مقایسهٔ مدل‌های اندازه‌گیری (کلاسیک و سؤال ـ پاسخ) از لحاظ برآورد پارامترهای سؤال و توانایی

معرفی مقاله

نوشتهٔ مهدی فراهانی

نظریه‌های اندازه‌گیری از یک بُعد به دو دستهٔ اساسی تقسیم می‌شود: نظریهٔ کلاسیک و نظریه‌های جدید اندازه‌گیری (IRT)[1]. این پژوهش از طریق داده‌های حاصل از اجرای سه آزمون ریاضیات، فیزیک و درس فنی که در مورد ۴۳۰ نفر از داوطلبان کنکور داخلی وزارت نیرو اجرا شده بود، به مقایسهٔ نظریات کلاسیک و جدید از لحاظ برآورد پارامترهای سؤال و توانایی پرداخته است.

تحلیل‌های کلاسیک داده‌ها از طریق برنامهٔ SPSS و برآورد پارامترهای سؤال و توانایی بر حسب مدل‌های جدید اندازه‌گیری از طریق نرم افزار رایانه‌ای BILOG در کالیفرنیای آمریکا انجام گرفته‌است. مقایسهٔ مدل‌های مختلف اندازه‌گیری از نظر برآوردهای متفاوت یا مشابه پارامترهای سؤال به وسیلهٔ تابع آگاهی و آزمون t وابسته، به آزمون و ارزیابی گذاشته شد. نتایج نشان داد که مدل کلاسیک و مدل‌های IRT برآوردهای متفاوتی برای پارامترهای سؤال به دست می‌دهند و برآورد پارامترهای سؤال بر پایهٔ مدل‌های

IRT، به‌ویژه مدل سه پارامتری، دقیق‌تر از برآورد پارامترهای سؤال بر اساس مدل کلاسیک و مدل‌های ساده‌تر IRT است. مقایسهٔ مدل‌ها از نظر برآورد متفاوت پارامتر توانایی آزمودنی‌ها به وسیلهٔ آزمون کای اسکوئر آزموده شد. نتایج این مرحله نیز تفاوت مدل‌ها را در برآورد پارامتر توانایی آشکار ساخت و نشان داد که مدل سه پارامتری نسبت به مدل کلاسیک و حتی مدل‌های یک و دو پارامتری IRT برآوردهای متفاوت‌تر و دقیق‌تری ارائه می‌دهد.

این مقاله خلاصه‌ای است از پایان‌نامهٔ تحصیلی دورهٔ کارشناسی‌ارشد آقای مهدی فراهانی، کارشناس آموزش وزارت نیرو که با راهنمایی آقای دکتر محمدکاظم سلیمی‌زاده، عضو هیأت علمی دانشگاه علامه طباطبایی تهیه شده و در اختیار فصلنامه قرار گرفته است. بدین‌وسیله از ایشان سپاس‌گزاری می‌شود.

«فصلنامه»

مقدمه

مدل یا نظریه، رکن اساسی هر رشتهٔ علمی محسوب می‌شود. در اندازه‌گیری و روان‌سنجی نیز نظریه‌ها و مدل‌هایی وجود دارد که از لحاظ تاریخی و سیر روند تکاملی، به دو دستهٔ کلاسیک (نظریهٔ ضعیف نمرهٔ حقیقی) و سؤال ـ پاسخ (به اختصار IRT) تقسیم می‌شوند (۱).

مبانی مدل کلاسیک اندازه‌گیری را در اوایل قرن حاضر، اسپیرمن معرفی و پایه‌ریزی کرد (۲). این مدل سپس در دو کتاب گالیکسن و لرد و ناویک به اوج توسعه و بسط خود رسید (۳ و ۴) اما هم‌زمان با بسط این مدل، ضعف‌های جدی آن بیش‌تر آشکار شد و روان‌سنجان و متخصصان آزمون‌سازی را بیش از پیش به سمت مدل‌های جدید سوق داد (۵). از نیمهٔ دوم قرن بیستم به‌تدریج زمینهٔ ارائهٔ نظریات جدید مطرح شد و کسانی مانند لرد (۶)، راش (به نقل از رایت، ۱۹۷۷) (۷)، برن بام (۸)، رایت و همبلتون (۹) در این مسیر گام‌های مؤثری برداشتند.

هم اینک فعالیت‌های آزمون سازی و اندازه‌گیری در زمینه‌های مختلف ـ از پیشرفت تحصیلی گرفته تا سنجش نگرش‌ها و ... ـ در آمریکا و اروپا و براساس مدل‌های جدید (IRT) بررسی و ساخته می‌شود و مدل کلاسیک اعتبار بیش‌تر برای بحث در مورد تاریخچهٔ نهضت آزمون‌سازی یا برآورد پارامترهای نظریات جدید به عنوان برآورد اوّلیه و مقدماتی مورد استفاده قرار می‌گیرد. معلمان هر چند در محدودهٔ کلاس درس به ظاهر کم‌تر می‌توانند از مدل‌های اندازه‌گیری بهره ببرند اما اطلاع از مبانی نظریِ طراحی، اجرا و تجزیه و تحلیل آزمون‌های پیشرفت تحصیلی و یافته‌های جدید، بینش و بصیرت بهتری در مورد یکی از وظایف مهم معلمان ـ یعنی سنجش محصلان ـ در اختیارآن‌ها قرار می‌دهد.

مطالعهٔ منابع مربوط به موضوع پژوهش

در بعضی از مطالعات قبلی، محققان به مقایسهٔ مدل‌ها از نظر دقت برآورد پارامترهای سؤال و توانایی پرداخته و برخی دیگر، تأثیر حجم نمونهٔ آزمودنی و سؤال یا نقض مفروضات را در برآورد پارامترها بررسی کرده‌اند. با توجه به محدودیت مقالهٔ حاضر، به‌طور عمده نتایج دستهٔ اوّل (مقایسهٔ مدل‌ها در برآورد پارامترها) ارائه می‌شود.

