
-۱	۰/۴۹	۰/۴۰	۹/۴	۰/۸۲	-۱/۰۳	۰/۴۷	۰/۷۴	۴۳ ب
-۰/۸	۱/۱۵	۰/۴۴	۵/۹	۱/۲۶	-۰/۸۱	۰/۶۰	۰/۷۳	۴۴ الف
-۰/۸	۰/۵۹	۰/۵۶	۶/۷	۰/۹۰	-۰/۸۱	۰/۵۳	۰/۷۰	۴۴ ب
-۰/۷	۱	۰/۲۴	۹/۱	۱/۱۸	-۰/۶۷	۰/۶۷	۰/۶۸	۴۵a
-۰/۹	۰/۸۳	۰/۳۷	۷/۶	۱/۰۷	-۰/۹۵	۰/۵۹	۰/۷۵	۴۵b
۰/۲	۰/۹۶	۰/۰۰	۲۳/۶	۱/۱۵	۰/۲۲	۰/۰۷	۰/۴۱	۴۶
۰/۲	۱/۰۱	۰/۶۳	۵/۲	۱/۱۸	۰/۲۳	۰/۷۸	۰/۴۲	۴۷ الف
-۰/۲	۱/۴۳	۰/۴۳	۵/۹	۱/۴۱	-۰/۲۱	۰/۸۴	۰/۵۵	۴۷ ب
۰/۴	۰/۹۱	۰/۹۰	۳/۵	۱/۱۲	۰/۳۷	۰/۷۶	۰/۳۸	۴۷ پ
-۰/۸	۲/۶۶	۰/۴۰	۴	۱/۹۲	-۰/۸۳	۰/۶۵	۰/۷۶	۴۸
-۰/۵	۱/۵۳	۰/۰۳	۱۲/۲	۱/۴۶	-۰/۵۴	۰/۷۵	۰/۶۶	۴۹
-۰/۵	۱/۷۰	۰/۰۱	۱۶/۱	۱/۵۳	-۰/۴۶	۰/۸۳	۰/۶۳	۵۰
-۰/۳	۱/۴۹	۰/۵۲	۵/۲	۱/۴۴	-۰/۳۲	۰/۸۰	۰/۵۹	۵۱
-۰/۲	۱/۶۷	۰/۶۷	۴/۱	۱/۵۲	-۰/۲۲	۰/۸۵	۰/۵۴	۵۲
-۰/۷	۱/۹۳	۰/۲۷	۶/۴	۱/۶۴	-۰/۷۱	۰/۷۱	۰/۷۱	۵۳
				۱/۰۳	-۰/۶۹	۰/۵۷	۰/۶۵	میانگین
-2 LOG LIKELIHOOD <sub>2pt</sub> = ۳۳۵۳۸/۰۱۹۲								

جدول ۲. توزیع سؤالات ادبیات فارسی بر اساس ضرایب دشواری و تمیز کلاسیک

ضریب دشواری			تمیز کلاسیک
کمتر از ۰/۳	بین ۰/۳ تا ۰/۷	بیشتر از ۰/۷	
۰/۲ کمتر از	۱ (۱/۶۴٪)	-	۰/۲ کمتر از
بین ۰/۲ تا ۰/۳	-	۴ (۶/۵۶٪)	بین ۰/۲ تا ۰/۳
بیشتر از ۰/۳	۳۶ (۵۹/۰۲٪)	۲۰ (۳۲/۷۹٪)	بیشتر از ۰/۳

درس زیست‌شناسی. ضریب اعتبار کل آزمون با استفاده از فرمول آلفای کرونباخ،  $(0/97) = \alpha$  به دست آمد. نتایج به دست آمده از محاسبه میزان اعتبار آزمون بعد از حذف هر سؤال نشان داد که با حذف سؤالات ۱ الف، ۱ ب، ۱ ج، ۱ د، ۲ الف، ۲ ج، ۴ ج، ۶ الف، ۸ ج، ۱۴ الف، ۱۷ الف، ۲۴ میزان اعتبار تغییر نکرده و با حذف باقی سؤالات، اعتبار آزمون فقط به میزان  $(0/001)$  کاهش یافته است.

با توجه به جدول ۳، بیشترین و کمترین ضریب تمیز به ترتیب متعلق به سؤالات ۲۲ الف و ۴ ج  $(0/88, 0/1)$  است که سؤال ۴ ج، ساده‌ترین سؤال آزمون زیست‌شناسی است. سؤالات ۱۱a و ۳۰ الف a نیز، به‌طور مشترک با ضریب دشواری  $(0/19)$ ، دشوارترین سؤالات هستند. توزیع فراوانی و درصد سؤالات آزمون برحسب ضرایب دشواری و تمیز به دست آمده از مدل کلاسیک اندازه‌گیری، در جدول ۴ نشان داده شده است. ۹ سؤال دارای ضریب دشواری بیش از  $0/7$  هستند که دامنه ضریب تمیز این سؤالات بین  $0/1$  تا  $0/6$  است. ۶۳ سؤال معادل  $(80/77)$  درصد از سؤالات دارای ضریب دشواری بین  $0/3$  تا  $0/7$  دارند که از این میان تنها یک سؤال دارای ضریب تمیز کمتر از  $0/3$  است. همچنین، ۶ سؤال دارای ضریب دشواری کمتر از  $0/3$  هستند که ضریب تمیز آنها بین  $0/5$  تا  $0/67$  است.  $97/43$  درصد سؤالات زیست‌شناسی دارای ضریب تمیز بالاتر از  $0/3$  هستند که بیانگر وضعیت بسیار مطلوبی است. در این درس، سؤالات ۴ ج و ۲۴ دارای ضریب تمیز کمتر از  $0/3$  هستند و قادر به تفکیک آزمودنی‌های قوی و ضعیف نبوده و نامناسب‌اند.

جدول ۳. تحلیل سؤالات آزمون زیست‌شناسی بر اساس مدل کلاسیک و مدل پرسش پاسخ به همراه مقادیر آگاهی

سؤال	مدل کلاسیک			مدل دو پارامتری			مقادیر پارامتر آگاهی	
	ضریب دشواری	ضریب تمیز	دشواری	تشخیصی (نسب)	ضریب دو	سطح معناداری	پیشینه آگاهی	ثبات ماکسیم
الف ۱	۰/۷۵	۰/۳۳	-۱/۵۲	۰/۴۹	۷	۰/۷۳	۰/۱۸	-۱/۵
ب ۱	۰/۸۰	۰/۳۸	-۱/۵۰	۰/۷۰	۸/۴	۰/۳۹	۰/۳۶	-۱/۵
ج ۱	۰/۷۷	۰/۴۵	-۱/۳۳	۰/۶۸	۱۶/۶	۰/۰۵	۰/۳۳	-۱/۳
د ۱	۰/۴۵	۰/۴۰	۰/۳۵	۰/۳۶	۱۳/۹	۰/۱۸	۰/۰۹	۰/۴، ۰/۳
الف ۲	۰/۷۹	۰/۳۳	-۱/۸۵	۰/۴۶	۱۲/۵	۰/۲۵	۰/۱۵	-۱/۹، -۱/۸
ب ۲	۰/۵۱	۰/۶۹	-۰/۰۸	۰/۷۵	۱۰/۸	۰/۲۹	۰/۴۱	-۰/۱
ج ۲	۰/۵۱	۰/۵۰	-۰/۰۹	۰/۵۰	۱۱/۲	۰/۳۴	۰/۱۸	-۰/۱
د ۲	۰/۷۲	۰/۶۰	-۰/۸۳	۰/۹۸	۸/۹	۰/۴۵	۰/۶۹	-۰/۸
۳a	۰/۴۰	۰/۶۶	۰/۳۸	۰/۸۴	۱/۶	۰/۹۹	۰/۵۲	۰/۴
۳b	۰/۴۷	۰/۶۶	۰/۰۷	۰/۷۹	۴/۳	۰/۸۳	۰/۴۵	۰/۱
الف ۴	۰/۵۹	۰/۶۴	-۰/۳۹	۰/۸۰	۲/۴	۰/۹۹	۰/۴۶	-۰/۴
ب ۴	۰/۵۳	۰/۷۹	-۰/۱۳	۱/۰۳	۸/۷	۰/۳۷	۰/۷۷	-۰/۱
ج ۴	۰/۹۲	۰/۱۰	-۳/۶۷	۰/۴۴	۶/۴	۰/۶۰	۰/۱۳	-۳
۵a	۰/۷۹	۰/۴۵	-۱/۲۲	۰/۹۸	۱۱/۷	۰/۱۷	۰/۶۹	-۱/۲
۵b	۰/۸۴	۰/۴۰	-۱/۲۸	۱/۲۹	۴/۴	۰/۷۳	۱/۲۰	-۱/۳
الف ۶	۰/۵۹	۰/۵۴	-۰/۴۵	۰/۵۷	۱۰/۵	۰/۳۹	۰/۲۳	-۰/۵، -۰/۴
ب ۶	۰/۶۳	۰/۶۴	-۰/۵۶	۰/۸۳	۶/۲	۰/۷۹	۰/۵	-۰/۶
الف ۷	۰/۶۶	۰/۵۴	-۰/۷۳	۰/۷۰	۶/۱	۰/۷۳	۰/۳۵	-۰/۷
ب ۷	۰/۵۵	۰/۶۵	-۰/۲۴	۰/۷۴	۷/۱	۰/۶۲	۰/۴۰	-۰/۲
الف ۸. a	۰/۵۹	۰/۷۵	-۰/۳۵	۱/۱۰	۲/۸	۰/۹۴	۰/۸۷	-۰/۴
الف ۸. b	۰/۵۵	۰/۸۵	-۰/۲۱	۱/۲۷	۱/۶	۰/۹۸	۱/۱۷	-۰/۲
ب ۸	۰/۳۶	۰/۷۳	۰/۴۳	۱/۱۹	۱۰/۸	۰/۱۵	۱/۰۲	۰/۴

