فصلنامه پژوهش‌های نوین روانشناختی

سال پانزدهم شماره ۵۷ بهار ۱۳۹۹

مقایسه ویژگی‌های روانسنجی مدل‌های دارای مجانب پایین (3PL)، بالا (3Plu) و هر دو مجانب (4PL) براساس داده‌های آزمون‌های سراسری ورود به دانشگاه

جواد محمدلو[۱]، بلال ایزانلو[۲]

۱-ارشد تحقیقات آموزشی، دانشکده روانشناسی و علوم تربیتی دانشگاه خوارزمی

۲-استادیار دانشکده روانشناسی و علوم تربیتی دانشگاه خوارزمی، تهران، ایران

تاریخ وصول: ۱۳۹۸/۰۳/۲۰ **تاریخ پذیرش: ۱۳۹۸/۰۹/۱۰**

چکیده

هدف پژوهش حاضر مقایسه ویژگی‌های روان‌سنجی مدل لوجستیک سه پارامتری دارای مجانب پایین (3PL)، مدل سه پارامتری دارای مجانب بالا (3Plu) و مدل لجستیک چهار پارامتری با هر دو مجانب پایین و بالا (4PL) است. برای مقایسه بهتر، مدل‌های 1PL و 2PL نیز در تحلیل‌ها لحاظ و نتایج هر پنج مدل مقایسه شد. به منظور مقایسه از داده‌های شرکت‌کنندگان آزمون سراسری ایران سال ۱۳۹۵-۱۳۹۴، برگزار شده توسط سازمان سنجش آموزش (NOET) استفاده شد. به طوره ویژه، در گروه علوم ریاضی آزمون فیزیک و معارف در گروه تجربی آزمون شیمی و معارف و انسانی آزمون ادبیات تخصصی و معارفبرای تحلیل استفاده شد. از هر گروه یک نمونه تقریبا ۶۰۰۰ نفری به صورت تصادفی انتخاب و تحلیل شد. مفروضه تک‌بعدی بودن با نرم افزار NOHARM بررسی شد که در همه آزمون‌ها برقرار بود. سپس برای پاسخ‌گویی به سوال‌ها از بسته mirt در نرم افزار R استفاده شد. نتایج برازش هر یک از مدل‌ها در سطح سوال نشان داد در آزمون‌های عمومی سوال‌ها بیشتر با مدل‌های 3PL و 4PL و در آزمون‌های اختصاصی سوال‌ها بیشتر به ترتیب با مدل 2PL و $3PL_U$ برازش دارند. برازش در سطح مدل بر اساس شاخص DIC نشان داد که بجز شیمی در سایر آزمون‌ها مدل 3PL نسبت به مدل 4PL مناسبت‌تر است (البته در بین پنج مدل، مدل 2PL نسبت به همه مدل‌های استفاده شده بهتر است). نتایج برازش در سطح مدل براساس شاخص بیز (BF) در کل با نتایج شاخص DIC همخوانی داشت. همبستگی بین نمره‌های خام هر آزمون و توانایی برآورد شده آن بر اساس هر پنج مدل بالا است، که حاکی از رابطه خطی قوی بین متغیرها است. توانایی برآورد شده مدل‌ها تفاوت معناداری باهم نداشت. از نظر آگاهی در سطوح توانایی بالا، مدل ۳ پارامتری نسبت به مدل ۴ پارامتری آگاهی‌دهنده‌تر است. بطور کلی برای مدل ۴ پارامتری نسبت به مدل ۳ پارامتری مزیت خاصی پیدا نشد. مگر در سوال‌های خاصی که برازش این مدل با آنها بهتر بود. البته تاثیر حجم نمونه بر نتایج مشخص نیست و لازم است در پژوهش‌های آتی نتایج مدل‌ها بر اساس حجم نمونه نیز مقایسه شوند.

کلیدواژه: نظریه سوال- پاسخ، مدل ۴ پارامتری (4PL)، مدل ۳ پارامتری با مجانب پایین (3PL)، مدل سه پارامتری با مجانب بالا (3PLu)، مجانب پایین، مجانب بالا

---

* (نویسنده مسئول) Email: izan.b@khu.ac.ir

مقدمه

در نظریه سوال- پاسخ[1] (IRT) برای تحلیل داده‌های بدست آمده از مقیاس‌ها، پرسشنامه‌ها و آزمون‌های روانی از مدل‌های ریاضی استفاده می‌شود. این مدل‌ها برای ارائه و نمایش بهتر خم ویژه سوال[2] به کار می‌روند. این‌گونه مدل‌ها در حقیقت برای توسعه و پیشرفت دقیق یک نظریه اندازه‌گیری و انتقال اطلاعات درباره ویژگی‌های فنی سوال‌ها ابزار مفیدی به شمار می‌آیند. انتخاب یک مدل باید بر پایه ملاحظات نظری و تجربی مانند برازش مدل با داده‌ها[3] صورت گیرد (بیکر[4]، ۲۰۰۲). سه مدل تک‌بعدی نظریه سوال پاسخ، بخاطر تعداد پارامترهایی که برای توصیف داده‌ها استفاده می‌کنند، به مدل‌های یک (1PL)، دو (2PL) و سه پارامتری (3PL) معروف هستند. این مدل‌ها که برای داده‌های دوارزشی مناسبند از نظر تعداد پارامترهای استفاده شده برای توصیف سوال‌ها متفاوتند. مدل یک پارامتری (1PL) که اغلب مدل راش نیز نامیده می‌شود، ولی از برخی جهات تفاوت‌هایی نیز با مدل راش دارد، یکی از پر استفاده‌ترین مدل‌های نظریه سوال پاسخ است. در این مدل احتمال پاسخ درست هر فرد به یک سوال از طریق معادله (۱) بدست می‌آید.

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1, 2, \dots, n \tag{۱}$$

که در آن $P_i(\theta)$ احتمال آنکه یک آزمودنی با توانایی θ به سوال i پاسخ درست دهد را نشان می‌دهد، b_i پارامتر دشواری سوال، n تعداد سوال‌های آزمون و e یک عدد اصم معروف به عدد اُیلر[5] است که مقدار تقریبی آن تا سه رقم اعشار برابر ۲/۷۱۸ است. در مدل یک پارامتری دشواری سوال، تنها ویژگی سوال است که جایگاه سوال در پیوستار توانایی (۳+ تا ۳-) را نشان می‌دهد. در این مدل فرض می‌شود که شیب همه سوال‌ها برابر یک است ($a = 1$). به علاوه احتمال آنکه آزمودنی‌های با توانایی بسیار پایین سوال را بدرستی پاسخ دهند صفر است. بنابراین هیچ عبارتی برای حدس در این مدل وجود ندارد (گائو[6]،۲۰۱۱؛ چنگ[7] و لیو[8]، ۲۰۱۶). اگر درمدل‌های لجستیک نظریه سوال پاسخ به جای شیب(a) از عبارت Da استفاده شود، تابع لجستیک تقریب بسیار نزدیکی از تابع نرمال خواهد بود. در این پژوهش در تابع لجستیک استفاده شده برای توصیف مدل‌ها از a برای شیب استفاده شده است. اگر شیب همه سوال‌ها نابرابر فرض شود (که بر اساس شواهد تجربی منطقی‌تر به نظر می‌رسد) مدل دو پارامتری به دست می-آید. برنبام[9] (۱۹۶۸) تابع دو پارامتری لُجستیک (2PL) را جایگزین تابع دو پارامتری اجایو نرمال کرد. زیرا توابع لجستیک نسبت به توابع اجایو نرمال دارای مزیت هستند. مثلا کار کردن با آنها به لحاظ محاسباتی راحت‌تر است. بر این اساس، احتمال پاسخ درست به سوال توسط یک آزمودنی در مدل دو پارامتری برنبام به همان صورت معادله (۱) بیان می‌شود. با این تفاوت که در اینجا شیب برابر یک نیست. هر چه شیب بیشتر باشد، توانایی سوال برای جدا کردن آزمودنی‌های با سطوح مختلف توانایی بیشتر است. روشن است که در شیب شدید، سوال در یک نقطه افراد را به دو گروه تقسیم می‌کند. در صورتی که در مدل دو پارامتری مجانب پایین منحنی ویژگی سوال را برابر صفر در نظر نگیریم به مدل سه پارامتری (3PL) می‌رسیم. هدف از گنجاندن این پارامتر تبیین پاسخ درست به برخی از

---

1. Item Response Theory
2. Item Characteristic Curve
3. Model-Data Fit
4. Baker
5. Euler number
6. Gao
7. Cheng
8. liu
9. Birnbaum

سوال‌های دشوار توسط افراد ضعیف است (ریس و والر[1]، ۲۰۰۳؛ چنگ و لیو، ۲۰۱۶). مدل سه پارامتری از سه پارامتر دشواری، شیب و حدس‌پذیری سوال برای تبیین عملکرد فرد در سوال استفاده می‌کند. بیان ریاضی مدل سه پارامتری بصورت معادله (۲) است.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1, 2, \dots, n \qquad (۲)$$

که در آن c_i مجانب پایین غیرصفر منحنی ویژگی سوال (ICC) است و احتمال اینکه آزمودنی‌های با توانایی پایین، به یک سوال (معمولا دشوار) پاسخ درست دهند را نشان می‌دهد. باید توجه کرد که در این مدل ارزش c_i به عنوان تابعی از سطح توانایی تغییر نمی‌کند. بنابراین اگرچه ضعیف‌ترین و تواناترین آمودنی‌ها احتمال یکسانی برای درست پاسخ دادن به سوال از طریق حدس زدن را دارند ولی حضور این پارامتر در مدل بیشتر به نفع افراد ضعیف است تا قوی (گائو، ۲۰۱۱). این موضوع به راحتی از طریق رسم چند منحنی ویژگی سوال با دشواری و شیب یکسان ولی حدس‌پذیری متفاوت قابل مشاهده است. در خصوص پارامتر مجانب پایین می‌توان به چند نکته اشاره کرد. اولین مورد در این خصوص برچسب‌های متفاوتی است که به پارامتر C اختصاص داده می‌شود. پارامتر C در اصل پارامتر حدس سوال (لرد[2]،۱۹۸۰، ص ۱۲) نام دارد. با این وجود، چون پارامتر C معمولا کمتر از آن چیزی است که توسط مدل حدسی تصادفی (معکوس تعداد گزینه‌ها) پیش بینی می‌شود، پس پارامتر C به پارامتر شبه حدس[3] موسوم شده است. تفاوت بین مقدار C و مقدار پیش‌بینی مدل حدس تصادفی به دلیل تفاوت در جذابیت گزینه‌ها است. مدل حدس تصادفی فرض می‌کند که همه گزینه‌ها جذابیت برابری دارند. ولی براساس تحلیل سوال کلاسیک می‌دانیم که گزینه‌های سوال در میزان جذابیت‌شان برای افراد فرق دارند. برای مثال استفاده از کلید واژگان در گزینه‌ها، روشی معمول برای افزایش جذابیت گزینه‌ها است. بعلاوه، عواملی و منابعی که افراد را برای آزمون آماده می‌کنند، به آزمون دهندگانی که جواب سوال را نمی‌دانند یاد می‌دهند که طولانی‌ترین گزینه را انتخاب کنند، زیرا این نوع گزینه‌ها معمولا پاسخ درست‌اند. در نتیجه پیش فرض‌های مدل حدس تصادفی در داده‌های حاوی پاسخ‌های مشاهده منعکس نمی‌شود و معمولا بین پارامتر حدس و احتمال حدس تصادفی همیشه اختلاف وجود دارد. دومین مورد به ماهیت پارامتر C مربوط است. کارکرد C در مدل، انعکاس این مسئله است که برخی افراد با توانایی پایین ممکن است به سوال پاسخ درست داده و امتیاز یک بگیرند، درحالی که طبق مدل دو پارامتری نباید این طور باشد. اینگونه پاسخ‌ها نشان دهنده تعامل بین ویژگی‌های فرد و سوال (ازجمله قالب سوال) است. در مورد ابزارهای مهارتی، ویژگی‌های فرد نه تنها توانایی وی را منعکس می‌کند بلکه گرایشات خطرپذیری و آزمون خبرگی (آگاهی آزمون) را نیز شامل می‌شود. این دو عامل اخیر متغیرهای فردی پنهان غیرمستقیمی هستند که اهمیت چندانی ندارند. بنابراین اگرچه C به عنوان پارامتر سوال در نظر گرفته می‌شود، ولی بهتر است به عنوان ویژگی فرد (یک پارامتر فردی دیگر) در نظر گرفته شود تا ویژگی سوال، یا دست کم تعامل بین ویژگی‌های فرد و سوال محسوب گردد. یعنی، پاسخ‌های افراد با توانایی کم، تعامل بین توانایی فرد و ویژگی‌های سوال را نشان می‌دهد. سومین مورد به پیش فرض ضمنی مربوط است که به هنگام استفاده از C و تاثیر آن بر برآورد مد نظر قرار می‌گیرد. در مدل سه پارامتری حاوی C، فرض می‌شود، صرف نظر از جایگاه فرد، تمایل طبیعی وی به حدس زدن در تمام پیوستار توانایی ثابت است (C به عنوان تابعی از توانایی تغییر نمی‌کند). این پیش فرض ممکن است در تمام شرایط معقول نباشد. در خصوص تاثیرات آن باید گفت که C غیرصفر برآورد جایگاه فرد را کاهش داده و از میزان آگاهی سوال می‌کاهد. بنابراین اگرچه ما C را در مدل قرار می‌دهیم تا مقدار آن برآورد شود ولی مطلوب‌تر آن است که مقدار آن به صفر نزدیک باشد. در این مورد مدل ۲ پارامتری می‌تواند بازنمایی قابل قبولی از داده‌ها ارائه دهد. یعنی این فرض برای پارامتر حدس که هر آزمودنی احتمال یکسانی جهت حدس درست در یک سوال را دارد ممکن است انعکاسی از موقعیت حدس واقعی نباشد