یکی از نتایج مطالعات مربوط به مقایسهٔ دقت برآورد پارامترها در مدل‌های اندازه‌گیری حاکی از آن‌است که مدل منطقی سه پارامتری نسبت به مدل‌های یک و دو پارامتری با آزمون‌های ۲۰ سؤالی در برآورد جایگاه افراد در صفت مکنون و رتبه‌بندی آزمودنی‌ها براساس صفت مورد سنجش دارای قدرت بیش‌تری است. نکتهٔ دیگر آن‌که

پژوهشگران با تقسیم آزمودنی‌ها به دوگروه بالا و پایین (۰ تا $\theta_L = -۲/۵$ و $\theta_u = ۰$ تا ۲/۵) و مقایسهٔ دقت برآورد جایگاه حقیقی افراد بر حسب مدل‌های IRT، نشان دادند که افزایش دقت و قابلیت پیش‌بینی توانایی افراد توسط مدل سه پارامتری نسبت به سایر مدل‌ها، در گروه‌های با توانش پایین بیش‌تر از گروه‌های با توانش بالاست؛ زیرا امکان استفاده از حدس و شانس کاذب در گروه‌های با توانایی کم بیش‌تر است و تنها مدل سه پارامتری در برآورد پارامترها به عامل C_i (مجانب پایین ICC) توجه می‌کند. مقایسهٔ مدل‌های یک و دو پارامتری (بررسی تأثیر پارامتر قدرت تشخیص سؤال در برآورد توانایی افراد) تمایز و تفاوت جدی نشان نداد (۱۰).

راید برای تحلیل داده‌های چهار خرده آزمون مورد استفاده ـ شامل محاسبات عددی، تجسم فضای سه بعدی، خزانهٔ لغات و استدلال ریاضی ـ از نرم‌افزارهای BILOG و TESTFACT بهره گرفت و بر اساس یک نمونهٔ ۴۰۶ نفره از آزمودنی‌ها و با حذف سؤال‌هایی که کم‌تر از ۷۵ درصد آزمودنی‌ها به آن‌ها پاسخ داده بودند، نتیجه گرفت که هر چهار خرده آزمون از لحاظ سطح دشواری سؤال‌ها و قدرت تشخیص افراد در سطوح مختلف توانایی، تفاوت داشته است. بنابراین، مدل‌های یک و دو پارامتری از لحاظ برآورد پارامترها تفاوت معنی‌دار نشان داده‌اند (۱۱).

هومن در تحقیقی با استفاده از آزمون تهران ـ استنفرد ـ بینه (TSB) (۱۲) به بررسی و مقایسهٔ برآورد پارامترهای دشواری و توانایی پرداخت و در واقع، توانمندی مدل راش را در برآورد پارامترها در شرایط نقض مفروضات بررسی کرد. نتیجهٔ کلی تحقیق مذکور آن است که مدل راش برای برآورد پارامتر دشواری سؤال‌ها وقتی a_i سؤال‌ها تفاوت دارد، مناسب نیست اما برای برآورد توانایی افراد مناسب و خوب است (۱۳).

در مورد مدل دو پارامتری و تفاوت برآورد توانایی بر پایهٔ آن و مدل کلاسیک، انصارین به پژوهش دست زد. او با استفاده از داده‌های حاصل از اجرای آزمون هوش تهران ـ استنفرد ـ بینه (TSB) به برآورد منحنی ویژهٔ سؤال‌ها، پارامترهای دشواری و قدرت تشخیص سؤال‌ها و توانایی آزمودنی‌ها اقدام کرد. نتیجه آن‌که نمرات خام یکسان دارای برآورد یکسانی از توانایی و موقعیت آزمودنی بر روی پیوستار مکنون نبودند (۱۴ و ۱۵).

دیوجی طی یک بررسی، کاربرد مدل یک پارامتری راش را برای سؤال‌های چند گزینه‌ای مورد بررسی قرار داده است اما به رغم استفاده از مدل راش برای برآورد پارامترها در سؤال‌های چندگزینه‌ای، به عقیدهٔ دیوجی به علت نبود پارامتر حدس و قدرت تشخیص یکسان در مدل راش، این کاربردها درست نیست (۱۶). البته علاوه بر دیوجی،

محققان دیگری هم به نامناسب بودن استفاده از مدل راش برای سؤال‌های چندگزینه‌ای اشاره کرده‌اند؛ از جمله، آندرسن عدم برازش مدل را به نابرابری قدرت تشخیص نسبت داده است. همبلتون و تراب نشان داده‌اند که مدل دو پارامتری، توزیع نمرات را بهتر از مدل راش پیش‌بینی می‌کند (۹).

همبلتون و موری از طریق نمودار باقی‌مانده برای آزمون‌های ریاضی، برازش مدل سه پارامتری و عدم برازش مدل تک پارامتری را برای داده‌ها مطرح کرده‌اند (۱۶).

یکی از موضوع‌ها و نکات مهم در مورد مدل‌های IRT، حجم نمونهٔ آزمودنی و سؤال است. این دو عامل، به‌ویژه در مدل سه پارامتری، می‌تواند بر برآورد پارامترهای سؤال و توانایی تأثیرات جدی داشته باشد. همبلتون و کوک در یک مطالعه با انتخاب آزمون‌هایی با سه طول ۱۰، ۲۰ و ۸۰ سؤالی و نمونه‌هایی با حجم ۵۰، ۲۰۰ و ۱۰۰۰ آزمودنی به بررسی اثرات حجم نمونهٔ آزمودنی و ویژگی‌های خزانهٔ سؤال و تعداد سؤال بر خطای استاندارد برآورد توانایی اقدام کردند. طول آزمون ۱۰ سؤالی حداقل طول ممکن برای یک آزمون و آزمون ۸۰ سؤالی نیز از طول‌های متداول تست است. در مورد حجم نمونهٔ آزمودنی ۵۰ و ۱۰۰۰ نفر نیز به همین‌گونه استدلال کرده‌اند. سؤال‌ها از دو خزانهٔ سؤال استخراج شدند. در عمل، سؤال‌های خزانهٔ ۱ دارای دامنهٔ عریض‌تری برای پارامترهای دشواری و قدرت تشخیص سؤال بودند. پارامتر حدس هر دو خزانهٔ سؤال ۰/۲۵ در نظر گرفته شد. جمع‌بندی نتایج این مطالعه به صورت زیر است:

۱. حجم‌نمونهٔ پاسخ‌گویان و طول آزمون، دوعامل بسیارمهم در دقت منحنی‌های SE (θ) است. موارد نقض و استثنای این امر به نوسانات نمونه گیری مربوط می‌شود.