ویژگی‌های روان‌سنجی امتحانات نهایی سال سوم متوسطه و ... / ۵۹

۸ ج	۰/۸۲	۰/۳۴	-۱/۶۶	۰/۶۸	۱۲/۷	۰/۱۲	۰/۳۳	-۱/۷
۹ الف	۰/۴۵	۰/۶۹	۰/۱۳	۰/۸۱	۲۱/۱	۰/۰۱	۰/۴۷	۰/۱
۹ ب	۰/۵۵	۰/۷۷	-۰/۲۱	۱/۰۶	۱۴/۹	۰/۰۶	۰/۸۱	-۰/۲
۹ ج	۰/۵۴	۰/۷۹	-۰/۱۶	۱/۱۵	۱۱/۴	۰/۱۲	۰/۹۶	-۰/۲
۱۰ الف	۰/۴۸	۰/۷۳	۰/۰۶	۰/۹۵	۲/۵	۰/۹۸	۰/۶۶	۰/۱
۱۰ ب	۰/۵۹	۰/۷۵	-۰/۳۵	۱/۱۶	۱/۴	۰/۹۹	۰/۹۶	-۰/۳
۱۱a	۰/۱۹	۰/۵۴	۱/۱۵	۱/۱۵	۶/۳	۰/۵۰	۰/۹۵	۱/۱
۱۱b	۰/۲۷	۰/۶۷	۰/۷۹	۱/۲۸	۴/۸	۰/۶۸	۱/۱۹	۰/۸
۱۲a	۰/۵۱	۰/۷۳	-۰/۰۷	۱/۰۷	۵/۸	۰/۶۷	۰/۸۴	-۰/۱
۱۲b	۰/۴۹	۰/۸۲	۰/۰۱	۱/۲۹	۲/۹	۰/۸۹	۱/۲۰	۰
۱۳a	۰/۵۳	۰/۸۳	-۰/۱۲	۱/۵۵	۱/۲	۰/۹۹	۱/۷۴	-۰/۱
۱۳b	۰/۴۹	۰/۷۹	-۰/۰۳	۱/۳۲	۲/۵	۰/۹۲	۱/۲۵	۰
۱۴ الف	۰/۶۷	۰/۳۸	-۱/۱۲	۰/۴۴	۱۲/۳	۰/۲۶	۰/۱۴	-۱/۱
۱۴ ب	۰/۶۲	۰/۷۳	-۰/۴۵	۱/۰۹	۱/۱	۰/۹۹	۰/۸۶	-۰/۴
۱۵ الف	۰/۳۷	۰/۶۳	۰/۵۱	۰/۸۱	۱۱/۳	۰/۱۸	۰/۴۷	۰/۵
۱۵ ب	۰/۶۰	۰/۶۱	-۰/۴۳	۰/۷۸	۶/۶	۰/۵۸	۰/۴۴	-۰/۴
۱۵ ج	۰/۵۷	۰/۷۱	-۰/۲۹	۰/۹۶	۳	۰/۹۶	۰/۶۷	-۰/۳
۱۵ د	۰/۶۰	۰/۸۱	-۰/۳۷	۱/۲۴	۹/۱	۰/۲۴	۱/۱۱	-۰/۴
۱۶a	۰/۵۹	۰/۷۰	-۰/۳۵	۰/۹۸	۳/۳	۰/۹۱	۰/۶۹	-۰/۳
۱۶b	۰/۶۳	۰/۷۰	-۰/۴۶	۱/۱۹	۳/۴	۰/۹۰	۱/۰۱	-۰/۵
۱۷ الف	۰/۳۱	۰/۴۹	۰/۹۱	۰/۶۲	۱۴/۳	۰/۰۷	۰/۲۸	۰/۹
۱۷ ب	۰/۶۱	۰/۷۲	-۰/۴۶	۰/۹۴	۲۵	۰/۰۱	۰/۶۴	-۰/۵
۱۸ الف	۰/۴۰	۰/۶۰	۰/۳۳	۰/۶۶	۱۰/۳	۰/۲۴	۰/۳۲	۰/۳
۱۸ a.ب	۰/۴۱	۰/۷۴	۰/۳۱	۰/۹۸	۱۳/۵	۰/۱۰	۰/۶۹	۰/۳
۱۸ b.ب	۰/۵۱	۰/۸۳	-۰/۰۵	۱/۲۷	۳/۸	۰/۸۰	۱/۱۵	-۰/۱
۱۹ الف	۰/۵۳	۰/۷۵	-۰/۱۴	۱/۱۰	۴/۸	۰/۷۸	۰/۸۷	-۰/۱
۱۹ ب	۰/۴۸	۰/۵۸	۰/۰۵	۰/۶۳	۲۲/۱	۰/۰۱	۰/۲۸	۰/۱
۱۹ ج	۰/۴۳	۰/۵۹	۰/۳۰	۰/۶۶	۲۱/۲	۰/۰۱	۰/۳۲	۰/۳
۱۹ د	۰/۲۲	۰/۵۷	۱/۰۴	۱/۱۶	۳/۸	۰/۸۰	۰/۹۷	۱
۲۰ الف	۰/۴۹	۰/۷۷	۰/۰۱	۱/۱۰	۴/۴	۰/۸۲	۰/۸۷	۰