[1] . Waller
[2] . Lord
[3] . Pseudo-Guessing

(دی‌آیالا[1] ، ۲۰۰۸). عدم دقت برآورد پارامتر حدس نیز نگرانی‌های دیگری را سبب می‌شود. رنولدز[2] (۱۹۸۶) در یک مطالعه شبیه‌سازی نشان داد که پارامتر حدس نمی‌تواند به صورت دقیق برآورد شود، اگر چه او اندازه نمونه و طول آزمون را افزایش و توزیع توانایی را تغییر داد. مطالعه انجام شده توسط ری[3] (۱۹۷۹) در مورد دقت برآورد پارامتر حدس نشان داد که دقت پارامتر حدس حتی با ۲۰۰۰ آزمودنی و آزمون ۸۰ سوال شبیه‌سازی شده باز هم پایین است. باید عوامل دیگری بر دقت پارامتر حدس تاثیر داشته باشند، زیرا حجم نمونه و طول آزمون عوامل اصلی موثر بر برآورد پارامتر حدس نیستند (گائو، ۲۰۱۱). در اوایل سال ۱۹۸۱، بارتون و لرد مجانب بالا منحنی (d)، که خطای ناشی از بدشانسی آزمون‌دهنده‌های قوی در پاسخ به سوال‌های نسبتا آسان را نشان می‌دهند، به مدل سه پارامتری اضافه کردند و آن را شبیه مجانب پایین که خوش شانسی دانش آموزان با توانایی کم در حدس زدن پاسخ درست را حساب می‌کند، مورد بحث قرار دادند. معادله مدل چهار پارامتری لجستیک (4PL)، که برای هر سوال پارامتر d متفاوتی را در نظر می‌گیرد، بصورت معادله (۳) است.

$$P_i(\theta) = C_i + (d_i - C_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad (۳)$$

که در آن علاوه بر سه پارامتر ذکر شده، پارامتر d_i ، مجانب بالاتر (سطح عدم توجه یا بی‌دقتی) منحنی ویژه سوال است. در حالی که احتمال پاسخ درست در مدل سه پارامتری بین c (مجانب پایین منحنی ویژه سوال) و ۱ است، این احتمال در مدل چهار پارامتری بین c و مجانب بالای منحنی، d، قرار داد. مجانب بالای منحنی، بی‌دقتی افراد در پاسخ دادن به سوال خاصی را نشان می‌دهد (ین و همکاران، ۲۰۱۲). در مدل ۴ پارامتری، d یا پارامتر بی‌دقتی، مجانب بالای منحنی ویژه سوال را منعکس می‌کند، که نشان می‌دهد افراد توانا بنا به دلایل مختلف به سوال‌هایی که در سطح توانایی‌شان است، پاسخ نادرست می‌دهند. دلیل وجود پارامتر d می‌تواند عوامل گوناگونی به غیر از استرس و بی‌دقتی باشد. ممکن است آزمون "سرعتی" باشد (وست، ۲۰۱۵؛ سدریدیس، سوئسس و الهربی[4]، ۲۰۱۶). بنابراین آزمودنی‌های توانا سوال‌ها را به خاطر نبود وقت اشتباه جواب می‌دهند. آزمون دهنده‌هایی که به شیوه غیرمعمول و خلاقانه به سوالات پاسخ می‌دهند در کلید نمره‌گذاری لحاظ نشده و به همین دلیل این پاسخ‌ها به عنوان نادرست نمره‌گذاری می‌شوند (وست، ۲۰۱۵ به نقل از کارباسوس[5]، ۲۰۰۳). از میان سایر موارد، عواملی مثل عدم توجه، خستگی یا فقدان انگیزه، تلاش ناکافی، یا ناتوانی در پردازش سوال‌های دارای کلمات معکوس را می‌توان نام برد (سدریدس و همکاران، ۲۰۱۶). با این وجود، تعیین دلیل رواج پارامتر d بسیار مبهم است و ممکن است نتایج نامحتملی در این خصوص ارایه شود (وست، ۲۰۱۵). با توجه به مطالب بالا تعیین علت رواج پارامتر d بسیار مبهم و احتمالی است. یعنی مدل نمی‌تواند دقیقا توضیح دهد که چرا چنین پدیده‌ای رخ می‌دهد.

در دهه‌های اخیر، از میان مدل هایی که ذکر شد، در آزمون‌های چندگزینه‌ای که به صورت دو ارزشی نمره‌گذاری می‌شوند، مدل ۳ پارامتری (3PL) و مدل‌های ساده‌تر یک پارامتری (1PL) و دو پارامتری (2PL) بیشترین توجه را دریافت کرده‌اند. گرچه بارتون و لرد[6] (۱۹۸۱) نوع بسط یافته مدل ۳ پارامتری با مجانب بالا، که به آن مدل چهار پارامتری (4PL) می‌گویند، را نیز مطرح کرده بودند، ولی این مدل تا این اواخر توجه زیادی را دریافت نکرد. همانطور که لاکن و رولیسن[7] (۲۰۱۰) اشاره کرده‌اند. تسلط قوی مدل

[1] . De Ayala
[2] . Renolds
[3] . Ree
[4] . Sideridis, Tsaousis & Al Harbi
[5] . Karabatsos
[6] . Barton & Lord
[7] . Loken and Rulison

۳ پارامتری در پیشینه، فقدان توافق عام درباره مفید بودن مدل ۴ پارامتری و مشکلات فنی در برآورد دقیق خط مجانب بالا، از دلایل قوی علیه استفاده از مدل ۴ پارامتری بودند. علی‌رغم این موانع مفهومی، مدل 4PL اخیرا در پیشینه مورد توجه قرار گرفته است. توسعه نرم افزارهای مدل‌سازی آماری دقیق و افزایش در قدرت و توان محاسباتی یکی از دلایل این توجه است (مگیس[۱]، ۲۰۱۳). بارتون و لرد (۱۹۸۱) از تحقیق‌شان نتیجه گرفتند که مدل چهار پارامتری بطور منظم برآورد بیشینه درست‌نمایی را بهبود نمی‌بخشد و برآوردهای سطوح توانایی را تغییر نمی‌دهد. به علاوه، به دلیل پیچیدگی ریاضی مدل برآوردهای آن وقت‌گیر است. ذکر این نکته ارزشمند است که آنها پارامتر d را به طور مستقیم برآورد نکردند بلکه خوبی برازش مدل‌ها را با تثبیت پارامتر d (به ترتیب به ۱، ۰/۹۹ و ۰/۹۸) مقایسه کرده‌اند. اینکار بدلیل مشکلات محاسباتی و عدم وجود نرم افزار مناسب در آن زمان صورت گرفته است (والر[۲] و ریس[۳]، ۲۰۱۰). همبلتون و سوامیناتان (۱۹۸۵؛ صص. ۴۹-۴۸) معتقدند که مدل چهار پارامتری ارزش عملی ندارد. بنابراین تحقیق روی ویژگی‌های روان‌سنجی و کاربردی مدل چهار پارامتری بدلیل مشکلات برآورد (حتی امروزه نیز برآورد پارامتر مجانب پایین مشکل در نظر گرفته می‌شود؛ لاکن و رولیسن، ۲۰۱۰) متوقف و کار بر روی آن بی‌فایده قلمداد شده است. بعد از گذشت ۲۵ سال، امروزه علاقه به مدل چهار پارامتری دوباره زنده شده است. در ابتدا روش‌های محاسبات بیزین[۴] مطرح شدند. این روش‌ها قطعا می‌توانند برآورد را سرعت و تسهیل ببخشند، اگرچه آنها همه مشکلات مفهومی مدل چهار پارامتری را حل نمی‌کنند (وست، ۲۰۱۵). علاقه به مدل چهار پارامتری در رشته روان‌شناسی بالینی و شخصیت نیز بوجود آمده است، اندازه‌گیری دقیق صفت پنهان در نمره‌های کرانه‌ای[۵] بسیار مهم است (والر و رایس، ۲۰۱۰). زمینه تحقیق عمومی نیز کاربردهای مدل‌های IRT و بویژه مدل چهار پارامتری را مورد توجه قرار داده است. یک کاربرد عملی جدید مدل چهار پارامتری برای سنجش انطباق کامپیوتری[۶] (CAT) است. به این صورت که به منظور کاهش تاثیر اشتباهات اولیه آزمودنی‌ها روی برآورد سطوح تواناییشان، از مدل ۴ پارامتری استفاده شده است به طوری که مدل چهار پارامتری در مقایسه با مدل سه پارامتری در کاهش دادن چنین تاثیری موثرتر است (لاکن و رولیسن، ۲۰۱۰؛ لیائو[۷] و همکاران،۲۰۱۲). لیائو، هن، ین، و چن (۲۰۱۲) نیز نشان دادند که سطح توانایی با استفاده از مدل چهار پارامتری در طول CAT به اندازه کافی برآورد می‌شود. مدل چهار پارامتری مانع از افت اولیه برآوردهای توانایی ایجاد شده بوسیله پاسخ‌های اشتباه در دو سوال اول می‌شود. مگیس (۲۰۱۳) نشان داد که مدل چهار پارامتری بدلیل وزن تابع درست‌نمایی، برآورد قوی‌تری از توانایی ارائه می‌کند (یعنی به پاسخ‌های اشتباه سوال، وزن کمی داده می‌شود و تاثیر کمتری روی برآورد توانایی دارند). اگر مدل‌های ساده‌تر نظریه سوال پاسخ به داده‌های حاوی پارامتر d برازش داده شوند چه اتفاقی می‌افتد؟ لاکن و رولیسون (۲۰۱۰) نشان دادند که وقتی که مدل سه پارامتری به چنین داده‌های برازش داده می‌شود، پارامتر تشخیص سوال‌ها کاهش می‌یابد. دشواری سوال‌ها به بالاتر منتقل می‌شود (حدود ۰/۵ واحد انحراف استاندارد (SD=۰/۵)). وقتی پارامترهای مدل چهار پارامتری با پارامترهای مدل سه پارامتری مقایسه شدند، پارامترهای C (پارامتر شبه حدس) بدست آمده در مدل سه پارامتری، جذر میانگین مربعات خطای[۸] (RMSE) نسبتا بالاتر و همبستگی نسبتا پایین‌تری با نمره واقعی داشتند. اگر مدل دو پارامتری (به داده‌های حاوی پارامتر d) برازش داده شود، ارزش میانگین پارامترهای تشخیص به اندازه ۰/۵ تغییر کرد و پارامترهای دشواری به صفر نزدیک‌تر می‌شود. براین اساس، اگرچه تغییر بین پارامترها ثابت می‌شود، ولی همبستگی بین توانایی‌های بدست آمده از مدل‌های دو، سه و چهار پارامتری مثبت و بسیار بالا (۰/۹۸)