۲. در کرانه‌های پیوستار توانایی، دقت منحنی‌های SE (θ) حتی با وجود نمونه‌های بزرگ آزمودنی، بسیار پایین است.

۳. در اکثر موارد با نمونه‌های ۲۰۰ آزمودنی و ۲۰ سؤال، دقت برآورد خطای استاندارد توانایی قابل قبول خواهد بود. البته این نکته بیش‌تر در دامنهٔ وسط توانایی [۱+ ، ۱-] صادق است.

۴. افزایش طول تست از ۱۰ به ۲۰ سؤال بیش از افزایش آن از ۲۰ به ۸۰ سؤال، دقت SE را بهبود می‌بخشد.

۵. در مورد حجم نمونه نیز افزایش افراد از ۵۰ به ۲۰۰، بیش از ۲۰۰ به ۱۰۰۰ نفر دقت برآورد SE را ارتقا می‌دهد (۱۷).

لرد طی یک مطالعه و ضمن مقایسهٔ مدل‌های یک و دو پارامتری IRT در برآورد نمرهٔ حقیقی آزمودنی‌ها، تلاش کرده است تأثیر حجم نمونه را بررسی کند. داده‌های مطالعه

شامل پاسخ ۳۰۰۰ دانش‌آموز کلاس ششم به آزمون خزانهٔ لغات متروپولیتن با برنامه LOGIST تجزیه و تحلیل شده است. نتایج مطالعه نشان داد وقتی حجم نمونه کوچک باشد، پارامتر قدرت تشخیص (a_i) سؤال‌ها و پارامتر مجانب یا حدس سؤال‌ها (c_i) را نمی‌توان به‌دقت تعیین کرد. از این رو، در بعضی موقعیت‌های معین و محدود و با حجم نمونهٔ کوچک‌تر از ۱۰۰ یا ۲۰۰ آزمودنی، برآوردکنندهٔ نمره حقیقی X آزمودنی در مدل راش (یک پارامتری) می‌تواند اندکی بهتر از برآوردکنندهٔ نمرهٔ حقیقی بر پایهٔ مدل دو پارامتری باشد (٦).

بیان اهداف و فرضیه‌های پژوهش

هدف اصلی پژوهش حاضر مقایسهٔ مدل‌های اندازه‌گیری (کلاسیک و سؤال ـ پاسخ) از لحاظ برآوردهای متفاوت یا مشابهی است که برای پارامترهای سؤال‌های آزمون و توانایی آزمودنی‌ها به دست می‌دهند.

فرضیه‌های این پژوهش که آزمون آن‌ها مورد توجه است، عبارت اند از:

۱. کاربرد مدل‌های سؤال ـ پاسخ (با تعداد پارامتر مناسب) برای برآورد دقیق مشخصات سؤال بر مدل کلاسیک برتری دارد.

۲. برآورد توانایی آزمودنی‌ها با استفاده از مدل‌های IRT (با تعداد پارامتر مناسب) از برآورد توانایی افراد بر پایهٔ مدل کلاسیک دقیق‌تر است.

۳. در صورت وجود عامل حدس در پاسخ دادن به سؤال‌ها، افزودن پارامتر حدس (C_i) به مدل IRT، مدل برازنده‌تری برای داده‌ها ایجاد می‌کند.

روش اجرای پژوهش

الف. آزمودنی‌ها

جامعهٔ این پژوهش را همهٔ داوطلبان آزمون ورودی دوره‌های داخلی وزارت نیرو تشکیل می‌دهد و نمونهٔ تحقیق تعداد ۵۵۳ نفر از داوطلبانی هستند که در آزمون داخلی گزینش دانشجو (مورخ ۱۳۷۳/۳/٦) در رشتهٔ قدرت (مقطع کاردانی) شرکت کرده‌اند. برای نمونه‌برداری همهٔ شرکت‌کنندگان انتخاب شدند. در واقع، شرکت‌کنندگان در آزمون مذکور نمونه‌ای از همهٔ داوطلبان فرض شده‌اند. حجم نمونهٔ اوّلیه ۵۵۳ نفر بود اما تعدادی از پاسخ‌نامه‌ها به علت مخدوش و غیرقابل استفاده بودن از نمونه حذف شد و حجم نمونهٔ نهایی به ۴۳۰ نفر کاهش یافت.

ب. روش‌های آماری

برای تجزیه و تحلیل داده‌ها ابتدا در ایران مشخصه‌های کلاسیک سؤال‌ها و آزمون‌ها محاسبه شد. سپس داده‌ها که شامل پاسخ‌های ۴۳۰ آزمودنی به سؤال‌های چهارگزینه‌ای سه خرده آزمون ۲۰ سؤالی بود، از طریق نظام شبکهٔ جهانی اطلاعات (Internet) به یکی از دانشگاه‌های آمریکا (دانشگاه UCLA) ارسال گردید. پس از تحلیل داده‌ها با نرم‌افزار بای‌لوگ، خروجی کامپیوتر شامل برآورد پارامترهای سؤال و توانایی افراد، به ایران فرستاده شد.[۱۳] خلاصهٔ اطلاعات در مورد خرده آزمون ریاضیات به‌همراه نمودارهای دشواری و قدرت تشخیص سؤال‌ها در پایان مقاله ارائه شده است (نمودار شمارهٔ ۱و۲ و جدول شمارهٔ ۴).