۶۰ / فصلنامه اندازه گیری تربیتی، دوره ششم، شماره ۲۳، پاییز ۱۳۹۵

۰/۴	۰/۹۵	۰/۴۲	۸/۱	۱/۱۵	-۰/۴۴	۰/۷۲	۰/۶۲	۲۰ ب
-۰/۱	۱/۰۹	۰/۱۵	۱۰/۸	۱/۲۳	-۰/۰۵	۰/۸۵	۰/۵۰	۲۱a
۰/۴	۰/۷۳	۰/۰۱	۱۹/۲	۱/۰۱	۰/۳۶	۰/۷۵	۰/۴۰	۲۱b
۰/۲	۱/۵۷	۰/۱۹	۹/۹	۱/۴۷	۰/۱۸	۰/۸۸	۰/۴۳	۲۲ الف
۰/۵	۰/۷۶	۰/۲۵	۱۰/۱	۱/۰۳	۰/۴۶	۰/۷۴	۰/۳۷	۲۲ ب
۰/۳	۱/۰۵	۰/۸۲	۴/۴	۱/۲۱	۰/۳۳	۰/۷۹	۰/۴۰	۲۲ ج
۰/۵	۰/۸۰	۰/۵۸	۵/۶	۱/۰۵	۰/۴۷	۰/۷۳	۰/۳۶	۲۳ الف
۰	۰/۳۷	۰/۰۱	۲۱/۲	۰/۷۱	-۰/۰۱	۰/۶۵	۰/۵۰	۲۳ ب
۱/۶	۰/۱۷	۰/۹۱	۴	۰/۴۸	۱/۶۳	۰/۲۸	۰/۳۰	۲۴
۰/۹	۰/۷۰	۰/۲۲	۱۰/۶	۰/۹۹	۰/۸۷	۰/۶۲	۰/۲۷	۲۵ الف
۰/۶	۱/۴۸	۰/۷۰	۳/۸	۱/۴۳	۰/۶۱	۰/۷۵	۰/۳۱	۲۵ ب
۰/۴	۰/۵۱	۰/۰۵	۱۳/۹	۰/۸۴	۰/۳۹	۰/۶۸	۰/۴۰	۲۵ ج
۰	۱/۰۴	۰/۸۴	۳/۴	۱/۲۰	۰/۰۲	۰/۸۱	۰/۴۸	۲۶ الف
۰/۴	۱/۴۰	۰/۰۵	۱۳/۸	۱/۳۹	۰/۴۵	۰/۸۰	۰/۳۶	۲۶ ب
۰/۴	۱/۰۵	۰/۸۰	۴/۶	۱/۲۱	۰/۳۷	۰/۷۹	۰/۳۸	۲۶ ج. a
۰/۷	۰/۷۰	۰/۶۷	۵/۸	۰/۹۸	۰/۷۴	۰/۶۵	۰/۳۰	۲۶ ج. b
-۰/۱	۰/۶۸	۰/۲۹	۹/۷	۰/۹۷	-۰/۱۰	۰/۷۳	۰/۵۲	۲۷ الف
۰/۳	۰/۷۶	۰/۳۵	۸/۹	۱/۰۲	۰/۲۹	۰/۷۶	۰/۴۲	۲۷ ب. a
۰/۳	۱/۰۶	۰/۷۴	۵/۱	۱/۲۱	۰/۳۴	۰/۷۹	۰/۳۹	۲۷ ب. b
۰/۱	۰/۴۲	۰/۰۳	۱۸/۳	۰/۷۷	۰/۱۰	۰/۶۳	۰/۴۷	۲۸a
-۰/۶	۰/۶۴	۰/۲۱	۱۰/۹	۰/۹۴	-۰/۵۹	۰/۶۷	۰/۶۵	۲۸b
۰/۴	۰/۳۶	۰/۸۰	۵/۴	۰/۷۰	۰/۴۱	۰/۵۸	۰/۴۰	۲۹a
۰/۹	۰/۷۰	۰/۰۶	۱۳/۶	۰/۹۹	۰/۸۹	۰/۵۹	۰/۲۷	۲۹b
۱/۲	۰/۹۹	۰/۲۷	۷/۶	۱/۱۷	۱/۲۰	۰/۵۰	۰/۱۹	۳۰ الف. a
۰/۱	۱/۳۱	۰/۷۷	۴/۱	۱/۳۵	۰/۱۳	۰/۸۴	۰/۴۵	۳۰ الف. b
۰/۱	۰/۶۰	۰/۲۷	۱۱/۱	۰/۹۱	۰/۱۲	۰/۶۹	۰/۴۶	۳۰ ب
				۰/۹۶	-۰/۰۹	۰/۶۵	۰/۵۰	میانگین

$$-2 \text{ LOG LIKELIHOOD}_{2pt} = 45762/7608$$

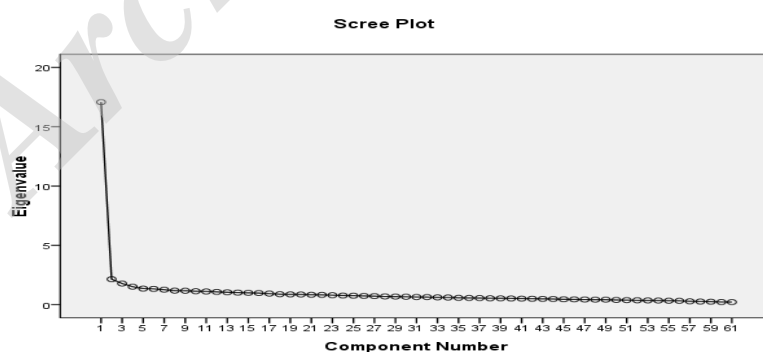


جدول ۴. توزیع سؤالات زیست‌شناسی بر اساس ضرایب دشواری و تمیز کلاسیک

ضریب دشواری			
کمتر از ۰/۳	بین ۰/۳ تا ۰/۷	بیشتر از ۰/۷	
کمتر از ۰/۲	-	۱ (۱/۲۸٪)	}
بین ۰/۲ تا ۰/۳	۱ (۱/۲۸٪)	-	
بیشتر از ۰/۳	۶۲ (۷۹/۴۹٪)	۸ (۱۰/۲۶٪)	

**تحلیل روان‌سنجی سؤالات امتحانات نهایی بر اساس مدل‌های IRT.** در این تحقیق، تحلیل روان‌سنجی سؤالات مطابق با مراحل زیر صورت گرفته است که به دلیل برازش بهتر مدل دو پارامتری، تحلیل تنها بر اساس این مدل ارائه می‌گردد. مراحل تحلیل عبارت‌اند از: (۱) بررسی مفروضات (۲) برازش مدل با داده‌ها (۳) برآورد پارامترهای هر سؤال به همراه آزمون‌های دو و سطح معناداری (۴) بررسی برازندگی تک‌تک سؤالات با مدل‌ها (۵) توصیف و تحلیل مقادیر پارامترهای سؤال در چارچوب ملاک‌های تفسیر (۶) منحنی ویژگی سؤالات (۷) مقادیر بیشینه آگاهی سؤالات و تنای ماکسیمم (۸) تابع آگاهی آزمون و خطای استاندارد اندازه‌گیری (۹) هیستوگرام توزیع توانایی آزمودنی‌ها.

**درس ادبیات فارسی.** در راستای بررسی مفروضات، نمودار اسکری آزمون ادبیات فارسی نشان می‌دهد، آزمون مذکور دارای یک عامل غالب است و به عبارتی تک‌بعدی است. در نتیجه استقلال موضعی نیز در آن رعایت شده است.



نمودار ۱. نمودار اسکری آزمون ادبیات فارسی

در هر دو مدل یک پارامتری و دو پارامتری، شاخص RMS به دست آمده (۰/۲۴) کمتر از ۲ است که بیانگر برازش کلی مدل‌های مذکور با داده‌ها است. به منظور تعیین برازش مدل با داده‌ها (مقایسه مدل PL۱ با 2PL) در آزمون ادبیات فارسی، خی دو به صورت زیر محاسبه گردید؛

$$\chi^2 = (-2 \text{ LOG LIKELIHOOD}_{1pl}) - (-2 \text{ LOG LIKELIHOOD}_{2pl})$$

$$\chi^2 = ۳۴۰۲۳/۵۸۶۱ - ۳۳۵۳۸/۰۱۹۲ = ۴۸۵/۵۷$$

از آنجا که خی دو محاسبه شده (۴۸۵/۵۷) از خی دو بحرانی (۸۸/۳۸) با درجه آزادی ۶۱ در سطح اطمینان ۹۹ درصد بزرگ‌تر است، بنابراین نتیجه گرفته می‌شود که مدل دو پارامتری برازش بهتری با داده‌های آزمون دارد. مقادیر برآورد شده‌ی پارامترهای سؤال بر اساس مدل دو پارامتری به همراه آزمون خی دو و سطح معناداری در جدول ۱ نشان داده شده است. چنانچه در مدلی، سطح معناداری سؤال از مقدار آلفای موردنظر کوچک‌تر باشد، نشان‌دهنده‌ی عدم برازش سؤال با مدل است. بررسی برازش سؤال با مدل نشان داد؛ در مدل یک پارامتری، ۱۰ سؤال در سطح (۰/۰۱) و ۱۸ سؤال در سطح (۰/۰۵) با مدل برازش نداشتند. در مدل دو پارامتری نیز، سؤالات ۱۵، ۴۶، ۴۹، ۵۰ در سطح (۰/۰۵) با مدل برازش نداشتند که به دلیل برازش بهتر این مدل با داده‌ها، بهتر است این سؤالات از آزمون حذف شوند. سؤال ۱۵ در سطح ۰/۰۵ با هیچ مدلی برازش نداشت.