---

[1] . Magis
[2] . Waller
[3] . Reise
[4] . Bayesian
[5] . Extremes
[6] .. Computer Adaptive Testing
[7] . Liao
[8] . Root-Mean-Squared Error

است. در برآورد پارامترها تحت مدل‌های ساده‌تر IRT سوگیری منظمی وجود ندارد و به طور کلی برآورد صفت فردی تحت هر سه مدل (2PL، 3PL و 4PL) تقریبا یکسان است. برازش مدل‌های ساده‌تر IRT به داده‌های دارای پارامتر d می‌تواند روی تابع آگاهی آزمون (TIF) و مدل‌سازی خطاهای استاندارد تاثیر گذار باشد. وقتی سطح توانایی پایین باشد، سطح آگاهی آزمون با مدل سه پارامتری کمتر برآورد می‌شود. برای سطوح بالای توانایی، آگاهی مبتنی بر مدل سه پارامتری بالا برآورد می‌شود. مدل‌سازی خطاهای استاندارد به دست آمده از مدل سه پارامتری نیز متاثر می‌شوند، از این نظر که برای افراد با توانایی پایین خطاهای استاندارد به اندازه کافی بالا نیستند (فواصل اطمینان بسیار گسترده‌اند). در برازش مدل دو پارامتری به داده‌های به دست آمده از مدل ۴ پارامتری، اگر چه برای مدل 2PL، پارامتر شیب بسیار پایین‌تر برآورد می‌شود، اما سطح آگاهی کلی نسبت به مدل 4PL باز هم بالاتر است. دشواری سوالات به مرکز[1] توزیع کشیده می‌شود و آگاهی در سطوح میانی صفت پنهان انباشته می‌شود. به علاوه چون براساس مدل 2PL نه در کرانه پایین حدس وجود دارد و نه در کرانه بالا برای بی دقتی در پاسخ پارامتری لحاظ می‌شود، پس در کرانه‌های توزیع توانایی کاهش یا از دست رفتن اطلاعات وجود ندارد و چنین به نظر می‌رسد که میزان آگاهی مدل 2PL در کرانه‌ها نیز بیشتر از 4PL است. نتیجه نهایی این است که تابع آگاهی مدل ۲ پارامتری نسبت به مدل ۴ پارامتری بالاتر است و دقت مدل دو پارامتری برای سطوح میانی صفت پنهان در بالاترین حد است (لاکن و رولیسن، ۲۰۱۰). با توجه به اینکه در یک آزمون چندگزینه‌ای پاسخ-های فرد در یکی از دسته‌های زیر قرار می‌گیرد: الف) پاسخ‌هایی که توانایی واقعی فرد را نشان می‌دهد، ب) پاسخ‌های درستی که ناشی از حدس شانسی است و ج) پاسخ‌های نادرست ناشی از اضطراب، بیدقتی و یا حواس پرتی است. پاسخ‌های مربوط به دو مورد ب و ج، باعث خطا در برآورد توانایی می‌شود، زیرا این پاسخ‌ها دانش واقعی آزمودنی‌ها را نشان نمی‌دهند (ین[2] و همکاران، ۲۰۱۲). از آنجایی که مدل ۳ پارامتری برای زمانی که می‌خواهیم بطور صحیح توانایی آزمودنی‌های ضعیف (که با شانس و حدس به سوال‌های دشوار پاسخ درست داده‌اند) را برآورد کنیم مفید است. اما مدل ۳ پارامتری این واقعیت را که ممکن است دانش آموزان با توانایی بالا نیز بخاطر استرس و بی‌دقتی به سوال‌های آسان پاسخ نادرست دهند را نادیده می‌گیرد. در حالی که مدل ۴ پارامتری علاوه بر لحاظ کردن پاسخ‌های درست ناشی از حدس در افراد دارای توانایی ضعیف، پاسخ (یا رفتار) آزمودنی‌های دارای توانایی بالا به هنگام پاسخ نادرست به سوال‌های ساده را با برآورد مجانب بالای منحنی لجستیک (d) در نظر می‌گیرد (وست[3]، ۲۰۱۵). برای مقایسه این دو مدل، تا جایی که بررسی ما نشان داد، دست کم در پژوهش‌های داخلی، مشخص نشده که مدل چهار پارامتری در مقایسه با مدل سه پارامتری چه اطلاعاتی فراهم می‌کند. اینکه افزودن پارامتر چهارم به معادله مدل سه پارامتری به چه نتایجی منجر می‌شود؟ آیا این نتایج به لحاظ عملی کارآمدند؟ برای پاسخ دادن به سوالات فوق لازم است مدل چهار پارامتری در داده‌های واقعی استفاده و نتایج آن با مدل سه پارامتری مقایسه شود. مشخص نشدن نقش پارامتر چهارم (d) و اهمیت عملی آن بر اساس داده‌های واقعی، باعث ابهام در نتایج تحلیل داده‌ها با این مدل خواهد شد. با توجه با مطالب ذکر شده، هدف اصلی پژوهش حاضر مشخص کردن تفاوت ویژگی-های روان‌سنجی مدل ۴ پارامتری (4PL) با مدل ۳ پارامتری حاوی مجانب پایین (3PL) و مدل سه پارامتری حاوی مجانب بالا (3PLu) است (اگر در مدل چهار پارامتری با مجانب پایین و بالا مجانب پایین حذف شود معادله این مدل به دست خواهد آمد. به بیان دیگر اگر به معادله مدل دو پارامتری مجانب بالا اضافه شود معادله این مدل به دست می‌آید). برای کامل بودن مقایسه مدل‌های یک و دو پارامتری نیز در تحلیل لحاظ شده‌اند. لذا، سؤال پژوهش حاضر این است که کدامیک از مدل‌های دارای مجانب (4PL, 3PL, 3PLu) بهترین برازش را با داده‌های مشاهده شده دارد؟ بعلاوه، افزودن پارامتر چهارم چه تاثیری بر آگاهی سوال و آزمون دارد؟ برآورد پارامتر توانایی تحت کدام مدل دقیق‌تر است؟ و در نهایت، چه رابطه‌ای بین نمره خام و توانایی حاصل از هریک از مدل‌ها وجود دارد؟

[1]. Middle
[2]. YenT
[3]. Swist

روش

با توجه به این که این پژوهش به دنبال مقایسه مدل‌ها، بررسی رابطه بین متغیرها و برآورد پارامترها بود، جزو پژوهش‌های توصیفی- همبستگی بود. جامعه آماری این پژوهش شامل همه داوطلبان شرکت‌کننده گروه‌های علوم ریاضی و فنی، تجربی و علوم انسانی در آزمون سراسری سال ۱۳۹۴ بود. از تعداد مجموع کل داوطلبان شرکت کننده، ۱۸۱ هزار و ۸۴۶ نفر در علوم ریاضی، ۴۹۸ هزار و ۸۲۲ نفر در علوم تجربی، ۱۸۲ هزار و ۲۳۹ نفر در علوم انسانی قرار داشتند. اندازه نمونه‌های انتخاب شده در این پژوهش در گروه علوم ریاضی و فنی ۶۲۰۳ نفر، در گروه علوم تجربی ۶۷۹۵ نفر و در گروه علوم انسانی ۶۰۲۴ نفر است. چون داده‌های حاصل از اجرای آزمون‌ها قبلا توسط سازمان سنجش جمع آوری شده طرح جمع‌آوری داده‌های حاصل از اجرای این پژوهش جزو طرحهای ثانویه است. به دلیل زیاد بودن آزمون‌های تخصصی و عمومی هر یک از سه گروه، از آزمون‌های تخصصی گروه تجربی آزمون شیمی، گروه ریاضی آزمون فیزیک و گروه انسانی آزمون ادبیات فارسی انتخاب و داده‌های آنها مورد تحلیل قرار گرفت. علاوه بر این آزمون عمومی معارف اسلامی که در هر یک از گروه‌ها به صورت جداگانه و متفاوت اجرا می‌شود جزو آزمون‌های انتخاب شده بود.

برای بررسی پیش فرض تک‌بعدی بودن مدل‌های استفاده شده در این پژوهش از نرم افزار NOHARM (فراسر و مک دونالد[1]، ۱۹۹۸) استفاده شد. نرم افزار NOHARM با استفاده از نوعی تحلیل عاملی غیرخطی هم به صورت اکتشافی و هم تاییدی میزان برازش داده‌ها با مدل تک‌بعدی را بررسی می‌کند و در صورتی که مدل با داده‌ها برازش مناسب داشته باشد می‌توان آزمون را تک‌بعدی در نظر گرفت. برای تحلیل داده‌ها و پاسخ‌گویی به سوال‌های پژوهش از بسته mirt (چالمرز[2]، ۲۰۱۲، نسخه 1.23) در نرم افزار R (۲۰۱۷، تیم هسته R، نسخه 3.3.3) برای برآورد پارامترهای مدل‌های تک‌بعدی نظریه سوال پاسخ استفاده شده است.