برای تعیین برتری مدل‌های سؤال ـ پاسخ بر مدل کلاسیک در برآورد دقیق مشخصات سؤال (فرضیهٔ ۱ پژوهش) از یک ابزار نیرومند نظریات جدید اندازه‌گیری به نام تابع آگاهی[۱۴] استفاده شد. تابع آگاهی نظامی است که ورودی آن پارامترهای سؤال و خروجی آن میزان آگاهی‌دهندگی آزمون می‌باشد. متفاوت بودن تابع آگاهی یک آزمون برحسب مدل‌های اندازه‌گیری، بیانگر تفاوت مدل‌ها از برآوردهایی است که برای پارامترهای سؤال‌های آزمون مذکور محاسبه شده است. برای بررسی معناداری تفاوت تابع آگاهی آزمون‌ها به علت وابسته بودن داده‌ها، از آزمون t برای داده‌های وابسته استفاده شد و تابع آگاهی هر آزمون که براساس مدل‌های یک و دو پارامتری و کلاسیک برآورد شده بود، با تابع آگاهی همان آزمون در مدل سه پارامتری مقایسه گردید. معناداری تفاوت این توابع آگاهی با t آزمون شد.

به منظور آزمون فرضیهٔ دوم ـ تفاوت یا عدم تفاوت مدل‌ها در برآورد توانایی و سطح صفت مکنون آزمودنی‌ها ـ از آزمون مجذور کای از نوع نیکویی برازش استفاده شد تا مدل‌های یک و دو پارامتری و کلاسیک را در مقایسه با مدل سه پارامتری آزمون کند. برای این منظور، ابتدا نمرات به مقیاس استاندارد یا Z برده شدند و سپس به ۱۲ طبقه از [۲/۵ـ ، ۳ـ] تا [۳، ۲/۵] تقسیم گردیدند. آن‌گاه فراوانی طبقات براساس آزمون نیکویی برازش با درجات آزادی $d_f = k - 1$ مقایسه و در دو سطح معناداری ۰/۰۵ و ۰/۰۱ بررسی شد.

برای بررسی و تعیین برازش یا عدم برازش هر یک از مدل‌های IRT با داده‌ها، یعنی به منظور آزمون فرضیهٔ ۳ پژوهش مبنی بر برازندگی بیش‌تر مدل سه پارامتری نسبت به سایر مدل‌های IRT، نرم‌افزار BILOG از نوعی آزمون مجذور کای استفاده می‌کند. فرض صفر آزمون نیکویی برازش مدل ـ داده‌ها بر تناسب و برازش مدل با داده‌ها تأکید می‌ورزد و در

واقع، تفاوت مدل با داده‌ها را انکار می‌کند و فرض خلاف عدم برازش را بیان می‌دارد. در عمل، سطح احتمال مجذورکای محاسبه شدهٔ هر سؤال با ۰/۰۱ و ۰/۰۵ مقایسه می‌شود. در صورت بزرگ‌تر بودن سطح احتمال هر سؤال از ۰/۰۱ یا ۰/۰۵، نتیجه گرفته می‌شود که سؤال با مدل دارای برازش نسبی ($۰/۰۵ > \alpha > ۰/۰۱$) یا برازش کامل ($\alpha > ۰/۰۵$) است.

پ. ابزار گردآوری داده‌ها

ابزار مورد استفاده برای گردآوری داده‌های پژوهش شامل سه آزمون بیست سؤالی ریاضیات، فیزیک و درس فنی به صورت چهارگزینه‌ای بوده است. درس‌های مذکور جزء مواد امتحانی اصلی و مشترک آزمون‌های گزینش دانشجو در وزارت نیرو به شمار می‌روند. سایر مواد امتحانی تعداد آزمودنی کمی دارد. این آزمون‌ها به وسیلهٔ متخصصان حیطه‌های مزبور که معمولاً جزء کارشناسان وزارت نیرو هستند، طراحی و تهیه می‌شود. در مورد نحوهٔ اجرای ابزار تا حد قابل قبولی می‌توان شرایط تقریباً استاندارد و اصولی را برای آن‌ها در نظر گرفت. از جمله در ابتدای دفترچهٔ آزمون دربارهٔ شیوهٔ پاسخ‌گویی به سؤال‌ها، مدت زمان اجرا، وجود نمرهٔ منفی در آزمون، تعداد سؤال‌های هر خرده آزمون و ... توضیحاتی ذکر شده است.

یافته‌های پژوهش

برای آزمون فرضیهٔ ۱، مقایسهٔ مدل سه پارامتری و مدل کلاسیک از نظر برآوردهایی که برای پارامترهای سؤال‌ها به دست می‌دهند، از تابع آگاهی خرده آزمون‌ها بهره گرفته‌شد. تفاوت توابع آگاهی هر آزمون که حاصل مدل‌های مختلف بود، محاسبه و با آزمون t برای داده‌های وابسته آزمون شد. نتایج نشان داد که نه تنها بین تابع آگاهی هر آزمون براساس مدل کلاسیک تفاوت قابل توجهی دیده می‌شود و این تفاوت حتی در سطح $\alpha < ۰/۰۱$ و با بیش از ۹۹ درصد اطمینان معنادار است بلکه توابع آگاهی مدل‌های دو پارامتری و یک پارامتری نیز برای آزمون از تابع آگاهی هر خرده آزمون براساس مدل کلاسیک در همین سطح تفاوت معناداری دارد. در واقع، با درجهٔ آزادی ۳۲ و مقدار t مبین ۲/۴۵ تنها توابع آگاهی آزمون‌ها بین مدل‌های یک و دو پارامتری معنادار نبود و سایر مقادیر همه تفاوت معناداری داشتند. بنابراین، می‌توان گفت علاوه بر مدل سه پارامتری IRT، حتی مدل‌های دو و یک پارامتری نیز پارامترهای سؤال‌ها را دقیق‌تر و مناسب‌تر از مدل کلاسیک برآورد می‌کنند و فرضیهٔ صفر رد و فرضیهٔ اوّل پژوهش با بیش از ۹۹ درصد اطمینان تأیید می‌شود (جدول شمارهٔ ۱).