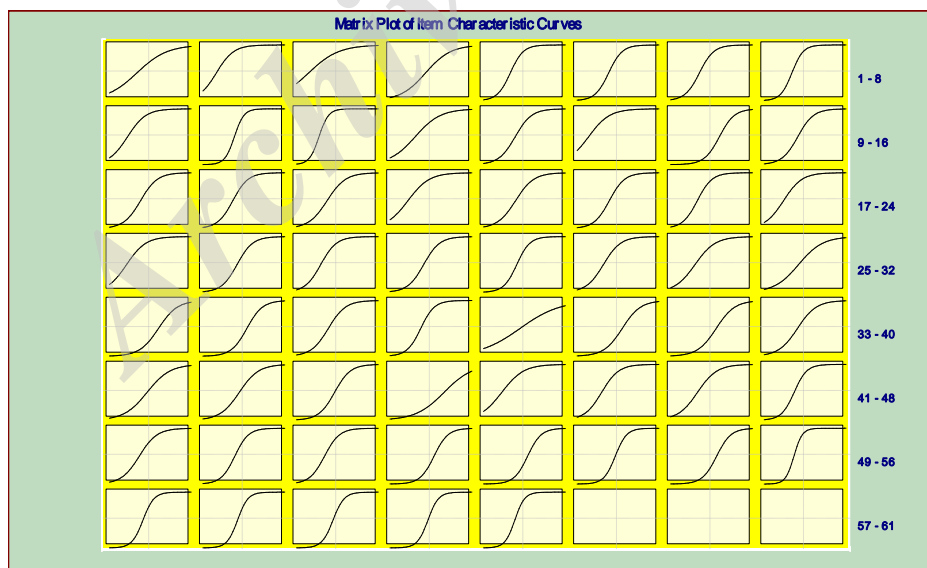
در جدول ۵، توزیع فراوانی و درصد سؤالات آزمون ادبیات فارسی برحسب پارامترهای برآورد شده‌ی دشواری و تشخیص حاصل از مدل دو پارامتری، در چارچوب ملاک‌های تفسیر ارائه شده است. با توجه به اطلاعات جداول ۱ و ۵، به توصیف و تحلیل مقادیر پارامترهای سؤال پرداخته می‌شود. ۴۴ سؤال (۰/۷۲/۱۳) در دامنه دشواری متوسط قرار دارند و مابقی سؤالات نیز سؤالات آسانی هستند. سؤال ۱۴ با دشواری (۰/۱۶-) آسان‌ترین سؤال آزمون ادبیات فارسی است. این سؤال در سطوح پایین توانایی از قدرت تشخیص متوسطی (۰/۷۶) برخوردار است. سؤال ۴۱ نیز سخت‌ترین سؤال است که با داشتن دشواری (۰/۹۷)

دارای قدرت تشخیص متوسطی (۰/۵۷) است. قابل ذکر است که هیچ کدام از سؤالات آزمون ادبیات فارسی از قدرت تشخیص منفی، خیلی ضعیف و ضعیف برخوردار نیستند. سؤالات ۴۸ و ۳۶ با ضریب تشخیص (۱/۹۲) و (۰/۴۰) به ترتیب بالاترین و کمترین قدرت تشخیص را در بین سؤالات این آزمون دارند.

جدول ۵. توزیع سؤالات آزمون ادبیات فارسی بر اساس پارامترهای سؤال در مدل دو پارامتری

پارامتر	دشواری		
	بیشتر از ۱ (دشواری)	۱ تا -۱ (متوسط)	کمتر از -۱ (آسان)
خیلی ضعیف	-	-	-
ضعیف	-	-	-
متوسط	-	۶ (۹/۸۴٪)	۵ (۸/۲٪)
بالا	-	۱۱ (۱۸/۰۳٪)	۸ (۱۳/۱۱٪)
خیلی بالا	۱ و بالاتر	۲۷ (۴۴/۲۶٪)	۴ (۶/۵۶٪)

در ادامه، منحنی ویژگی سؤالات ارائه و بررسی گردید. با مشاهده نمودار ۲ می‌توان دریافت که در مدل دو پارامتری، ۵۰ درصد سؤالات دارای قدرت تشخیص خیلی بالا هستند و بیش از نیمی از سؤالات در دامنه دشواری متوسط قرار دارند.



نمودار ۲. منحنی ویژگی سؤالات آزمون ادبیات فارسی در مدل دو پارامتری

در جدول ۱، به بیشینه آگاهی هر سؤال همراه آن سطح توانایی که این بیشینه آگاهی در آن اتفاق می‌افتد، اشاره گردید. در آزمون ادبیات فارسی بیشترین و کمترین میزان آگاهی (۲/۶۶، ۰/۱) به ترتیب به سؤالات ۴۸ و ۳۶ اختصاص دارد. در جدول ۶، توزیع فراوانی و درصد سؤالات برحسب دامنه‌ی آگاهی ارائه شده است.

جدول ۶. توزیع سؤالات آزمون ادبیات فارسی برحسب دامنه‌ی آگاهی در مدل دو پارامتری

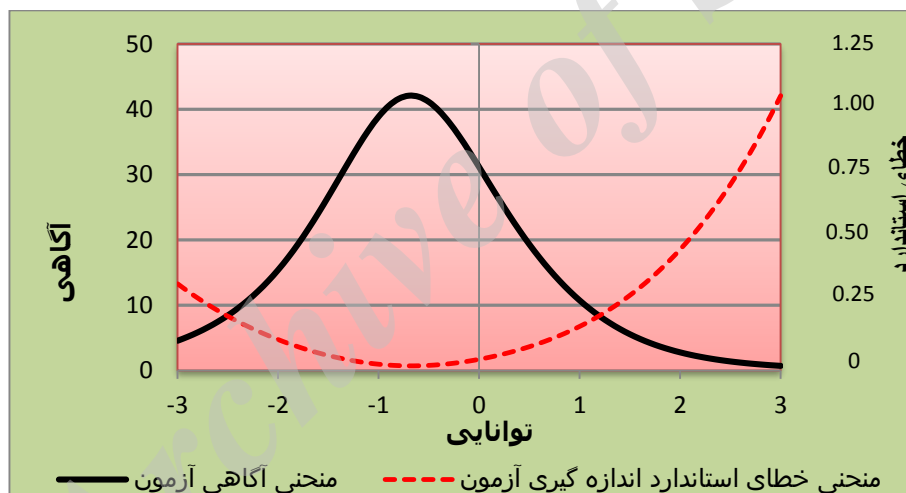
دامنه آگاهی سؤالات			تعداد و درصد سؤالات
۰ - ۱	۱ - ۲	۲ و بیش از آن	
۴۵ (۷۳/۷۷٪)	۱۳ (۲۱/۳۱٪)	۳ (۴/۹۲٪)	

مقادیر آگاهی و خطای استاندارد اندازه‌گیری آزمون ادبیات فارسی در سطوح توانایی در جدول ۷ و منحنی آن‌ها در نمودار ۳ نشان داده شده است. با مشاهده و بررسی جدول مذکور و منحنی آگاهی آزمون، می‌توان دریافت که بیشینه میزان آگاهی دهندگی این آزمون در سطح توانایی ۰/۷- به میزان ۴۲/۱۱ قرار دارد که بیانگر این مطلب است؛ توانایی افرادی که در این سطح قرار دارند در مقایسه با سطوح دیگر با دقت بیشتری برآورد می‌شود. خطای معیار اندازه‌گیری در سطح توانایی مذکور برابر ۰/۷۱ است که کمترین میزان خطای آزمون است. کمینه میزان آگاهی آزمون (۰/۷۱) اختصاص به سطح توانایی ۳ دارد.