برای بررسی تعداد سوالات برازش یافته با هر یک از مدل‌ها از ملاک کای‌دو، برای مقایسه و انتخاب مناسب‌ترین مدل برازش یافته به داده‌ها از شاخص برازش معیار انحراف اطلاعات[3] (DIC) و عامل بیز[4] (BF) استفاده شد[5]. به طور سنتی از ملاک آکایک[6] (AIC) برای مقایسه و انتخاب مدل استفاده می‌شود و براساس آن مدلی که دارای کمترین مقدار آکایک است به عنوان مناسب‌ترین مدل انتخاب می‌شود. در گروه معیارهای مقایسه‌ای بیزی یکی از ملاک‌های مورد استفاده، معیار انحراف اطلاعات (DIC) است که توسط اشپیگل هارتر[7] و همکاران (۲۰۰۲) ارائه شده است. بر اساس این معیار نیز مدلی که دارای کمترین مقدار DIC باشد، به عنوان بهترین مدل انتخاب می‌شود. عامل بیز (BF) یکی دیگر از شاخص‌های مورد استفاده برای مقایسه مدل‌هاست. با فرض وجود دو مدل آماری، صرف نظر از درست بودن یا نبودن آنها، می‌توان با استفاده از رویکرد بیز مدل‌ها را مقایسه و یکی را انتخاب کرد. عامل بیز میزان حمایت از یک مدل نسبت به مدل دیگر توسط داده‌ها را، صرف نظر از درست بودن آنها، به‌صورت کمی به ما نشان می‌دهد. به شرط داده‌ها، عامل بیز، $B_{10} = \frac{p(D|H1)}{p(D|H0)}$، نسبت درست‌نمایی مدل دارای پارامترهای کمتر (یا مدل مربوط به فرضیه جایگزین) بر درست‌نمایی مدل دارای پارامترهای بیشتر (یا مدل مربوط به فرضیه صفر) است، البته ممکن است تعداد پارامترهای دو مدل یکسان باشند، ولی باز هم امکان مقایسه دو مدل با هم براساس این شاخص وجود دارد، چرا که مدل مربوط به فرضیه جایگزین در صورت و مدل مربوط به فرضیه صفر در مخرج قرار گرفته و مقایسه می‌شوند. تحت این شرایط اگر نسبت درست‌نمایی دو مدل بزرگتر از یک باشد نشان می‌دهد که مدل دارای پارامترهای کمتر نسبت به مدل دارای پارامترهای بیشتر، توسط داده‌ها بیشتر حمایت می‌شود. اگر

---

1. Fraser & McDonald
2. Chalmers
3. Deviance Information Criterion
4. Bayes Factor
5. به دلیل این که در برخی آزمون‌های تحلیل شده در این پژوهش تحلیل براساس مدل‌های ۳ و ۴ پارامتری به همگرایی لازم نرسید، پس برای همه مدل‌های استفاده شده پارامترها براساس ارایه مقادیر پیشین (Prior) به دست آمدند. به همین دلیل مقایسه مدل‌ها براساس ملاک‌های مبتنی بر رویکرد بیز صورت گرفت.
6. Akaike Information Criteion
7. Spigelhalter

نسبت مساوی یک باشد نشان می‌دهد که هیچ یک از مدل‌ها بر دیگری برتری ندارد و اگر کوچکتر از یک باشد یعنی داده‌ها از مدل دارای پارامترهای بیشتر حمایت می‌کند (جفری[1]، ۱۹۶۱). برای مقایسه میزان آگاهی دهندگی هر یک از مدل‌ها بیشینه آگاهی تک تک سؤال‌ها، بیشینه آگاهی آزمون و کمینه خطای استاندارد در هر ۵ مدل و هچنین دامنه‌ای از توانایی که بیشترین میزان آگاهی دهندگی آزمون در آن قرار دارد، محاسبه شد. برای مقایسه میانگین توانایی و خطای استاندارد مدل‌ها از تحلیل واریانس و برای بررسی رابطه بین توانایی و خطای استاندارد مدل‌های مختلف از همبستگی پیرسون استفاده شده است.

یافته‌ها

پیش از تحلیل داده‌ها، نتایج سه شاخص مجموع مجذورات باقی مانده‌ها[2](SSR)، ریشه دوم میانگین مجذورات باقیمانده‌ها[3] (RMSE) و تاناکا[4] برای بررسی پیش فرض تک‌بعدی بودن با استفاده از NOHARM حاکی از قابل دفاع بودن این مفروضه در داده‌ها است، چرا که برای همه آزمون‌ها، SSR کوچکتر مساوی ۰/۰۱، RMSR کمتر از ۰/۰۵ و شاخص تاناکا بزرگتر از ۰/۹۵ است. میانگین و انحراف استاندارد شیب، دشواری، مجانب پایین و بالا در جدول (۱) ارایه شده است. همان طور که مشخص است در تمام آزمون‌ها میانگین شیب سوال‌ها به ترتیب براساس مدل ۴ پارامتری، مدل ۳پارامتری با مجانب پایین، ۳ پارامتری با مجانب بالا و ۲ پارامتری به طور تدریجی کاهش یافته، هر چند که این اختلاف‌ها ناچیز است. این در حالی است که میانگین دشواری سوال‌های آزمون‌های مختلف براساس ۵ مدل وضعیت مشخصی ندارد. مثلا در بین مدل‌های دارای مجانب میانگین دشواری مدل ۳ پارامتری مجانب پایین بیشتر از مدل ۴ پارامتری و ۳ پارامتری مجانب بالا است. در آزمون شیمی تجربی، معارف ریاضی و ادبیات انسانی میانگین دشواری مدل یک پارامتری بیشتر از سایر مدل‌ها است. بین مجانب بالای دو مدل ۴ پارامتری و ۳ پارامتری با مجانب بالا تفاوت زیادی وجود ندارد. مجانب پایین مدل‌های ۴ پارامتری و ۳ پارامتری با مجانب پایین به جز آزمون فیزیک ریاضی، معارف ریاضی و معارف تجربی در سایر آزمون‌ها به هم نزدیک است. مجانب پایین مدل چهار پارامتری در این آزمون‌ها بیشتر از مدل ۳ پارامتری با مجانب پایین است.

برای پاسخ‌گویی به این سوال که کدام یک از مدل‌ها بهترین برازش را با داده‌ها دارد، برازش در سطح سوال و مدل مورد بررسی قرار گرفت. تعداد سوال‌های دارای برازش با هر کدام از مدل‌ها به تفکیک آزمون‌ها در سطح معناداری ۰/۰۵ با آزمون کای‌دو در جدول (۱) ارایه شده است. همانطور که مشخص است در همه آزمون‌ها و در بین همه مدل‌ها در کل تعداد سوال‌های کمتری با مدل یک پارامتری برازش دارند. در آزمون فیزیک به ترتیب مدل‌های 3PLU و 2PL، در آزمون معارف گروه ریاضی به ترتیب مدل‌های 3PL و 4PL، در آزمون شیمی به ترتیب مدل‌های 2PL و 3PLU، در آزمون معارف گروه تجربی به ترتیب مدل‌های 3PL و 4PL، در آزمون ادبیات انسانی به ترتیب مدل‌های 3PL و 4PL و در آزمون معارف انسانی به ترتیب مدل‌های 4PL ، 3PL و 3PLU بیشترین تعداد سوال برازش یافته با مدل را دارند. هر چند که صرف نظر از مدل یک پارامتری در کل تفاوت در تعداد سوال-های دارای برازش در آزمون‌های مختلف براساس مدل‌های مختلف زیاد نیست.

[1] . Jeffrey
[2] . Sum of squars of residuals
[3] . Root mean squars of residuals
[4] . Tanaka index of goodness of fit

جدول (۱) میانگین، انحراف استاندارد پارامترهای سوال و تعداد سوال‌های دارای برازش براساس پنج مدل در آزمون‌های مختلف

درس	مدل	A	B	C	D	تعداد سوال برازش یافته
فیزیک	مدل 4PL	۱,۷۹(۰,۵۶)	۱,۷۶(۰,۷۶)	۰,۰۰۴(۰,۰۰۳)	۰,۹۲۵ (۰,۰۰۵)	۳۲
	مدل 3PL	۱,۷۵(۰,۵۳)	۱,۸۴(۰,۷۳)	۰,۰۰۵(۰,۰۰۳)	۱	۳۰
	مدل 3PLU	۱,۷۰(۰,۵۶)	۱,۷۸(۰,۷۹)	۰	۰,۹۲۵ (۰,۰۰۵)	۳۶
	مدل 2PL	۱,۶۷(۰,۵۳)	۱,۸۶(۰,۷۷)	۰	۱	۳۳
	مدل 1PL	۱(۰)	۰,۹۱(۰,۹۳)	۰	۱	۱۰
معارف ریاضی	مدل 4PL	۱,۸۲(۰,۵۱)	۰,۵۵(۰,۸۰)	۰,۰۴۱(۰,۰۳۵)	۰,۹۲۳ (۰,۰۵۲)	۱۹
	مدل 3PL	۱,۶۶(۰,۴۷)	۰,۶۸(۰,۸۵)	۰,۰۳۶(۰,۰۳۳)	۱	۲۰
	مدل 3PLU	۱,۶۰(۰,۵۰)	۰,۵۰(۰,۸۹)	۰	۰,۹۲۴ (۰,۰۶۱)	۱۸
	مدل 2PL	۱,۴۸(۰,۴۷)	۰,۶۵(۰,۹۴)	۰	۱	۱۶
	مدل 1PL	۱(۰)	۰,۷۹(۱,۱۶)	۰	۱	۸
شیمی	مدل 4PL	۱,۵۵(۰,۳۷)	۱,۹۳(۰,۸۴)	۰,۰۰۴(۰,۰۰۲)	۰,۹۲۳(۰,۰۰۲)	۱۷
	مدل 3PL	۱,۵۱(۰,۳۵)	۲,۰۲(۰,۸۱)	۰,۰۰۴(۰,۰۰۳)	۱	۱۹
	مدل 3PLU	۱,۴۹(۰,۳۶)	۱,۹۵(۰,۸۵)	۰	۰,۹۲۴(۰,۰۰۲)	۲۰
	مدل 2PL	۱,۴۶(۰,۳۴)	۲,۰۴(۰,۸۳)	۰	۱	۲۴
	مدل 1PL	۱(۰)	۲,۸۱(۰,۹۵)	۰	۱	۷
معارف تجربی	مدل 4PL	۱,۳۶(۰,۴۸)	۰,۷۱(۰,۹۴)	۰,۰۳۵(۰,۰۳۹)	۰,۹۲۸(۰,۰۱۷)	۱۵
	مدل 3PL	۱,۲۲(۰,۳۸)	۰,۸۵(۰,۹۶)	۰,۰۲۹(۰,۳۱۹)	۱	۱۵
	مدل 3PLU	۱,۱۹(۰,۳۶)	۰,۶۵(۰,۹۹)	۰	۰,۹۲۹(۰,۰۱۸)	۱۲
	مدل 2PL	۱,۱۰(۰,۳۱)	۰,۸۱(۱,۰۱)	۰	۱	۱۴
	مدل 1PL	۱(۰)	۰,۷۵(۰,۹۶)	۰	۱	۵
ادبیات	مدل 4PL	۱,۲۸(۰,۳۱)	۲,۰۵(۰,۸۶)	۰,۰۱۱(۰,۰۱۰)	۰,۹۲۴ (۰,۰۰۳)	۲۲
	مدل 3PL	۱,۲۵(۰,۳۱)	۲,۱۶(۰,۸۶)	۰,۰۱۱(۰,۰۰۹)	۱	۲۴
	مدل 3PLU	۱,۱۵(۰,۲۵)	۲,۰۸(۰,۸۹)	۰	۰,۹۲۴(۰,۰۰۴)	۲۰

	مدل 2PL	۱,۱۳(۰,۲۴)	۲,۲۰(۰,۸۸)	۰	۱	۲۱
	مدل 1PL	۱(۰)	۲,۳۵(۰,۷۲)	۰	۱	۹
معارف انسانی	مدل 4PL	۱,۳۲(۰,۶۵)	۱,۳۵(۱,۴۲)	۰,۰۲۰(۰,۰۱۴)	۰,۹۲۴(۰,۰۱۷)	۱۰
	مدل 3PL	۱,۲۳(۱,۲۳)	۱,۵۳(۱,۵۳)	۰,۰۱۹(۰,۰۱۹)	۱	۹
	مدل 3PLU	۱,۱۵(۰,۵۶)	۱,۴۲(۱,۶۴)	۰	۰,۹۲۵(۰,۰۱۵)	۹
	مدل 2PL	۱,۰۸(۰,۵۴)	۱,۵۹(۱,۶۵)	۰	۱	۸
	مدل 1PL	۱(۰)	۱,۳۳(۱,۰۴)	۰	۱	۰

مقایسه برازش مدل‌ها با یکدیگر بر اساس شاخص‌های DIC (جدول ۲) نشان داد در همه آزمون‌ها بجز آزمون شیمی، مدل ۲ پارامتری به عنوان مناسبترین مدل و مدل یک پارامتری به عنوان نامناسب‌ترین مدل است. وضعیت برازش مدل‌های دارای مجانب (3PLU، 3PL و 4PL) براساس شاخص DIC به این قرار است: در آزمون فیزیک به ترتیب مدل‌های 3PLU، 3PL و 4PL، در آزمون معارف ریاضی به ترتیب مدل‌های 3PL، 3PLU و 4PL، در آزمون شیمی تجربی به ترتیب مدل‌های 3PLU، 4PL و 3PL، در آزمون معارف تجربی به ترتیب مدل‌های 3PL، 4PL و 3PLU، در آزمون ادبیات انسانی به ترتیب مدل‌های 3PLU، 3PL و 4PL و در آزمون معارف انسانی به ترتیب مدل‌های 3PL، 4PL و 3PLU کمترین مقادیر DIC را دارند. مدلی که کمترین مقدار DIC را داشته باشد بعنوان بهترین و مناسبترین مدل محسوب می‌شود.