جدول شمارهٔ ۱ ـ محاسبهٔ آزمون t گروه‌های وابسته برای تفاوت بین توابع آگاهی آزمون‌ها برحسب مدل‌های مختلف

تابع آگاهی	آزمون ریاضیات			آزمون فیزیک			آزمون درس فنی		
DIF(I-F)	MEAN	STD	T(Ob)	MEAN	STD	T(Ob)	MEAN	STD	T(Ob)
DF (۱, ۳)	۱/۶۶	۳/۶۲	۲/۶۲	۱/۸۱	۳/۶۹	۲/۸۲	۱/۵۵	۲/۹۹	۲/۹۷
DF (۲, ۳)	۱/۶۴	۳/۷۴	۲/۵۲	۱/۵۷	۳/۲۴	۲/۷۹	۱/۱۷	۲/۳۷	۲/۸۳
DF (۱, ۲)	۰/۰۲	۰/۵۷	۰/۱۵	۰/۲۴	۰/۹۴	۱/۴۸	۰/۳۸	۱/۱۹	۱/۸۳
DF (cl, ۳)	۲/۹۲	۴/۷۱	۳/۵۶	۲/۸۱	۴/۵۹	۳/۵۲	۲/۵۸	۳/۷۹	۳/۹۱
DF (cl, ۲)	۱/۲۸	۱/۶۳	۴/۵۰	۱/۲۴	۱/۸۷	۳/۸۲	۱/۴۱	۱/۹۲	۴/۲۳
DF (cl, ۱)	۱/۲۶	۱/۳۳	۵/۴۴	۱/۰۰	۱/۰۵	۵/۴۵	۱/۰۳	۰/۹۷	۶/۱۴

$n = ۳۳$ $df = ۳۲$ $t(cr)\ ۰/۰۵ = ۱/۶۹$ $t(cr)\ ۰/۰۱ = ۲/۴۵$

فرضیهٔ ۲ مدعی است که مدل سه پارامتری نظریهٔ سؤال ـ پاسخ توانایی افراد را نسبت به مدل کلاسیک اندازه‌گیری به‌گونه‌ای متفاوت و دقیق‌تر برآورد می‌کند. به این ترتیب، باید برای آزمون این فرضیه، تفاوت مدل سه پارامتری از مدل کلاسیک در برآورد توانایی بررسی شود. به این منظور، آزمون مجذور کای بین مدل سه پارامتری و مدل کلاسیک (البته مدل‌های دو و یک پارامتری نیز) اجرا شد تا تفاوت فراوانی‌های افراد در طبقات مختلف آزمون شود. نتایج نشان داد که نه تنها توانایی برآورد شده برای افراد براساس مدل سه پارامتری از مدل کلاسیک متفاوت است بلکه حتی مدل‌های دو و یک پارامتری نیز در برآورد پارامتر توانایی آزمودنی‌ها با مدل سه پارامتری تفاوت و تمایز دارد. این امر در سطح $\alpha < ۰/۰۱$ هم معنی‌دار بود. از سوی دیگر، از آن‌جاکه مدل سه پارامتری از لحاظ تعداد پارامتر مناسب‌ترین مدل برای داده‌های حاضر محسوب می‌شود و طبق نتایج آزمون خوبی برازندگی، برازش و مناسبت کاملاً آشکاری با داده‌های پژوهش دارد، توانایی برآورد شده برای افراد راکه از مدل سه پارامتری به دست آمده است، می‌توان دقیق‌تر و مناسب‌تر از توانایی حاصل از سایر مدل‌های اندازه‌گیری تلقی کرد. بنابراین، فرضیهٔ صفر رد و فرضیهٔ دوم پژوهش حاضر نیز با بیش از ۹۹ درصد اطمینان تأیید می‌شود (جدول شمارهٔ ۲).

جدول شمارهٔ ۲ـ برازندگی مدل ـ داده‌ها برحسب مدل‌های IRT (تعداد سؤال‌ها)

مدل‌ها / آزمون‌ها	۱ ـ PAR			۲ ـ PAR			۳ ـ PAR		
	بدون برازش $\alpha<$/۰۱	برازش نسبی /۰۱<α<۰۵/	برازش کامل $\alpha>$/۰۵	بدون برازش $\alpha<$/۰۱	برازش نسبی /۰۱<α<۰۵/	برازش کامل $\alpha>$/۰۵	بدون برازش $\alpha<$/۰۱	برازش نسبی /۰۱<α<۰۵/	برازش کامل $\alpha>$/۰۵
آزمون ریاضیات	۱٦	۳	۱	۱۸	۱	۱	۱	۲	۱۷
آزمون فیزیک	۱۹	۱	۰	۱۷	۱	۲	۲	۲	۱٦
آزمون درس فنی	۱۸	۲	۰	۱٦	۱	۳	۱	۲	۱۷

۲ ≤ df ≥ ۸ ۲ ≤ df ≥ ۹ ۲ ≤ df ≥ ۹

جدول شمارهٔ ۳ـ مجذورکای محاسبه شده بین پارامتر توانایی مدل سه پارامتری و مدل‌های دو و یک پارامتری و کلاسیک

آزمون‌ها	ریاضی			فیزیک			درس فنی		
مدل‌ها	دو پارامتری	یک‌پارامتری	کلاسیک	دو پارامتری	یک‌پارامتری	کلاسیک	دو پارامتری	یک‌پارامتری	کلاسیک
سه پارامتری	۲۰۰۹	۵۰/۶	۲۱۹/۶	۲۲/۹	۳۱/۸	۸۲	۲۹/۵	۱۲/۴	۸۶/۱

df = ۱۱ x^2(cr) ۰/۰۵ = ۱۹/٦۷۵ x^2(cr) ۰/۰۱ = ۲۴/۷۲۵

برای آزمون فرضیهٔ ۳ پژوهش از نوعی آزمون خوبی برازندگی استفاده شد و برازش سؤال‌های آزمون با مدل‌های یک، دو و سه پارامتری بررسی گردید. براساس نتایج به دست آمده، مدل سه پارامتری برازش چشم‌گیری با داده‌ها داشت؛ در حالی‌که نه تنها مدل دو پارامتری بلکه مدل یک پارامتری نیز بدون برازش شناخته شد. بنابراین، فرضیهٔ سوم پژوهش نیز با بیش از ۹۹ درصد اطمینان و در سطح $\alpha < ۰/۰۱$ معنادار شناخته شد و فرض صفر رد و فرض پژوهش (خلاف) تأیید گردید. بدین‌ترتیب مدل سه پارامتری، برای داده‌های آزمون‌های مورد استفاده در این پژوهش نسبت به سایر مدل‌های IRT برازنده‌تر و مناسب‌تر شناخته شد (جدول شمارهٔ ۳).