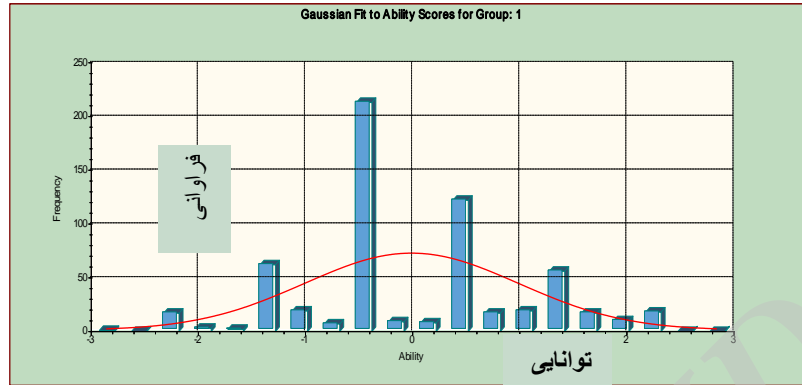
با توجه به اینکه میانگین و انحراف استاندارد نمرات در درس ادبیات فارسی به ترتیب ۱۳/۱۹ و ۴/۸۰ است، نمره‌ی ملاک ۹/۸۳ معادل نقطه‌ی تا ۰/۷- است. اگر نمره‌ی ۹/۸۳ به‌عنوان نمره‌ی ملاک انتخاب شود، خطای کمتری در تصمیم‌گیری راجع به قبولی یا رد آزمودنی‌ها خواهیم داشت. بیشترین میزان آگاهی دهندگی آزمون در دامنه‌ی توانایی (۰/۳-، ۱-)، قرار دارد که معادل دامنه‌ی نمرات (۱۱/۷۵، ۸/۳۹) است. از آنجاکه نمره‌ی ملاک ۱۰ به‌عنوان نمره‌ی قبولی در نظام آموزش و پرورش در این دامنه قرار می‌گیرد و فاصله‌ی بسیار کمی با نمره‌ی ملاک پیشنهادی ۹/۸۳ دارد، بنابراین به‌طور قابل قبولی می‌توان به نتایج آزمون اعتماد کرد. همچنین، طبق نمره‌ی ملاک (۹/۸۳)، ۷۵/۱۷ درصد و طبق نمره‌ی ملاک (۱۰)، ۷۴/۸۳ درصد افراد در این آزمون قبول‌اند.

جدول ۷. مقادیر برآورد شده‌ی آگاهی و خطای استاندارد اندازه‌گیری آزمون ادبیات فارسی در سطوح مختلف توانایی برحسب مدل دو پارامتری

خطای استاندارد اندازه‌گیری	آگاهی	$\theta$	خطای استاندارد اندازه‌گیری	آگاهی	$\theta$
۳/۶۵	۱۹/۳۱	۰/۵	۱۳/۲۷	۴/۵۶	-۳
۶/۷۵	۱۰/۷۳	۱	۸/۳۰	۸/۴۵	-۲/۵
۱۱/۵۲	۵/۵۵	۱/۵	۴/۷۳	۱۵/۴۲	-۲
۱۸/۵۶	۲/۷۷	۲	۲/۲۹	۲۶/۶۸	-۱/۵
۲۸/۵۱	۱/۳۹	۲/۵	۰/۹۵	۳۸/۹۱	-۱
۴۲/۱۱	۰/۷۱	۳	۰/۷۸	۴۱/۱۴	-۰/۵
			۱/۷۱	۳۱/۱۴	۰



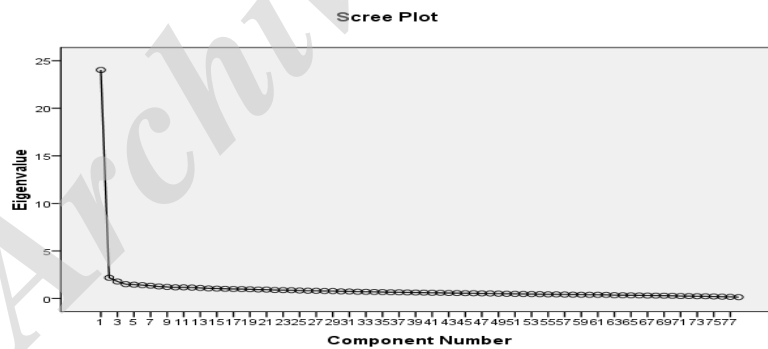
نمودار ۳. منحنی آگاهی و خطای استاندارد اندازه‌گیری آزمون ادبیات فارسی برحسب مدل دو پارامتری



نمودار ۴. هیستوگرام توزیع توانایی آزمودنی‌ها در آزمون ادبیات فارسی برحسب مدل دو پارامتری

با بررسی توزیع توانایی آزمودنی‌ها، کمینه و بیشینه نمره توانایی در درس ادبیات فارسی (۲/۷۷-، ۲/۱۶) به دست آمد. میانگین توانایی دختران و پسران در این درس (۰/۳۱، ۰/۳۱- ) بود. نمودار ۴، هیستوگرام توانایی برآورد شده‌ی آزمودنی‌ها را نمایش می‌دهد. میانگین و انحراف استاندارد توزیع توانایی افراد در مدل دو پارامتری به ترتیب ۰/۰۰۱ و ۱/۰۱۹۶ به دست آمد.

#### درس زیست‌شناسی



نمودار ۵. نمودار اسکری آزمون زیست‌شناسی

با توجه به نمودار اسکری، پیداست که آزمون مذکور تک‌بعدی بوده و در نتیجه مفروضه

استقلال موضعی نیز برقرار

است. پس از اطمینان از برقراری مفروضات، برازش مدل با داده‌ها بررسی گردید. در دو مدل یک پارامتری و دو پارامتری، شاخص RMS برابر ۰/۱۹ است که بیانگر برازش کلی مدل‌های مذکور با داده‌ها است. به منظور تعیین برازش مدل با داده‌ها (مقایسه مدل IPL با 2PL) در آزمون زیست‌شناسی، خی دو محاسبه شد؛

$$\chi^2 = 46535/3262 - 45762/7608 = 772/56$$

از آنجا که خی دو محاسبه شده (۷۷۲/۵۶) از خی دو بحرانی (۱۱۲/۳۳) با درجه آزادی ۷۸ در سطح اطمینان ۹۹ درصد بزرگ‌تر است، بنابراین نتیجه گرفته می‌شود که مدل دو پارامتری برازش بهتری با داده‌های آزمون دارد. مقادیر برآورد شده‌ی پارامترهای سؤال در مدل دو پارامتری به همراه آزمون خی دو و سطح معناداری در جدول ۳ ارائه گردیده است که با توجه به اطلاعات مذکور، میزان برازندگی سؤال با مدل بررسی گردید. در مدل یک پارامتری، ۲۰ سؤال در سطح (۰/۰۱) و ۳۳ سؤال در سطح (۰/۰۵) با مدل برازش نداشتند. در مدل دو پارامتری، سؤالات (۱ ج، ۹ الف، ۱۷ ب، ۱۹ ب، ۱۹ ج، ۲۱b، ۲۳ ب، ۲۵ ج، ۲۶ ب، ۲۸a) در سطح (۰/۰۵) با مدل برازش نداشتند که به دلیل برازش بهتر این مدل با داده‌ها، بهتر است این سؤالات از آزمون حذف شوند. در میان سؤالات مذکور، سؤالات ۹ الف، ۱۷ ب، ۱۹ ب، ۱۹ ج، ۲۱b، ۲۳ ب، ۲۶ ب، ۲۸a در سطح ۰/۰۵ با هیچ‌یک از مدل‌ها برازش نداشتند.

مقادیر پارامترهای سؤال، ارائه شده در جدول ۳ نشان می‌دهد؛ سؤال ۴ ج با دشواری (۳/۶۷-) و سؤال ۲۴ با دشواری (۱/۶۳) آسان‌ترین و سخت‌ترین سؤالات آزمون زیست‌شناسی هستند که از قدرت تشخیص متوسطی برخوردارند. سؤالات ۱۳a و ۱ د با ضریب تشخیص (۱/۵۵) و (۰/۳۶) به ترتیب بالاترین و کمترین قدرت تشخیص را در بین سؤالات این آزمون دارند.