جدول (۲) مقادیر DIC آزمون‌ها بر اساس مدل‌های مختلف (کمترین مقدار DIC بترتیب از راست به چپ)

	DIC	DIC	DIC	DIC	DIC
فیزیک ریاضی	2PL(160295.4)	3PLU(160330.7)	3PL(160804)	4PL(160831.5)	1PL(163218)
معارف ریاضی	2PL(154546.9)	3PL(154667.6)	3PLU(154776.7)	4PL(154863.2)	1PL(156460.4)
شیمی تجربی	3PLU(139878)	2PL(139883.2)	4PL(140285.1)	3PL(140296)	1PL(141019.3)
معارف تجربی	2PL(183772)	3PL(183796.3)	4PL(183893.6)	3PLU(183911)	1PL(141019.3)
ادبیات انسانی	2PL(124519.6)	3PLU(124542.8)	3PL(124685.6)	4PL(124702.3)	1PL(125029)
معارف انسانی	2PL(143074.5)	3PL(143063.5)	4PL(143082.2)	3PLU(143111.4)	1PL(145676.4)

جدول (۳) مقادیر عامل بیز (BF) و مدل برتر برای مقایسه هر زوج مدل

مقایسه مدل‌ها	فیزیک ریاضی	معارف ریاضی	شیمی تجربی	معارف تجربی	ادبیات انسانی	معارف انسانی
(4PL, 3PL)	۰	۴/۰۳۵	۰	BF>۱۰۰	۰	۰
مدل برتر	4PL	3PL	4PL	3PL	4PL	4PL
(4PL, 3PLu)	BF>۱۰۰	BF>۱۰۰	۱/۵۲۵	۰	۳/۹۴۸	۰
مدل برتر	3PLu	3PLu	3PLu	4PL	3PLu	4PL
(4PL, 2PL)	BF>۱۰۰	۹/۵۲۵	۷/۲۶۱	BF>۱۰۰	۴/۱۲۰	۰
مدل برتر	2PL	2PL	2PL	2PL	2PL	4PL
(4PL, 1PL)	۰	۰	۰	۰	۰	۰
مدل برتر	4PL	4PL	4PL	4PL	4PL	4PL
(3PL, 3PLu)	۰	۴/۸۷۰	۰	۷/۷۱۸	۰	BF>۱۰۰
مدل برتر	3PLu	3PL	3PLu	3PL	3PLu	3PL
(3PL, 2PL)	BF>۱۰۰	۲/۳۶۱	۲/۶۸۴	۰	۱/۰۲۸	۰
مدل برتر	2PL	2PL	2PL	3PL	2PL	3PL
(3PL, 1PL)	۰	۰	۰	۰	۰	۰
مدل برتر	3PL	3PL	3PL	3PL	3PL	3PL
(3PLu, 2PL)	۰	۱/۱۵۰	۰	۲/۰۸۶	۰	۰/۰۰۱
مدل برتر	2PL	3PLu	2PL	3PLu	2PL	2PL
(3PLu, 1PL)	۰	۰	۰	۰	۰	۰
مدل برتر	3PLu	3PLu	3PLu	3PLu	3PLu	3PLu
(2PL, 1PL)	۰	۰	۰	۰	۰	۰
مدل برتر	2PL	2PL	2PL	2PL	2PL	2PL

* مقادیر صفر به برتری مدل حاوی پارامتر بیشتر در مخرج عامل بیز قرار گرفته اشاره دارد

مقایسه مدل‌ها بر اساس عامل بیز (BF) در جدول (۳) نشان می‌دهد، که همانند نتیجه شاخص DIC، در همه آزمون‌ها به جز شیمی، مدل 2PL مناسب‌ترین مدل و مدل 1PL نامناسب‌ترین مدل است. از ۱۸ مقایسه‌ای که بین ۳ مدل 4PL، 3PLu و 3PL صورت گرفت، در ۱۵ مقایسه نتایج عامل بیز هم راستا با نتایج شاخص DIC است، اما در ۳ مورد (در آزمون‌های فیزیک

ریاضی، ادبیات انسانی و معارف انسانی) با توجه به عامل بیز شواهد قطعی مبنی بر اینکه مدل 4PL بهتر از مدل 3PL است وجود دارد، در حالی که نتیجه DIC در این سه مورد معکوس است. مقایسه توانایی برآورد شده هر پنج مدل در هر شش آزمون با تحلیل واریانس نشان داد به لحاظ آماری تفاوت معناداری بین توانایی برآورد شده مدل‌ها وجود ندارد (جدول (۴))، ولی بین خطای استاندارد اندازه‌گیری مدل‌ها تفاوت معناداری وجود دارد (جدول (۴)).

جدول (٤) نتایج تحلیل واریانس بر روی نمره های توانایی و خطاهای استاندارد اندازه‌گیری ۵ مدل به تفکیک آزمون

آزمون		SS	df	MS	F	سطح معنی‌داری
فیزیک	توانایی برآورد شده	۱/۲۹۷۰	۴	۰/۳۲۴۳	۰/۲۹۵	۰/۸۸۱
	خطای استاندارد	۶۰/۹۳	۴	۲۴۳/۷	۱۹۰۴	۰/۰۰۰۱
معارف ریاضی	توانایی برآورد شده	۰/۰۷۶۰	۴	۰/۰۱۹۰	۰/۰۱۷	۰/۹۹۹
	خطای استاندارد	۱۲۳/۷	۴	۲۰/۹۲۶	۴۸۶۲	۰/۰۰۰۱
شیمی	توانایی برآورد شده	۰/۳۱۵۲	۴	۰/۰۷۸۸	۰/۰۸۲	۰/۹۸۸
	خطای استاندارد	۱۸۶/۸	۴	۴۶/۷۱	۱۸۰۳	۰/۰۰۰۱
معارف تجربی	توانایی برآورد شده	۰/۰۱۷۶	۴	۰/۰۰۴۴	۰/۰۰۵	۱
	خطای استاندارد	۱۰/۲۸	۴	۲/۵۶۹۹	۸۲۱/۲	۰/۰۰۰۱
ادبیات	توانایی برآورد شده	۰/۰۶۲۴	۴	۰/۰۱۵۶	۰/۰۲۱	۰/۹۹۹
	خطای استاندارد	۱۲	۴	۲/۹۹۲۱	۲۱۱/۱	۰/۰۰۰۱
معارف انسانی	توانایی برآورد شده	۰/۱۱۶۰	۴	۰/۰۲۹۰	۰/۰۳۹	۰/۹۹۷
	خطای استاندارد	۱۸۶/۸	۴	۴۶/۷۱	۱۸۰۳	۰/۰۰۰۰۱

نتایج آزمون تعقیبی توکی در جدول (۵) نشان می‌دهد که در آزمون‌های فیزیک ریاضی، معارف ریاضی، شیمی تجربی، معارف تجربی و ادبیات انسانی فقط تفاوت بین مدل‌های ۳ و ۴ و نیز مدل‌های ۲ و ۳ پارامتری با مجانب بالا معنادار نیست. در آزمون معارف انسانی این نتیجه فرق دارد و بجز مدل ۳ و ۱ بین سایر مدل‌ها تفاوت معنادار است.

جدول (۶) همبستگی نمره خام هریک از آزمون‌ها با نمره توانایی مربوط به همان آزمون براساس مدل‌های مختلف را نشان می-دهد. با توجه به نتایج جدول ()۶، همبستگی بین نمرات خام و نمره توانایی مربوط به مدل‌های مختلف در آزمون فیزیک ریاضی برابر ۰/۹۳، در آزمون معارف ریاضی برابر ۰/۹۹، در آزمون شیمی تجربی تقریبا ۰/۹۵، در آزمون معارف تجربی برابر ۰/۹۹، در آزمون ادبیات تقریبا برابر ۰/۹۶ و در آزمون معارف انسانی برابر ۰/۹۷ است. جدول (۷) همبستگی توانایی هر زوج از مدل‌ها را نشان می‌دهد. در همه آزمون‌ها بجز آزمون معارف انسانی همبستگی بین مدل‌ها بزرگتر یا مساوی ۰/۹۹ است. در آزمون معارف انسانی نیز رابطه بین مدل یک پارامتری با سایر مدل‌ها برابر ۰/۹۶ و همبستگی بین سایر مدل‌ها با یکدیگر مساوی ۰/۹۹ یا بزرگتر است. برای مقایسه میزان آگاهی هر یک از آزمون‌ها براساس مدل‌های مختلف، بیشینه آگاهی سوال و بیشینه آگاهی آزمون مورد بررسی قرار گرفت. با توجه به شکل (۱) راست در آزمون فیزیک ریاضی، مقادیر اگاهی اکثر سوال‌ها در بازه ۰/۴ تا ۱/۴ است و سوال‌های میانی نسبت

سوال‌های دو انتهای آزمون آگاهی بیشتری دارند. آگاهی اکثر سوال‌های تحت مدل 3PL نسبت به سایر مدل‌ها بیشتر و تحت مدل 1PL نسبت به سایر مدل‌ها کمتر است. بعد از مدل 3PL، آگاهی اکثر سوال‌ها به ترتیب تحت مدل‌های 4PL، 2PL و 3PLU بیشتر است. همچنین آگاهی سوال‌های ۱۷، ۱۸، ۳۳ و ۳۵ بیشتر از سایر سوال‌ها است. سوال‌ها ۵، ۱۲، ۱۹، ۲۰، ۴۴ و ۴۵ آگاهی کمتری دارند.

در آزمون معارف ریاضی (شکل (۲) چپ) مقادیر آگاهی اکثر سوال‌ها در بازه ۰/۳ تا ۰/۹ قرار دارد. در اکثر سوال‌ها آگاهی ایجاد شده بر اساس مدل 4PL و 3PL اندکی بیشتر از سایر مدل‌ها است. بطور کلی مقادیر آگاهی سوال‌ها تحت مدل‌های مختلف (به جز مدل 1PL، که آگاهی سوال‌ها تحت آن برابر ۰/۲۵ است) نزدیک به هم است. سوال‌های ۴ و ۱۵ آگاهی بیشتر و سوال‌های ۶، ۱۱ و ۱۴ آگاهی کمتری دارند.