نتیجه‌گیری

مطابق نتایج تحقیقات و مطالعات قبلی در زمینهٔ نظریه‌های جدید اندازه گیری، IRT و مقایسهٔ آن با نظریهٔ کلاسیک آزمون، این مطالعه نیز نشان داد که در سطح بالایی از اطمینان می‌توان گفت:

۱. مدل‌های IRT نسبت به مدل کلاسیک اندازه گیری در برآورد پارامترهای سؤال‌ها و توانایی آزمودنی‌ها دارای مناسبت، دقت و برازندگی بیش‌تری است. این نتیجه در مورد آزمون‌های پیشرفت تحصیلی چهارگزینه‌ای که مبنای مطالعهٔ حاضر بوده است، با بیش از ۹۹ درصد اطمینان صدق می‌کند.

۲. در بین مدل‌های نظریهٔ سؤال ـ پاسخ (IRT) برای برآورد پارامترهای سؤال و توانایی افراد، مدل سه پارامتری در مورد آزمون‌های پیشرفت تحصیلی چندگزینه‌ای این پژوهش، نسبت به مدل‌های یک و دو پارامتری برتری خاصی نشان داد.

۳. در بین مدل‌های یک و دو پارامتری IRT از لحاظ برآورد پارامترهای سؤال براساس داده‌های خرده آزمون‌های حاضر، تفاوت چشم‌گیری ملاحظه نمی‌شود. از این رو می‌توان تأثیرگذاری پارامتر قدرت تشخیص سؤال‌ها را بر برآورد پارامتر دشواری چندان شدید و جدی ندانست. به زبان دیگر، سؤال‌های مورد استفاده در خرده آزمون‌های این مطالعه دارای قدرت تشخیص خیلی متفاوت و مؤثری نیستند اما در مورد برآورد پارامتر توانایی آزمودنی‌ها، می‌توان نتیجه گرفت که براساس داده‌های حاضر، بین مدل‌های یک و دو پارامتری در مقام مقایسه با مدل سه پارامتری، برآورد پارامتر توانایی تا حدی متفاوت است و در مورد آزمون‌های چندگزینه‌ای پارامتر شیب بر روی برآورد پارامتر توانایی تأثیر داشته است.

پی‌نوشت‌ها

۱. نظریات جدید اندازه‌گیری و روان‌سنجی نخست با اصطلاح صفت مکنون یا خصیصهٔ مکنون به‌شدت پیوند خورد اما به‌تدریج با عنوان نظریهٔ سؤال ـ پاسخ یا Item Response Theory (با علامت اختصاری IRT) یا نظریهٔ منحنی ویژهٔ سؤال رواج یافت که برای مقاصد آزمون‌سازی و تحلیل آماری داده‌ها مناسب‌تر به نظر می‌رسد (همبلتون، ۱۹۹۳).

2. Information Function

۳. برای تحلیل داده‌ها در آمریکا به وسیلهٔ نرم‌افزار BILOG، علاوه بر مساعی فراوان استاد محترم راهنما آقای دکترسلیمی‌زاده، نگارنده بر خود لازم می‌داند از عنایت استاد ارجمند دکترجمال عابدی، عضو هیأت علمی دانشگاه مذکور که تحلیل داده‌ها را مسیر کردند، تشکر و سپاسگزاری نماید.

منابع

1. Allen, J.M., & Yen, M.W. (1979). **Introduction to measurement thoery.** California: wadsworth.

۲. ثرندایک، رابرت؛ **روان‌سنجی کاربردی**، مترجم: حیدرعلی هومن، تهران، دانشگاه تهران، ۱۳٦۹.

3. Gulliksen, H. (1950). **Theory of mental tests.** New york: John wiley & sons.

4. Lord, F.M., & Novick, M.R. (1968). **Statistical Theories of mental test scores.** Reading, MA: Addison-wesley.

5. Hambleton, R.K., & van der linden, W.J. (1982). Advances in item response theory and applications: An introduction. **Applied Psychological Measurement,** 6. 373-378.

6. Lord, F.M. (1983). Small N justifies Rasch model. In D.J. weiss (ed.). **New Horizons in testing.** New York: Academic press Inc.

7. Wright, B.D. (1977a). Misunderstanding of the Rasch model. **Journal of Educational Measurement,** 14,219-226.

8. Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee's

ability. In F.M. Lord & M.R. Novick. **Statistical theories of mental test scores, MA:** Addison - wesley.

9. Hambleton, R.K. (1993). Principles and selected applications of item response theory. In R.L. Linn (Ed.), **Educational Measurement.** U.S.A.: oryx press.

10. Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. **Journal of Educational Measurement,** 14, 75-96.

11. Reid, C.A. (1993). **Latent trait modeling of the general test battery used with a rehabilitation client population: An investigation of Model - Data Fit.** Source: DAI-B 54/12. p. 6497. Jun 1994.

۱۲. هومن، حیدرعلی؛ **مقایسهٔ مدل تک‌پارامتری راش و مدل دوپارامتری**، پایان‌نامهٔ منتشر نشدهٔ دانشگاه آزاد اسلامی، ۱۳۷۳.

۱۳. هومن، حیدرعلی؛ **هوش آزمای انفرادی تهران ـ استنفرد ـ بینه. فصلنامهٔ** علوم‌تربیتی دانشگاه تهران، ویژه‌نامهٔ روان‌سنجی، دورهٔ جدید، سال یکم، شمارهٔ ۱ـ۴.