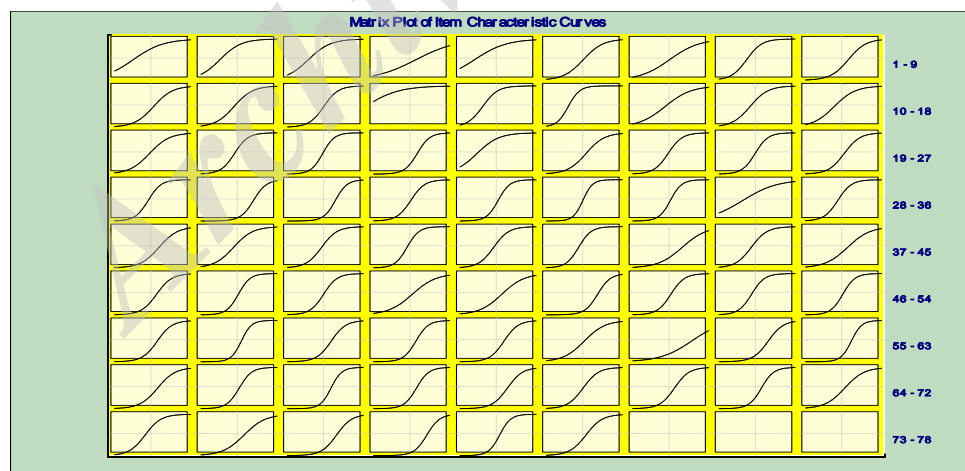
در جدول ۸، توزیع فراوانی و درصد سؤالات آزمون زیست‌شناسی برحسب پارامترهای برآورد شده‌ی دشواری و تشخیص حاصل از مدل دو پارامتری در چارچوب ملاک‌های

تفسیر ارائه شده است. در مدل دو پارامتری ۶۵ سؤال (۸۳/۳۳٪) در دامنه دشواری متوسط قرار دارند. ۴ سؤال از کل سؤالات نیز سؤالات دشوار و ۹ سؤال باقی‌مانده نیز سؤالات آسانی هستند. قابل ذکر است که هیچ کدام از سؤالات آزمون زیست‌شناسی از قدرت تشخیص منفی و خیلی ضعیف برخوردار نیستند و تنها یک سؤال از قدرت تشخیص ضعیف برخوردار است. ۵۶ سؤال معادل (۷۱/۷۹٪) سؤالات از قدرت تشخیص بالا و خیلی بالا برخوردارند.

جدول ۸. توزیع سؤالات آزمون زیست‌شناسی بر اساس پارامترهای سؤال در مدل دو پارامتری

دشواری			پارامتر	توزیع
کمتر از ۱- (آسان)	۱ تا ۱- (متوسط)	بیشتر از ۱ (دشوار)		
-	-	-	۰/۰۱ - ۰/۲۰	خیلی ضعیف
-	(۱/۲۸)٪	-	۰/۲۱ - ۰/۳۷	ضعیف
(۷/۸۹۷)٪	(۱۳/۱۶/۶۷)٪	(۱/۲۸)٪	۰/۳۸ - ۰/۷۸	متوسط
(۱/۲۸)٪	(۱۹/۲۴/۳۶)٪	-	۰/۷۹ - ۰/۹۹	بالا
(۱/۲۸)٪	(۳۲/۴۱/۰۲)٪	(۳/۸۵)٪	۱ و بالاتر	خیلی بالا

منحنی ویژگی تمامی سؤالات در نمودار ۶ نشان داده شده است که با مشاهده آن می‌توان دریافت که در مدل دو پارامتری، بیشتر سؤالات دارای قدرت تشخیص بالا و خیلی بالا هستند و در دامنه دشواری متوسط قرار دارند.



نمودار ۶. منحنی ویژگی سؤالات آزمون زیست‌شناسی در مدل دو پارامتری



مقادیر بیشینه آگاهی و تنای ماکسیمم سؤالات آزمون زیست‌شناسی در جدول ۳ آمده است. بیشترین و کمترین میزان آگاهی (۱/۷۴، ۰/۰۹) به ترتیب به سؤالات ۱۳a و ۱ اختصاص دارد. میزان آگاهی ۵۹ سؤال (معادل ۷۵/۶۴٪ از سؤالات)، کمتر از یک است که میزان آگاهی ۲۶ سؤال آن (معادل ۳۳/۳۳٪ از کل سؤالات)، کمتر از ۰/۵ است. در جدول زیر، توزیع فراوانی و درصد سؤالات آزمون برحسب دامنه‌ی آگاهی در مدل دو پارامتری ارائه شده است.

جدول ۹. توزیع سؤالات آزمون زیست‌شناسی برحسب دامنه‌ی آگاهی در مدل دو پارامتری

دامنه آگاهی سؤالات		
۰-۱	۱-۲	۲ و بیش از آن
۵۹ (۷۵/۶۴٪)	۱۹ (۲۴/۳۶٪)	-
تعداد و درصد سؤالات		

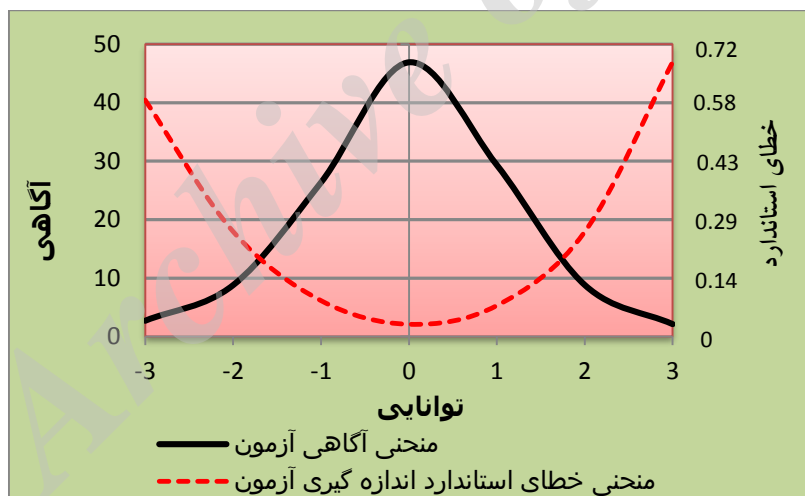
مقادیر آگاهی و خطای استاندارد اندازه‌گیری آزمون زیست‌شناسی در سطوح مختلف توانایی در جدول ۱۰ و منحنی آن‌ها در نمودار ۷ نشان داده شده است. با بررسی جدول مذکور و منحنی آگاهی آزمون می‌توان دریافت که بیشترین و کمترین میزان آگاهی‌دهندگی این آزمون به میزان (۲/۱۲، ۴۶/۹۶) به ترتیب متعلق به سطوح توانایی (۰/۱، ۳) است. بیشینه میزان آگاهی‌دهندگی این آزمون در سطح توانایی ۰/۱ بیانگر این مطلب است؛ توانایی افرادی که در این سطح قرار دارند در مقایسه با سطوح دیگر با دقت بیشتری برآورد می‌شود. خطای معیار اندازه‌گیری در سطح توانایی مذکور برابر ۲/۱۲ است که کمترین میزان خطای آزمون است.

میانگین و انحراف استاندارد نمرات در درس زیست‌شناسی، به ترتیب ۹/۹۷ و ۵/۱۲ و نمره‌ی ملاک ۱۰/۴۸ معادل سطح توانایی ۰/۱ است. اگر نمره‌ی ۱۰/۴۸ به‌عنوان نمره‌ی ملاک انتخاب شود، خطای کمتری در تصمیم‌گیری راجع به قبولی یا رد آزمودنی‌ها خواهیم داشت. بیشترین میزان آگاهی‌دهندگی آزمون در دامنه‌ی توانایی (۰/۵، ۰/۴-) قرار دارد که معادل دامنه‌ی نمرات (۱۲/۵۳، ۷/۹۲) است. از آنجا که نمره‌ی ملاک ۱۰ در این دامنه قرار می‌گیرد و تقریباً ۰/۵ نمره با نمره‌ی ملاک پیشنهادی ۱۰/۴۸ فاصله دارد، بنابراین به‌طور قابل

قبولی می توان به نتایج آزمون اعتماد کرد. همچنین، طبق نمره ی ملاک (۱۰/۴۸)، ۴۵/۱۷ درصد و طبق نمره ی ملاک (۱۰)، ۴۷/۱۷ درصد افراد در این آزمون قبول اند.