جدول (۵) نتایج آزمون تعقیبی توکی برای آزمون‌های فیزیک، شیمی و ادبیات معارف ریاضی، تجربی و انسانی

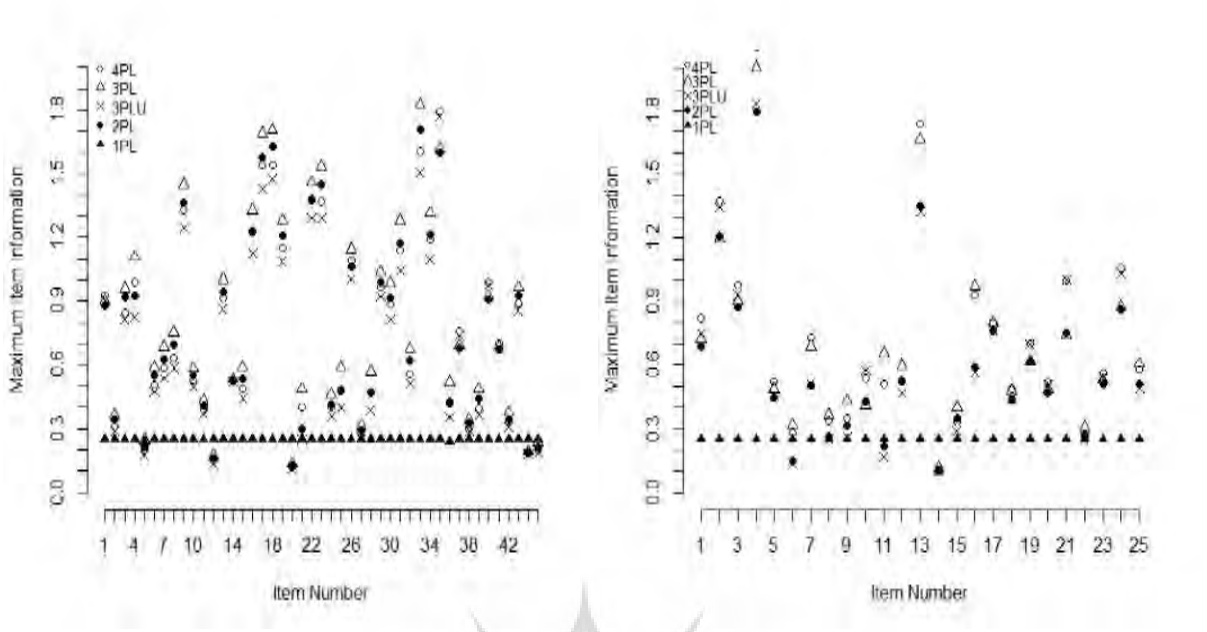
مقایسه زوجی	Diff	p adj	Diff	p adj	Diff	p adj	Diff	p adj	Diff	p adj	Diff	p adj
3PL-4PL	-۰/۰۰۳	۰/۰۶	۰/۰۰۲	۰/۹۲	۰/۰۰۱	۰/۸۹	-۰/۰۰۳	۰/۰۶	۰/۹۹	۰/۹۹	۵/۹۴	۰/۰۰۲
3PLu-4PL	-۰/۰۱	۰/۰۰	-۰/۰۱	۰/۰۰	-۰/۰۱۷	۰/۰۰	-۰/۰۱۱	۰/۰۰	۰/۰۰	۰/۰۰	-۱/۲۵	۰/۰۰۰
2PL-4PL	-۰/۰۱	۰/۰۰	-۰/۰۰۹	۰/۰۰	-۰/۰۱۶	۰/۰۰	-۰/۰۱۴	۰/۰۰	۰/۰۰	۰/۰۰	-۸/۰۲	۰/۰۰۰
1PL-4PL	-۰/۱۴	۰/۰۰	۰/۱۷	۰/۰۰	۰/۰۳۷	۰/۰۰	۰/۱۴۹	۰/۰۰	۰/۰۰	۰/۰۰	۶/۰۲	۰/۰۰
3PLu-3PL	-۰/۰۰۸	۰/۰۰	-۰/۰۱	۰/۰۰	-۰/۰۱	۰/۰۰	-۰/۰۰۸	۰/۰۰	۰/۰۰	۰/۰۰	-۱/۸۴	۰/۰۰
2PL-3PL	-۰/۰۱	۰/۰۰	-۰/۰۱	۰/۰۰	-۰/۰۱۸	۰/۰۰	-۰/۰۱۰	۰/۰۰	۰/۰۰	۰/۰۰	-۱/۳۹	۰/۰۰۰
1PL-3PL	-۰/۱۵	۰/۰۰	۰/۱۷	۰/۰۰	۰/۰۳۵	۰/۰۰	۰/۱۵	۰/۰۰	۰/۰۰	۰/۰۰	۸/۰۸	۰/۹۹
2PL-3PLu	-۰/۰۰۲	۰/۳۹	۰/۰۰۲	۰/۹۱	۰/۰۰۱	۰/۹۰	-۰/۰۰۲	۰/۳۹	۰/۹	۰/۹	۴/۴۸	۰/۰۴۲
1PL-3PLu	-۰/۱۶	۰/۰۰	۰/۱۹	۰/۰۰	۰/۰۵۵	۰/۰۰	۰/۱۶	۰/۰۰	۰/۰۰	۰/۰۰	۱/۸۵	۰/۰۰
1PL-2PL	-۰/۱۶	۰/۰۰	۰/۱۸	۰/۰۰	۰/۰۵۳	۰/۰۰	۰/۱۶	۰/۰۰	۰/۰۰	۰/۰۰	۱/۴۰	۰/۰۰

جدول (٦) همبستگی نمره خام هریک از آزمون‌ها با نمره توانایی همان آزمون براساس مدل‌های مختلف

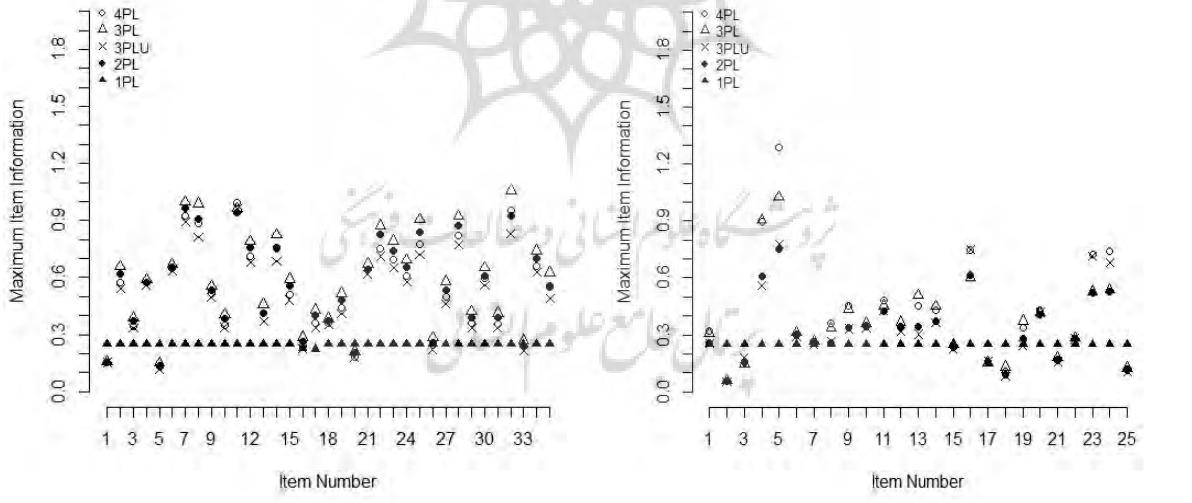
نمره توانایی / نمره خام	4PL	3PL	3PLU	2PL	1PL
فیزیک	۰/۹۳	۰/۹۳	۰/۹۳	۰/۹۳	۰/۹۳
معارف ریاضی	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹
شیمی	۰/۹۵	۰/۹۵	۰/۹۵	۰/۹۴	۰/۹۵
معارف تجربی	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹
ادبیات	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۷
معارف انسانی	۰/۹۷	۰/۹۷	۰/۹۷	۰/۹۷	۰/۹۹

جدول (۷) همبستگی نمره‌های توانایی مدل‌ها در آزمون‌های مختلف

	فیزیک	معارف ریاضی	شیمی	معارف تجربی	ادبیات	معارف انسانی
4PL-3PL	۱	۱	۱	۱	۱	۱
4PL-3PLU	۱	۱	۱	۱	۱	۱
4PL-2PL	۱	۱	۱	۱	۱	۰/۹۹
4PL-1PL	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۶
3PL-3PLU	۱	۱	۱	۱	۱	۱
3PL-2PL	۱	۱	۱	۱	۱	۱
3PL-1PL	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۶
3PLU-2PL	۱	۱	۱	۱	۱	۱
3PLU-1PL	۰/۹۹	۰/۹۹	۰/۹۹	۰/۹۹	۱	۰/۹۶
2PL-1PL	۰/۹۹	۰/۹۹	۱	۰/۹۹	۱	۰/۹۶



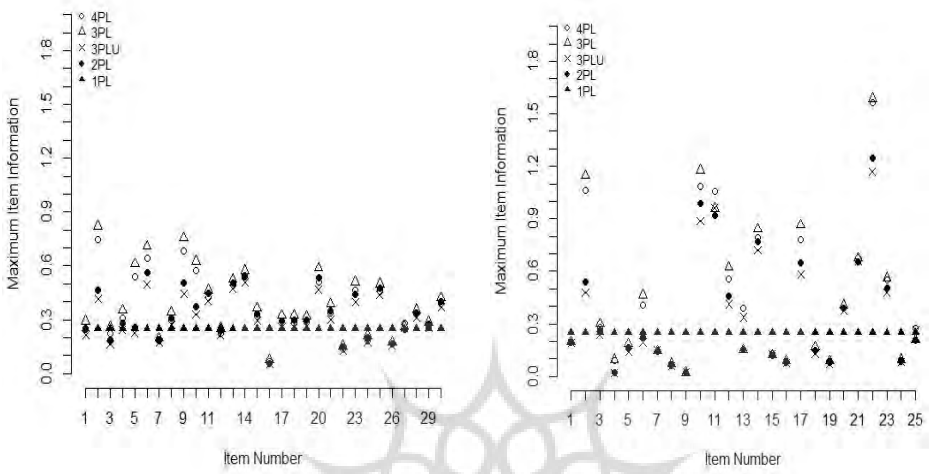
شکل (۱) بیشینه آگاهی سوال‌های فیزیک ریاضی (راست) و معارف ریاضی (چپ) براساس ۵ مدل در آزمون شیمی تجربی (شکل (۲) راست) مقادیر آگاهی همه سوالات در بازه ۰/۱ تا ۱/۲ قرار دارد.



شکل (۲) بیشینه آگاهی سوال‌های شیمی تجربی (راست) و معارف تجربی (چپ) براساس ۵ مدل

در اکثر سوال‌ها مدل ۳ پارامتری نسبت به سایر مدل‌ها آگاهی بیشتری نشان می‌دهد. بعد از مدل ۳ پارامتری بیشترین آگاهی را در اکثر سوال‌ها به‌ترتیب مدل‌های ۲ پارامتری، ۴ پارامتری، 3PLu و مدل ۱ پارامتری فراهم می‌کنند. سوال‌های ۳۲، ۷ و ۸ به ترتیب دارای بیشترین مقادیر آگاهی هستند. در مقابل سوال‌های ۵، ۱، ۲۰ بترتیب دارای کمترین مقادیر آگاهی هستند. البته اختلاف آگاهی سوال مدل‌های مختلف در ۱۸ سوال دوم اندکی بیشتر از ۱۸ سال اول است.

در آزمون معارف تجربی (شکل (۲) چپ) مقادیر آگاهی اکثر سوال‌ها در بازه ۰/۲ تا ۰/۷ است. مقادیر آگاهی در سوال‌های میانی تحت مدل‌های مختلف تقریبا نزدیک بهم است و در سوالات پایانی و آغازین مقدار آگاهی سوال‌ها کمتر است. بجز سوال‌های ۴ و ۵، در سایر سوال‌ها، آگاهی تحت مدل‌های مختلف نزدیک به هم است. سوال‌های ۵ و ۴ به ترتیب دارای بیشترین مقادیر آگاهی و در مقابل سوال‌های ۲ و۱۸ بترتیب دارای کمترین مقادیر آگاهی هستند.