۱۴. انصارین، علیرضا؛ **برآورد خم ویژهٔ سؤال و توانایی آزمودنی‌ها در مقیاس تهران ـ استنفرد ـ بینه بر پایهٔ مدل دوپارامتری صفت مکنون**، پایان‌نامهٔ منتشر نشدهٔ دانشگاه آزاد اسلامی، ۱۳۷۱.

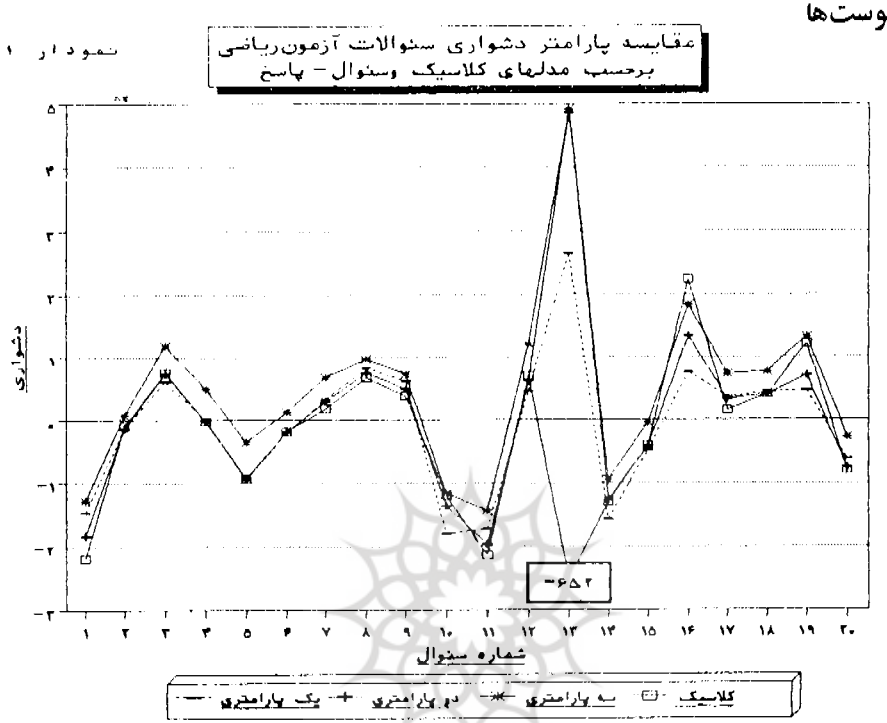
۱۵. هومن، حیدرعلی؛ **روش تهیهٔ آزمون هوش**، تهران، دانشگاه تهران، ۱۳۷۵.

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

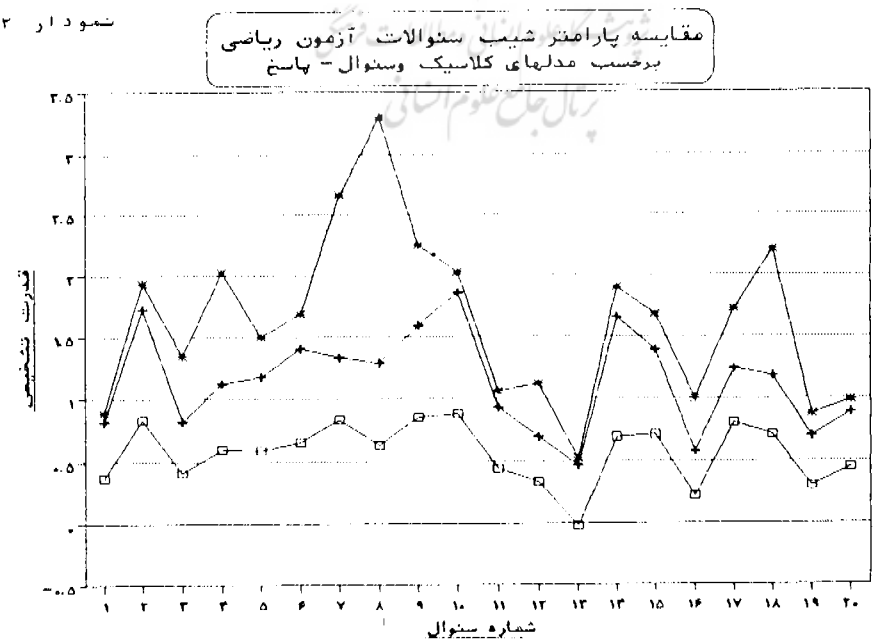
17. Hambleton, R.K., & Cook. L.L. (1983). The robustness of item response models and effects of test length and sample size on the precision of ability estimates. **In D.J.** Weiss (Ed.), **New horizons in testing** (pp. 31-49). New York: Academic Press.

پیوست‌ها

نمودار ۱



نمودار ۲



جدول شمارهٔ ۴ ـ آماره‌ها و پارامترهای سؤال‌های خرده آزمون ریاضی برحسب مدل‌های کلاسیک و سؤال ـ پاسخ (IRT)