جدول ۱۰. مقادیر برآورد شده ی آگاهی و خطای استاندارد اندازه گیری آزمون زیست شناسی در سطوح مختلف توانایی برحسب مدل دو پارامتری

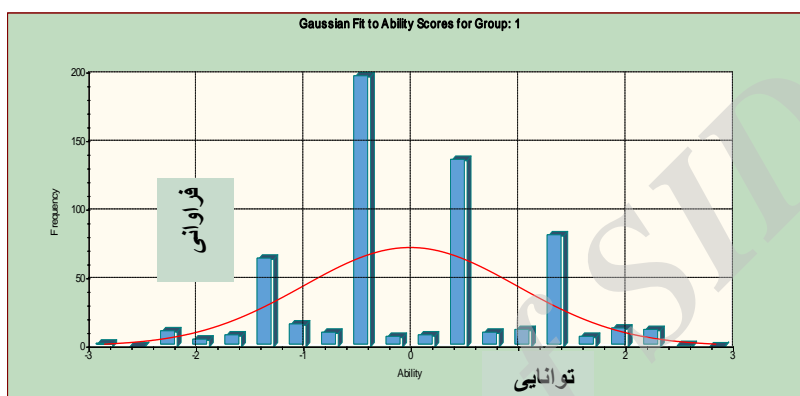
خطای استاندارد اندازه گیری	آگاهی	$\theta$	خطای استاندارد اندازه گیری	آگاهی	$\theta$
۲/۸۰	۴۲/۰۴	۰/۵	۴۰/۴۵	۲/۷۰	-۳
۵/۳۲	۲۹/۳۴	۱	۲۷/۷۸	۴/۸۱	-۲/۵
۱۰/۰۸	۱۷/۰۷	۱/۵	۱۷/۹۸	۸/۷۸	-۲
۱۷/۹۳	۸/۸۲	۲	۱۰/۹۲	۱۵/۷۲	-۱/۵
۲۹/۹۰	۴/۳۲	۲/۵	۶/۱۷	۲۶/۳۶	-۱
۴۶/۹۶	۲/۱۲	۳	۳/۲۴	۳۹/۲۹	-۰/۵
			۲/۱۳	۴۶/۹۱	۰



نمودار ۷. منحنی آگاهی و خطای استاندارد اندازه گیری آزمون زیست شناسی برحسب مدل دو پارامتری

پارامتر توانایی آزمودنی ها به همراه پارامترهای سؤال برآورد گردیده است که نمودار ۸، هیستوگرام توانایی برآورد شده ی آزمودنی ها را نمایش می دهد. میانگین و انحراف

استاندارد توزیع توانایی افراد در مدل دو پارامتری به ترتیب ۰/۰۱ و ۱/۰۳ به دست آمده است. بررسی توزیع توانایی آزمودنی‌ها و هیستوگرام آن نشان داد که کمینه و بیشینه نمره توانایی در درس زیست‌شناسی (۳/۰۶-، ۲/۷۴) است. همچنین، میانگین توانایی دختران و پسران در درس مذکور (۰/۳۲، ۰/۲۹-) به دست آمد.



نمودار ۸. هیستوگرام توزیع توانایی آزمودنی‌ها در آزمون زیست‌شناسی برحسب مدل دو پارامتری

مقایسه‌ی CTT و IRT از لحاظ پارامتر تشخیص. جداول ۱۱ و ۱۲، میزان توافق دو مدل CTT و IRT را در تعیین سؤالات مناسب و نامناسب از نظر قدرت تشخیص برای دروس ادبیات فارسی و زیست‌شناسی نشان می‌دهند. همان‌طور که جداول مذکور نشان می‌دهند، از نظر قدرت تشخیص؛ در آزمون ادبیات فارسی، ۶۰ سؤال و در آزمون زیست‌شناسی، ۷۲ سؤال در هر دو مدل مناسب و همچنین، در آزمون زیست‌شناسی در هر دو مدل، یک سؤال نامناسب شناسایی شد.

جدول ۱۱. میزان توافق دو مدل کلاسیک و IRT برای آزمون ادبیات فارسی

IRT		سؤال	CTT
نامناسب	مناسب		
۱	۶۰	مناسب	
۰	۱	نامناسب	

جدول ۱۲ میزان توافق دو مدل کلاسیک و IRT برای آزمون زیست‌شناسی

IRT		
نامناسب	مناسب	سؤال
۵	۷۲	مناسب
CTT		
۱	۰	نامناسب

### بحث و نتیجه‌گیری

استفاده از نتایج امتحانات نهایی در گزینش داوطلبان ورود به دوره‌های کارشناسی مراکز و مؤسسات آموزش عالی کشور، حساسیت و اهمیت آن‌ها را بیش‌ازپیش کرده است. با توجه به لزوم بالا بودن کیفیت چنین آزمون‌هایی، سؤالات آزمون‌های زیست‌شناسی و ادبیات فارسی با استفاده از مدل‌های کلاسیک و جدید اندازه‌گیری بررسی شدند که نتایج حاصل از این بررسی‌ها مورد بحث قرار می‌گیرد.

**ارزیابی کمی آزمون بر طبق CTT.** در آزمون‌های سرنوشت‌ساز که به تصمیم‌گیری‌های مهم درباره‌ی افراد مربوط می‌شود، ضریب اعتبار باید بیش از ۰/۹۵

باشد (شریفی، ۱۳۸۸). این اصل در هر دو آزمون مورد مطالعه، رعایت شده است. همچنین، نتایج حاصل از روش لوپ، نشان داد که این دو آزمون، فاقد سؤالاتی بوده‌اند که با حذفشان، اعتبار افزایش می‌یافت و این نشان از همسانی درونی مناسب سؤالات با هم دارد. ضریب اعتبار دروس مورد مطالعه، از اعتبار به‌دست‌آمده در مطالعه‌ی حشمتی و همکاران (۱۳۸۳)، بالاتر بوده است. ضریب تمیز نیز، نسبت به مطالعه‌ی رحیمی (۱۳۸۷)، سلیمانی (۱۳۸۴) و حشمتی و همکاران (۱۳۸۳) در وضعیت مطلوب‌تری قرار دارد. همچنین، ضریب دشواری به‌دست‌آمده با ضریب دشواری مطالعات مذکور، همخوان است. میانگین ضریب دشواری زیست‌شناسی در این تحقیق (۰/۵۰) پایین‌تر از میانگین ضریب دشواری زیست‌شناسی در استان لرستان (۰/۶۱) در مطالعه‌ی حشمتی و همکاران (۱۳۸۳) است که بیانگر دشواری بودن امتحان زیست‌شناسی خرداد ۹۰ نسبت به خرداد ۸۱ است.

**ارزیابی کمی آزمون بر طبق مدل‌های IRT.** نمودار اسکری نشان از برقراری مفروضات در دروس مورد مطالعه داشت. در این دروس همچنین، آزمون‌های دو نشان داد مدل دو

پارامتری نسبت به مدل یک پارامتری برازش بیشتری با داده‌ها دارد. انتظار می‌رود به‌منظور تفکیک افراد مسلط از غیر مسلط، آگاهی‌آزمون در اطراف نمره‌ی ملاک به مقدار زیادی قله‌ای بوده و از بیشترین میزان دقت در اندازه‌گیری توانایی افراد برخوردار باشد. با توجه به اینکه حداکثر میزان آگاهی‌بخشی در دو آزمون ادبیات فارسی و زیست‌شناسی در فاصله‌ی بسیار نزدیکی به نمره‌ی ملاک ۱۰ قرار دارد، می‌توان گفت که این اصل در دو آزمون مذکور رعایت شده است. با توجه به بیشینه مقدار آگاهی‌دروس مورد مطالعه، می‌توان گفت: این دروس برای افرادی که در محدوده‌ی توانایی متوسط قرار دارند مناسب است و توانایی آن‌ها در مقایسه با سایرین با خطای کمتر برآورد می‌شود.

بررسی توزیع توانایی‌آزمودنی‌ها نشان داد، عملکرد دختران در هر دو درس بهتر بوده است. با توجه به عملکرد تحصیلی بهتر دختران و درصد بالای قبولی آن‌ها در کنکور طی سال‌های اخیر می‌توان گفت؛ موفقیت دختران در آزمون ورودی دانشگاه‌ها نتیجه‌ی تلاش مستمر آن‌ها در دوران تحصیل است.