شکل (۳) بیشینه آگاهی سوال‌های ادبیات انسانی (راست) و معارف انسانی (چپ) براساس ۵ مدل

در آزمون ادبیات انسانی (شکل (۳) راست) در اکثر سوال‌ها مدل ۳ پارامتری نسبت به سایر مدل‌ها آگاهی بیشتری را نشان می‌دهد. بعد از مدل ۳ پارامتری بیشترین آگاهی اکثر سوال‌ها به‌ترتیب مربوط به مدل‌های 4PL، 2PL، 3PLu و 1PL است. مقدار آگاهی در ۱۵ سوال اول نسبت به ۱۵ سوال دوم بیشتر است. سوال۲ دارای بیشترین مقدار آگاهی و در مقابل سوال ۱۶ دارای کمترین مقدار آگاهی است. مقادیر آگاهی اکثر سوال‌های آزمون ادبیات انسانی در بازه ۰/۲۵ تا ۰/۶ است و آگاهی هیچ یک از سوالات از ۰/۹ بالاتر نیست. در آزمون معارف انسانی (شکل (۳) چپ) آگاهی بیشتر سوال‌ها در دامنه ۰/۰۲ تا ۰/۶ قرار دارد. بیشینه آگاهی نصف سوال‌ها کمتر ۰/۴ است و آگاهی اکثر سوال‌ها نزدیک بهم برآورد شده است. آگاهی سوال تحت مدل‌های مختلف در سوال‌ها آغازین و پایانی تقریبا نزدیک به هم است. در اکثر سوال‌ها مدل ۳ پارامتری نسبت به سایر مدل‌ها آگاهی بیشتری را نشان می‌دهد. بعد از مدل ۳ پارامتری بیشترین آگاهی را در اکثر سوال‌ها به‌ترتیب مدل‌های ۲ پارامتری، ۴ پارامتری، ۳ پارامتری با مجانب بالا فراهم می‌کنند. سوال ۲۲ بیشترین مقدار آگاهی و سوال ۴ کمترین مقدار آگاهی را دارند.

بیشینه آگاهی آزمون به تفکیک آزمون و مدل در جدول (۸) ارایه شده است. همانطور که انتظار می‌رود بیشینه آگاهی درس‌های فیزیک و شیمی نسبت به سایر درس‌ها بالاتر است، زیرا این دو آزمون تعداد سوالات بیشتری دارند. آگاهی آزمون معارف ریاضی نسبت به معارف گروه تجربی و انسانی (همگی با ۲۵ سوال) بیشتر است، که نشان می‌دهد آین آزمون سوال‌های مناسب‌تری دارد. همچنین در میان مدل‌های مختلف، مدل 3PL آگاهی آزمون نسبتا بیشتری دارد و بیشینه آگاهی آن در مقایسه با مدل‌های دیگر به سمت راست توزیع توانایی حرکت کرده است (تقریبا ۰/۵ واحد). همچنین آگاهی آزمون در مدل 1PL در همه دروس نسبت به سایر مدل‌ها کمتر است. مقادیر آگاهی آزمون مدل‌های 2PL و 4PL تقریبا نزدیک به هم بوده و شکل تابع آگاهی آنها شبیه به هم است. همچنین در مدل 3PLU بیشینه آگاهی آزمون به میانگین توزیع توانایی (صفر) نزدیک شده است. در همه آزمون‌ها در سطوح توانایی بالا (۴/۵+ تا ۲+) مدل‌های ۱، ۲ و ۳ پارامتری نسبت به مدل‌های ۴ پارامتری و ۳ پارامتری با مجانب بالا آگاهی بیشتری

فراهم می‌کنند و در سطوح توانایی خیلی پایین نیز مدل‌های ۳ و ۴ پارامتری نسبت به سایر مدل‌ها خطای استاندارد اندازه‌گیری بیشتری دارند (البته این اختلاف خطا بسیار کم است). شکل ۱ منحنی آگاهی آزمون فیزیک را به عنوان نمونه نشان می‌دهد.

جدول(۸) بیشینه آگاهی آزمون (و دامنه تقریبی توانایی مطابق با آن) براساس هریک از مدل‌ها در آزمون‌های مختلف

آزمون	4PL(دامنه توانایی)	3PL(دامنه توانایی)	3PLu(دامنه توانایی)	2PL(دامنه توانایی)	1PL(دامنه توانایی)
فیزیک ریاضی	۲۷/۰۳ (۱/۵–۲)	۳۰/۹۳(۱/۵–۲/۵)	۲۵/۳۱ (۱/۵–۲)	۲۸/ ۸۹(۱/۵–۲/۵)	۹/۴۱ (۲/۵–۳/۵)
معارف ریاضی	۱۲/۸۷ (۰–۱)	۱۲/۸۱(۰–۱)	۱۱/ ۶۸(۰–۵)	۱۱/ ۵۴(۰–۱)	۴/ ۹۶(۰/۵–۱/۵)
شیمی تجربی	۱۵/۱۷ (۱/۵–۲)	۱۷/۵۱(۱/۷۵–۲/۲۵)	۱۴/ ۲۶(۱/۵–۲)	۱۶/ ۴۴(۱/۷۵–۲/۲۵)	۷/۳۷ (۲/۲۵–۳/۲۵)
معارف تجربی	۷/۵۱ (۰/۲۵–۱/۲۵)	۷/۴۸(۰/۵–۱/۵)	۶/۶ (-۰/۷۵–۷۵۰)	۶/ ۶۰(۰/۲۵–۱/۲۵)	۵/۳۱ (۰/۵–۱/۵)
ادبیات انسانی	۹/۴۹ (۰/۵–۱/۵)	۱۰/۷۶(۱–۲)	۷/۸۱ (۰/۵–۱/۵)	۸/ ۸۷(۰/۵–۱/۵)	۶/۷۶ (۱–۲)
معارف انسانی	۸/۵۵ (۰/۷۵–۱/۲۵)	۹/۴۱(۱–۲)	۷/۱۹ (۰/۵–۱/۲۵)	۷/۹ (۰/۷۵–۱/۲۵)	۵/۱۷ (۱–۲)

بیشینه آگاهی آزمون فیزیک ریاضی براساس مدل‌های 3PL، 2PL، 4PL، 3PLu و 1PL به ترتیب برابر ۳۰/۹۳، ۲۸/۸۹، ۲۷/۰۳، ۲۵/۳۱ و ۹/۴۱ است. مقادیر بیشینه آگاهی برای مدل یک پارامتری تقریبا در سطح توانایی ۳+ و برای سایر مدل‌ها در دامنه توانایی ۱/۵+ تا ۲/۵+ قرار دارد. بیشینه آگاهی آزمون معارف ریاضی براساس هر یک از مدل‌های یک پارامتری، دو پارامتری، سه پارامتری، سه پارامتری با مجانب بالا و چهار پارامتری به ترتیب ۴/۹۶، ۱۱/۵۴، ۱۲/۸۱، ۱۱/۶۸ و ۱۲/۸۷ است که این مقادیر بیشینه در همه مدل‌ها تقریبا در دامنه توانایی ۰ تا ۱/۵+ قرار دارد. مقادیر آگاهی آزمون مدل‌های 4PL، 3PL، 3PLu و 2PL تقریبا نزدیک بهم است. مطابق انتظار آگاهی آزمون در مدل یک پارامتری کمتر از سایر مدل‌ها است.

در آزمون شیمی بیشینه آگاهی هریک از مدل‌ها ۱پارامتری، ۲ پارامتری، ۳ پارامتری، ۳ پارامتری با مجانب بالا و ۴ پارامتری به ترتیب ۷/۳۷، ۱۶/۴۴، ۱۷/۵۱، ۱۴/۲۶ و ۱۵/۱۷ است که نشان می‌دهد بیشینه آگاهی آزمون مدل ۳ پارامتری از سایر مدل‌ها بالاتر است. این مقادیر بیشینه در همه مدل‌ها تقریبا در دامنه توانایی ۲/۵+ تا ۱/۵+ قرار دارد. همان طور که مشاهده می‌شود آگاهی آزمون مبتنی بر مدل‌های 4PL، 3PL، 3PLu و 2PL تقریبا نزدیک بهم است.

در آزمون معارف تجربی بیشینه آگاهی آزمون مربوط به مدل ۴ پارامتری برابر ۷/۵۱ است. بعد از مدل ۴ پارامتری، بیشینه آگاهی آزمون سایر مدل‌ها بترتیب به مدل‌های ۳پارامتری، ۳ پارامتری با مجانب بالا، دو پارمتری و یک پارامتری به ترتیب با مقادیر ۷/۴۸، ۶/۶۰۷، ۶/۶۰۳ و ۵/۳۱ مربوط است که به هم نزدیک‌اند. مقادیر بیشینه آگاهی همه مدل‌ها در دامنه توانایی صفر تا ۲+ قرار دارد.

در آزمون ادبیات انسانی نیز بیشینه آگاهی آزمون مربوط به مدل ۳ پارامتری برابر ۱۰/۷۶ است. بعد از مدل ۳ پارامتری بیشینه آگاهی به ترتیب به مدل‌های ۴پارامتری، ۲پارامتری، 3PLu و یک پارامتری با مقادیر ۹/۴۹، ۸/۸۷، ۷/۸۱ و ۶/۷۶ مربوط است. بیشینه آگاهی این مدل‌ها تقریبا در دامنه ۰/۵+ تا ۲ قرار دارد. باز هم می‌توان گفت بیشینه آگاهی آزمون مدل‌های مختلف تقریبا نزدیک بهم است.

در آزمون معارف انسانی مدل ۳ پارامتری نسبت به سایر مدل‌ها بیشینه آگاهی بالاتری دارد. بعد از مدل ۳ پارامتری بیشینه آگاهی به ترتیب به مدل‌های ۴پارامتری، ۲پارامتری، 3PLu و یک پارامتری با مقادیر ۸/۵۵، ۷/۹، ۷/۱۹ و ۵/۱۷ مربوط است. این مقادیر بیشینه تقریبا در سطح توانایی ۰/۵+ تا ۱/۵ قرار دارند.

بحث و نتیجه‌گیری

هدف پژوهش حاضر مقایسه ویژگی‌های روان‌سنجی مدل‌های مختلف دارای مجانب (سه پارامتری با مجانب پایین، سه پارامتری با مجانب بالا و چهار پارامتری با مجانب پایین و بالا) است. یافته‌های پژوهش نشان داد که از مجموع ۱۸۵ سوال (مجموع سوال‌های فیزیک، معارف ریاضی، شیمی، معارف تجربی، ادبیات و معارف انسانی)، مدل ۳ پارامتری با ۱۱۷ سوال (۶۳ درصد) دارای بیشترین برازش است و بعد از آن مدل ۲ پارامتری با ۱۱۶ (۶۲/۵ درصد) و مدل‌های ۴ پارامتری و ۳ پارامتری با مجانب بالا با ۱۱۵ (۶۲ درصد) سوال قرار دارند. البته این اعداد نشان می‌دهند که بین این مدل‌ها از نظر برازش در سطح سوال فقط یک یا دو سوال اختلاف وجود دارد. که می‌توان نتیجه گرفت به لحاظ برازش در سطح سوال، هیچ مدلی بر دیگری برتری ندارد. مدل یک پارامتری نیز از مجموع ۱۸۵ سوال فقط با ۳۹ سوال (تقریبا ۲۱ درصد) برازش دارد. این تعداد سوال برازش یافته با مدل یک پارامتری بسیار کمتر از تعداد سوال برازش یافته با سایر مدل‌ها است.