پارامتر حدس Ci	پارامتر قدرت تشخیص Ai			پارامتر دشواری Bi			آماره‌های سؤال‌ها (مشخصه‌ها)					آزمون ریاضیات
۳ PAR	کلاسیک	۲ PAR	۳ PAR	کلاسیک	۱ PAR	۲ PAR	۳ PAR	دورشته‌ای نقطه‌ای	همبستگی دورشته‌ای	درصد صحیح	تعداد صحیح	تعداد کوشش	سؤال
۰/۲۵۱	۰/۳۷	۰/۸۲۴	۰/۸۹۵	-۲/۲۰۴	-۱/۴۷۴	-۱/۸۴۲	-۱/۲۲۸	۰/۲۴۳	۰/۳۴۷	۸۰/۱۰	۳۲۲	۴۰۲	۱
۰/۱۲	۰/۸۳۳	۱/۷۲۳	۱/۹۳۳	-۰/۰۹۹	-۰/۱۵۱	-۰/۱۴۵	۰/۰۶۸	۰/۵۱	۰/۶۴	۵۳۰۴۸	۱۹۲	۳۵۹	۲
۰/۱۸۳	۰/۴۱	۰/۸۲۳	۱/۳۴۲	۰/۷۴۵	۰/۵۹۴	۰/۷۱۷	۱/۱۷۹	۰/۲۹۷	۰/۳۷۹	۳۷۰۲۸	۱۰۴	۲۷۹	۳
۰/۲۲۶	۰/۵۹۶	۱/۱۳	۲/۰۲۷	-۰/۰۱۶	-۰/۰۱۱	-۰/۰۲۶	۰/۴۹۷	۰/۴۰۸	۰/۵۱۲	۵۰/۳۶	۱۴۰	۲۷۸	۴
۰/۲۷	۰/۵۹۱	۱/۱۷۴	۱/۴۹۷	-۰/۹۴	-۰/۹۵۹	-۰/۹۳۶	-۰/۳۶	۰/۳۸۲	۰/۵۰۹	۷۱/۹۶	۲۳۱	۳۲۱	۵
۰/۱۶	۰/۶۵۲	۱/۳۹۸	۱/۶۸۳	-۰/۱۹۲	-۰/۲۰۱	-۰/۱۹۱	۰/۱۳	۰/۴۳۴	۰/۵۴۶	۵۵/۲۶	۱۸۹	۳۴۲	۶
۰/۱۸۱	۰/۸۳۵	۱/۳۳۱	۲/۶۶۵	۰/۱۷۳	۰/۳۲۳	۰/۲۸	۰/۶۷۳	۰/۵۱	۰/۶۴۱	۴۳/۹۴	۸۷	۱۹۸	۷
۰/۱۵۵	۰/۶۳۲	۱/۲۸۵	۳/۳۰۳	۰/۶۵۵	۰/۸۲۱	۰/۷۴۹	۰/۹۶۷	۰/۴۱۱	۰/۵۳۴	۳۳/۰۶	۱۱۹	۳۶۰	۸
۰/۱۱۸	۰/۸۵۸	۱/۵۸۳	۲/۲۴۶	۰/۳۷۱	۰/۶۱۳	۰/۴۸۵	۰/۷۱۵	۰/۵۰۹	۰/۶۵۱	۳۶/۸۱	۱۱۳	۳۰۷	۹
۰/۲۰۳	۰/۸۷۶	۱/۸۶۱	۲/۰۳۱	-۱/۱۹۷	-۱/۸۱۴	-۱/۳۷	-۱/۱۷۳	۰/۴۲۵	۰/۶۵۹	۸۵/۶۴	۳۱۰	۳۶۲	۱۰
۰/۲۵۵	۰/۴۴	۰/۹۳۱	۱/۰۶	-۲/۱۴۸	-۱/۷۳۷	-۱/۹۸۳	-۱/۴۴۴	۰/۲۷	۰/۴۰۳	۸۳/۳۳	۳۳۰	-۳۹۶	۱۱
۰/۲۰۵	۰/۳۳۲	۰/۶۸۹	۱/۱۱۸	۰/۷۱۱	۰/۴۵۱	۰/۶۱۹	۱/۲۰۷	۰/۲۴۸	۰/۳۱۵	۴۰/۰۸	۱۰۳	۲۵۷	۱۲
۰/۰۳۷	-۰/۰۲۶	۰/۴۶۵	۰/۵۳۳	-۶۵/۲۲۶	۲/۶۴۷	۴/۸۵۸	۴/۸۹۴	-۰/۰۱۳	-۰/۰۲۶	۵/۳۴	۱۴	۲۶۲	۱۳
۰/۲۲۷	۰/۶۹۴	۱/۶۵۲	۱/۸۹۵	-۱/۳۰۸	-۱/۵۸۲	-۱/۲۶۴	-۰/۹۶۶	۰/۳۸۷	۰/۵۷	۸۲/۳۸	۳۰۴	۳۶۹	۱۴
۰/۲۰۱	۰/۷۱۸	۱/۳۹۳	۱/۶۸	-۰/۴۱۸	-۰/۴۸۸	-۰/۴۳۷	-۰/۰۵۲	۰/۴۵۶	۰/۵۸۳	۶۲/۵۴	۲۱۷	۳۴۷	۱۵
۰/۱۸۳	۰/۲۱۵	۰/۵۷۸	۱	۲/۲۴	۰/۷۵۴	۱/۳۲۱	۱/۸۲۷	۰/۱۶	۰/۲۱	۳۰/۶۴	۹۱	۲۹۷	۱۶
۰/۱۶۳	۰/۸۰۷	۱/۲۴	۱/۷۲۴	۰/۱۴۹	۰/۳۵۳	۰/۳۲۵	۰/۷۲۱	۰/۱۵	۰/۶۲۸	۴۴/۹۳	۹۳	۲۰۷	۱۷
۰/۱۷۷	۰/۷۱۸	۱/۱۷۸	۲/۲۰۳	۰/۳۹۴	۰/۴۳۷	۰/۴۰۲	۰/۷۶۱	۰/۴۵۷	۰/۵۸۳	۳۸/۲۱	۹۴	۲۴۶	۱۸
۰/۱۶۹	۰/۳۰۶	۰/۷۰۳	۰/۸۸۱	۱/۲۰۷	۰/۴۵۹	۰/۷۰۸	۱/۳۰۹	۰/۲۲۷	۰/۲۹۳	۳۴/۸۳	۶۲	۱۷۸	۱۹
۰/۲۰۱	۰/۴۵۳	۰/۸۹۵	۰/۹۸۴	-۰/۸۱۵	-۰/۶۳۴	-۰/۷۶	-۰/۲۸۳	۰/۳۲۱	۰/۴۱۳	۶۵/۲۴	۲۱۴	۳۲۸	۲۰
۰/۱۸۴	۰/۵۶۶	۱/۱۴۳	۱/۶۳۵	-۳/۴۰	-۰/۱۰۸۰	۰/۰۷۵	۰/۴۶۹						MEAN =
۰/۱۰۵۳	۰/۲۴۶	۰/۳۹۷	۰/۶۸۴	۱۴/۵۹	۱/۰۸۳	۱/۴۵۶	۱/۳۹۵						STD =