هرچند عملکرد دانش‌آموزان در درس زیست‌شناسی، نامناسب بوده اما سؤالات این درس از نظر ویژگی‌های روان‌سنجی (شاخص دشواری و ضریب تمیز) با ملاک‌های علمی طراحی سؤال و آزمون‌سازی انطباق دارد. پایین بودن میانگین درس زیست‌شناسی از نمره ملاک می‌تواند ناشی از عوامل مختلفی باشد. از این رو، به آموزش و پرورش پیشنهاد می‌شود در جامعه‌ی مورد مطالعه، به ریشه‌یابی این مشکل بپردازد و راهکارهایی برای بهبود عملکرد دانش‌آموزان در نظر بگیرد.

تحلیل سؤالات بر اساس CTT و IRT نشان داد که آزمون ادبیات فارسی فاقد سؤالات دشوار است و بیشتر سؤالات آسان آن در نیمه‌ی اول آزمون قرار دارد. آزمون زیست‌شناسی نیز با سؤالات آسان آغاز می‌شود، هرچند که سؤالات آسان کمی دارد. می‌توان نتیجه گرفت که اصل ترتیب سؤالات از ساده به مشکل در دو آزمون مذکور به نسبت رعایت شده است. همان‌طور که سیف (۱۳۹۱) بیان کرده است؛ وقتی یک آزمون با سؤالات آسان آغاز می‌شود، موجب افزایش اعتماد به نفس آزمودنی‌ها می‌شود، به طوری که افزایش کارایی آن‌ها را در عملکرد روی سؤالات پیچیده بعدی به دنبال خواهد داشت. میانگین شاخص‌های تحلیل

سؤال (ضرایب دشواری و تمیز) در هر دو درس نشان داد که سؤالات این دروس بر اساس CTT و IRT، دشواری متوسطی داشته‌اند و درعین‌حال در تفکیک آزمودنی‌های قوی و ضعیف موفق عمل کرده‌اند. به بیانی دیگر؛ سؤالات این دروس بر طبق دو مدل، از ویژگی‌های روان‌سنجی مطلوب برخوردارند. همچنین، از نظر قدرت تشخیص، دو مدل مذکور در آزمون ادبیات فارسی در ۹۸/۳۶ درصد موارد و در آزمون زیست‌شناسی در ۹۳/۵۹ درصد موارد با هم توافق داشتند.

یکی از اصلی‌ترین یافته‌های این مطالعه این است که اگرچه ویژگی‌های روان‌سنجی امتحانات نهایی مطلوب است، اما استفاده از آن‌ها برای مقاصد گزینشی جای بسی تردید دارد. همان‌طور که همبلتون و جونز (۱۹۹۳) مطرح کرده‌اند؛ در فرآیند انتخاب سؤال، علاوه بر در نظر گرفتن روایی محتوایی، همیشه سعی می‌شود سؤالات با پارامتر تمیز بالا انتخاب شوند و انتخاب سطح دشواری نیز، معمولاً تحت تأثیر هدف آزمون و توزیع توانایی پیش‌بینی‌شده گروهی که آزمون برای آن‌ها ساخته می‌شود، قرار می‌گیرد. از آنجاکه امتحانات نهایی در گروه آزمون‌های ملاک مرجع قرار دارند و با هدف دستیابی به اهداف آموزشی طراحی می‌شوند، انتظار می‌رود به منظور تفکیک افراد مسلط از غیر مسلط، آگاهی این آزمون‌ها در اطراف نمره‌ی ملاک (نمره ۱۰) بیشینه باشد. به این معنا که در امتحانات نهایی که با اهداف ارزشیابی پایانی صورت می‌گیرند، می‌بایست سؤالات دشواری متوسط داشته باشند. درحالی‌که در کنکور که با اهداف گزینشی برگزار می‌گردد، سؤالات طوری طراحی می‌شوند که تمیز میان همه‌ی آزمودنی‌ها را به حداکثر برساند و تنای ماکسیمم در تنای بالاتری قرار می‌گیرد؛ یعنی، دشواری سؤالات در سطح بالایی قرار دارد. با توجه به متفاوت بودن نمره‌ی برش امتحانات نهایی و کنکور، می‌توان گفت؛ در کل امتحانات نهایی چون آزمون‌های ملاک مرجع هستند، قابلیت گزینش داوطلبان ورود به مؤسسات و مراکز آموزش عالی را ندارند. تحقق هر کدام از اهداف ارزشیابی و گزینشی مستلزم امتحانات جداگانه‌ای است. همچنین، با توجه به خطیر بودن تصمیماتی که بر اساس نتایج کنکور گرفته می‌شود، ضرورت دارد مطالعات گسترده و علمی به منظور جایگزینی نظام جدید گزینش دانشجو صورت پذیرد و امتحانات نهایی از جنبه‌های مختلف، بارها مورد بررسی قرار گیرند.

در پایان به محققان آموزشی پیشنهاد می‌شود، در بررسی ویژگی‌های روان‌سنجی سؤالات امتحانات نهایی، مدل‌های جدید اندازه‌گیری را بکار گیرند. از آنجا که این امتحانات به دلیل سراسری بودن، دامنه‌ی زیادی از افراد را شامل می‌شوند، این خود امکان استفاده از مدل‌های پرسش پاسخ را که مستلزم نمونه‌های بزرگ است، مهیا می‌سازد. هرچند به دلیل تشریحی بودن این امتحانات، جمع‌آوری داده‌ها (ثبت پاسخ هر سؤال) کار بسیار وقت‌گیری خواهد بود، با این حال از مزایای IRT نمی‌توان صرف نظر کرد. کاربردهای IRT به محققان فرصت می‌دهد که ویژگی‌های روان‌سنجی امتحانات نهایی با دقت بیشتری نشان داده شود.

## منابع

- حشمتی، عبدالرضا؛ رافع، ابوالقاسم و جمشید نژاد، مهرداد. (۱۳۸۳). تجزیه و تحلیل نتایج امتحانات نهایی سال سوم متوسطه خرداد ۸۱ در سراسر کشور. گزارش طرح پژوهشی مصوب معاونت آموزش و پرورش نظری و مهارتی.
- رحیمی، محمود. (۱۳۸۷). ارزیابی و تحلیل سطوح حیطة شناختی و شاخص‌های روان‌سنجی سؤالات امتحانات نهایی پایه سوم راهنمایی استان خوزستان در خردادماه ۱۳۸۶. گزارش طرح پژوهشی مصوب سازمان آموزش و پرورش استان خوزستان.
- سلیمانی، علی. (۱۳۸۴). بررسی و تحلیل سؤالات امتحانی دروس ریاضی مقطع متوسطه نظری مراکز آموزشی استان کرمانشاه در امتحانات پایانی خردادماه ۸۳. گزارش طرح پژوهشی مصوب سازمان آموزش و پرورش استان کرمانشاه.
- سیف، علی‌اکبر. (۱۳۹۱). اندازه‌گیری، سنجش، و ارزشیابی آموزشی. تهران: دوران.
- شریفی، حسن پاشا. (۱۳۸۸). اصول روان‌سنجی و روان‌آزمایی. تهران: رشد.
- کیامنش، ع و حسنی، م. (۱۳۸۸). نقد و بررسی نظام ارزشیابی تحصیلی ایران از منظر ارزشیابی نظام آموزشی. فصلنامه نوآوری‌های آموزشی، ۳۰: ۱۰۱ - ۷۵.
- همبلتون، رونالد ک؛ سوامیناتان، اچ و جین راجرز، اچ. (۱۳۸۹). مبانی نظریه پرسش-پاسخ (ترجمه‌ی محمدرضا فلسفی نژاد). تهران: انتشارات دانشگاه علامه طباطبایی.

- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1): 1-21.
- Foley, B. P. (2010). *Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique* (doctoral dissertation, Nebraska-Lincoln University). Retrieved from <http://digitalcommons.unl.edu/cehsdiss/75>
- Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on. *Educational measurement: issues and practice*, 12(3): 38-47.
- Hambleton, R. K., & Traub, R. E. (1970). *Analysis of Empirical Data Using Two Logistic Latent Trait Models*. Retrived from ERIC database. (ED042816).
- Hambleton, R. K., & Van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4): 373-378.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks: Sage Publications.
- Suen, H. K., Lei, P. W. (2007). Classical versus Generalizability theory of measurement. *Educational measurement*, 4: 1-13.
- Van der Elst, W., Ouweland, C., van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The Shortened Raven Standard Progressive Matrices Item Response Theory-Based Psychometric Analyses and Normative Data. *Assessment*, 20(1): 48-59.

Archive