بررسی برازش سوال‌های آزمون‌های عمومی و اختصاصی بصورت جداگانه نتایج جالبی را بدست می‌دهد. تعداد سوال‌هایی که در آزمون‌های عمومی با هر کدام از مدل‌ها برازش دارند در جدول ۹ آمده است. در آزمون‌های عمومی مدل ۳ و ۴ پارامتری نسبت به سایر مدل‌ها برازش بهتری با سوال‌ها دارند. آزمونهای اختصاصی سوال‌های بیشتری به ترتیب با مدل‌های ۲ و ۳ پارامتری با مجانب بالا برازش دارند. همانطور که مشخص است از مجموع ۷۵ سوال مربوط به آزمونهای عمومی ۴۴ سوال (۵۹ درصد) با مدل‌های ۳ پارامتری و ۴ پارامتری برازش دارند. بعد از این دو مدل، مدل ۳ پارامتری با مجانب بالا با ۳۹ سوال (۵۲ درصد)، مدل ۲ پارامتری با ۳۸ سوال (۵۱ درصد) و مدل ۱ پارامتری با ۱۳ سوال (۱۷/۳) قرار دارند. بنابراین می‌توان نتیجه گرفت در آزمون‌های عمومی مدل ۳ و ۴ پارامتری نسبت به سایر مدل‌ها برازش بهتری با سوال‌ها دارند.

جدول (۹) فراوانی و درصد برازش ۷۵ سوال آزمون‌های عمومی و ۱۱۰ سوال اختصاصی با هرکدام از مدل‌ها

		4PL	3PL	3PLU	2PL	1PL
عمومی	(درصد) فراوانی	۷۱(۶۵)	۷۳(۶۶)	۷۶(۶۹)	۷۸(۷۱)	۲۶(۲۴)
اختصاصی	(درصد) فراوانی	۴۴(۵۹)	۴۴(۵۹)	۳۹(۵۲)	۳۸(۵۱)	۱۳(۱۷/۳)

نتایج شاخص DIC نشان داد که در بین مدل‌های ۳ پارامتری دارای مجانب پایین، ۳ پارامتری با مجانب بالا و ۴ پارامتری، در آزمون‌های عمومی مدل ۳ پارامتری با مجانب پایین و در آزمونهای اختصاصی مدل ۳ پارامتری با مجانب بالا مناسب‌ترین مدل‌ها هستند. بعلاوه با توجه شاخص DIC مدل ۱ پارامتری در همه آزمون‌ها از نظر برازش در رده آخر قرار دارد. شاخص دیگری که برای مقایسه برازش مدل‌ها استفاده شد، عامل بیز است که براساس آن اکثریت نتایجی که با DIC بدست آمده بود تایید شد. البته در آزمون‌های فیزیک، ادبیات و معارف انسانی با توجه به عامل بیز شواهد قطعی مبنی بر اینکه مدل ۴ پارامتری بهتر از ۳ پارامتری (3PL) است وجود دارد، در حالی که در این سه مورد شاخص DIC نتیجه برعکسی را نشان داد. بنابراین با توجه به شاخص‌های DIC، عامل بیز و همچنین تعداد سوال‌هایی که با هر کدام از مدل ها برازش دارند می‌توان نتیجه گرفت که از بین مدل‌های 4PL، 3PL و 3PLu در آزمون‌های عمومی بترتیب مدل‌های ۳ پارامتری با مجانب پایین، ۴ پارامتری و ۳ پارامتری با مجانب بالا بهترین برازش را دارند و در آزمون‌های تخصصی بترتیب مدل‌های ۳ پارامتری با مجانب بالا، ۳ پارامتری با مجانب پایین و ۴ پارامتری بهترین برازش را دارند.

بررسی بیشینه آگاهی هر سوال در آزمون‌های مختلف نشان داد که مدل ۳ پارامتری در اکثر سوال‌ها نسبت به سایر مدل‌ها بیشینه آگاهی سوال بالاتری دارد. طبق انتظار بیشینه آگاهی سوال و آزمون مدل یک پارامتری کمتر از سایر مدل‌هاست. نتایج نشان داد که در آزمون‌های فیزیک، شیمی، ادبیات و معارف انسانی مدل ۳ پارامتری (3PL) بیشینه آگاهی آزمون بالاتری نسبت به سایر

مدل‌ها دارد. در آزمونهای معارف ریاضی و معارف تجربی بیشینه آگاهی آزمون مدل ۴ پارامتری از سایر مدل‌ها بالاتر است که با مدل ۳ پارامتری (3PL) اختلاف اندکی دارد. بطور کلی می‌توان نتیجه گرفت که در آزمونهای اختصاصی بیشنه آگاهی آزمون مدل ۴ پارامتری کمتر از مدل ۳ پارامتری است. و در آزمونهای عمومی بیشینه آگاهی مدل ۴ پارامتری اندکی بیشتر از مدل ۳ پارامتری است (البته این اختلاف خیلی ناچیز است). در همه آزمون‌ها منحنی آگاهی آزمون مدل‌های ۱، ۲ و ۳ پارامتری نسبت به منحنی آگاهی آزمون مدل‌های ۳ پارامتری با مجانب بالا و ۴ پارامتری به سمت راست پیوستار توانایی تغییر مکان می‌دهد. همچنین در سطوح توانایی بالا مدل‌های ۱، ۲ و ۳ پارامتری نسبت به مدل‌های ۴ پارامتری و ۳ پارامتری با مجانب بالا آگاهی بیشتری ایجاد می‌کنند. و در سطوح توانایی پایین نیز مدل‌های ۳ و ۴ پارامتری نسبت به سایر مدل‌ها خطای استاندارد اندازه‌گیری بیشتری دارند (البته این اختلاف خطا بسیار کم است). در سطوح توانایی بالا نیز مدل ۴ پارامتری نسبت به مدل‌های ۳ ، ۲ و ۳ پارامتری با مجانب بالا خطای استاندارد اندازه‌گیری بیشتری دارد. نتایج یافته‌های این بخش تا حدی هم راستا با یافته‌های لاکن و رولیسن (۲۰۱۰) است که نشان دادند در سطوح توانایی بالا آگاهی مبتنی بر مدل سه پارامتری نسبت به مدل ۴ پارامتری کمی بیشتر است. ولی بر خلاف یافته‌های لاکن و رولیسون در سطوح پایین بین دو مدل اختلاف چندانی در میزان آگاهی نیست.

مقایسه توانایی برآورد شده در هر شش آزمون نشان داد که تفاوت معناداری بین توانایی برآورد شده مدل‌ها وجود ندارد. بین توانایی هر زوج مدل همبستگی بالا وتقریبا نزدیک به ۱ وجود دارد. لاکن و رولیسون (۲۰۱۰) نیز در مطالعه شبیه سازیشان نشان دادند که همبستگی بین توانایی‌های بدست آمده از مدل‌های دو، سه و چهار پارامتری مثبت و بسیار بالا (۰/۹۸) است. با این وجود بین خطای استاندارد اندازه‌گیری مدل‌ها به لحاظ آماری تفاوت معناداری وجود دارد. آزمون تعقیبی توکی نشان داد که در آزمون‌های فیزیک، معارف ریاضی، شیمی، معارف تجربی، ادبیات فقط تفاوت بین مدل‌های ۳ و ۴ و همچنین مدل‌های ۲ و ۳ پارامتری با مجانب بالا معنادار نیست و بین سایر زوج مدل‌ها تفاوت معنادار وجود دارد. در آزمون معارف انسانی این نتیجه فرق می‌کرد و بجز مدل ۳ و ۱ بین سایر مدل‌ها تفاوت وجود داشت. همچنین نتایج نشان داد که طبق انتظار یک نوع رابطه خطی بین نمره توانایی هر یک از مدل-ها با نمرات خام مربوطه در همه آزمون‌ها وجود دارد و همبستگی بین نمره توانایی و نمرات خام بسیار بالا است. با توجه به نتایج حاصل از تحلیل پیشنهادهای زیر برای افزایش کارایی سوال در آزمون سراسری مطرح می‌شود. ۱) با توجه به برازش زیاد سوال‌های آزمون‌های عمومی با مدل سه پارامتری این احتمال تقویت می‌شود که عامل حدس آگاهانه توسط افراد در پاسخ‌گویی به سوال‌های آزمون‌های عمومی بیشتر از آزمون‌های تخصصی دخیل است. از این رو توصیه می‌شود با طراحی گزینه‌های انحرافی بهتر تاثیر این عامل در آزمون‌های عمومی کاهش یابد. یکی از دلایل دیگر در این خصوص انتشار سوال‌ها بعد از اجرا است. چون با توجه به محدودیت منابع آزمون سراسری با انتشار سوال‌ها محدودیت‌های بیشتری برای طراحان پیش می‌آید و آنها باید هر بار سوال‌های جدیدی را بسازند که احتمالا در این فرایند به دلیل انتشار زیاد سوالات در طراحی گزینه‌های دچار اشتباه می‌شوند، به طوری که گزینه‌های انحرافی کارکرد خود را از دست می‌دهند. ۲) با توجه به برازش زیاد سوال‌های آزمونهای تخصصی با مدل سه پارامتری با مجانب بالا این احتمال تقویت می‌شود که دشواری بیش از حد این سوال‌ها باعث می‌شود عملکرد افراد به خوبی اندازه‌گیری نشود. در نتیجه تفاوت‌های فردی به خوبی انعکاس نیابد. از این رو توصیه می‌شود دشواری سوال‌ها تا حدی کاهش یابد تا افراد دارای توانایی بالا کمتر در سوال‌ها دچار اشتباه شوند. ۳) در نهایت این که چون مدل‌هایی مثل ۳ پارامتری با مجانب پایین، ۳ پارامتری با مجانب بالا و ۴ پارامتری بیشترین برازش را با داده‌های آزمون‌های عمومی دارند می‌توان گفت آزمون‌های بررسی شده تا حد زیادی به خاطر تاثیر عامل حدس و بی‌دقتی توانایی افراد در صفات اندازه‌گیری شده را به خوبی نشان ندهند. محدودیت عمده این پژوهش حجم نمونه بود. حجم نمونه می‌تواند بر روی نتایج تاثیر گزار باشد. در این پژوهش امکان بررسی و پاسخ دادن به سوال‌ها بر اساس چندین حجم نمونه مختلف وجود نداشت. از این رو لازم است این عامل در تعمیم یافته‌ها در نظر گرفته شود.

References

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. Princeton, NJ: Educational Testing Service.
Baker, F. B., & Kim, S. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.
Birnbaum, A. (1968). Some latent trait models and their use in inferring.an examinee's ability. In F. M.
Cheng, Y., & Liu, C. (2016). The Effect of Upper and Lower Asymptotes of IRT Models on Computerized Adaptive Testing. Applied Psychological Measurement, 39, 551-565.
De Ayala, R. J. (2008). The theory and practice of item response theory. New York, NY: Guilfords Publications.(P 126)
Gao, S. (2011). The exploration of the relationship between guessing and latent ability in IRT models (Doctoral dissertation).
Jeffreys H. (1961). Theory of Probability. Oxford University Press.
Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. Applied Measurement in Education, 16, 277-298.
Liao, W. W., Ho, R. G., Yen, Y. C. and Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. Social Behavior & Personality: an international journal, 40(10), 1679–1694.
Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. British Journal of Mathematical and Statistical Psychology, 63(3), 509–525.
Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269.
Magis, D. (2013). A note on the item information function of the four-parameter logistic model. Applied Psychological Measurement, 37(4), 304–315.
Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1–29.
Ree, J. M. (1979). Estimating item characteristic curve. Applied Psychological Measurement, 3, 371-385.
Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?Psychological Methods, 8, 164-184.
Reynolds, T. (1986). The effects of small sample size, short test length, and ability distribution upon parameter estimation. Unpublished paper.
Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? Applied Psychological Measurement, 33, 83-101.
Sideridis GD, Tsaousis I, Al Harbi K. (2016). The Impact of Non-attempted and Dually-Attempted Items on Person Abilities Using Item Response Theory. Front Psychol; 7:1572.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the royal statistical society: Series b (statistical methodology), 64(4), 583-639.
Swist, K. (2015). Item analysis and evaluation using a four-parameter logistic model. Edukacia 3, 77–97.
Waller, N. G. and Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (ed.), Measuring psychological constructs: advances in model-based approaches (pp. 147–173).
Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. Applied Psychological Measurement, 36, 75-87.