نظریه‌های قدیم و جدید اندازه‌گیری در علوم رفتاری و پزشکی: مروری بر روش‌شناسی، مزایا و تنگناها

مجتبی حبیبی[۱]، ابراهیم خدایی[۲]، بلال ایزانلو[۳]

چکیده

نظریه‌های کلاسیک آزمون و سؤال پاسخ در ساخت و تحلیل آزمون‌ها و پرسش‌نامه‌ها کاربرد زیادی دارند. نظریه سؤال پاسخ به دلیل وجود نقطه ضعف‌های متعدد نظریه کلاسیک پا به عرصه وجود گذاشت. اگر چه این نظریه نقطه ضعف‌های اصلی نظریه کلاسیک را جبران نمود و انجام کارهایی مثل سنجش انطباقی و تشکیل بانک سؤال در بطن آن را امکان‌پذیر کرد، ولی دلایل قطعی برای حذف کامل نظریه کلاسیک وجود ندارد و این نظریه هنوز در حوزه‌های متعدد علمی مورد استفاده قرار می‌گیرد. مطالعه پیشینه نشان می‌دهد که نظریه سؤال پاسخ توسعه نظریه کلاسیک در غالب مدل‌های پیچیده‌تر و کامل‌تر است نه رقیب آن. تفاوت اصلی این دو نظریه در این است که نظریه سؤال پاسخ، سؤال محور، ولی نظریه کلاسیک، آزمون محور است. این موضوع پیامدهای مختلفی در پی داشته است. به عنوان مثال، مدل‌سازی خطای اندازه‌گیری در نظریه سؤال پاسخ در سطح سؤال و در نظریه کلاسیک در سطح آزمون است. در عمل، نظریه سؤال پاسخ در مواجهه با مشکلات پیش روی نظریه کلاسیک مثل خطاهای استاندارد شرطی، طراحی آزمون‌های موازی، هم‌ترازسازی، بررسی سوگیری سؤال و سنجش انطباقی، عملکرد موفقی داشته است. در مقاله حاضر اصول، مبانی، مزایا و معایب هر دو نظریه مورد مقایسه قرار گرفته‌اند.

واژه‌های کلیدی: نظریه کلاسیک، نظریه سؤال پاسخ، مقایسه، مزایا و معایب

نوع مقاله: پژوهشی

دریافت مقاله: ۹۱/۲/۲۰ پذیرش مقاله: ۹۱/۸/۲۰

مقدمه

نظریه کلاسیک آزمون (Classical test theory) و سؤال پاسخ (Item response theory) در اندازه‌گیری از کاربرد و اهمیت زیادی برخوردارند (۱). نظریه کلاسیک آزمون به عنوان اولین نظریه منسجم در خصوص اندازه‌گیری‌های ذهنی در اوایل قرن ۲۰ شکل گرفت (۲). اجزای سازنده این نظریه از سه پیشرفت قابل توجه در ۱۵۰ سال قبل از شکل‌گیری آن تشکیل شده بود و بنیان نظریه کلاسیک به این ترتیب ایجاد شد. این سه پیشرفت مهم عبارتند از: ۱- شناسایی وجود خطا در اندازه‌گیری‌ها، ۲- در نظر گرفتن خطاها به عنوان یک متغیر تصادفی و ۳- مفهوم همبستگی و چگونگی شاخص‌گذاری (Index) آن. به دنبال این سه پیشرفت، Spearman (۳) نشان داد که چه طور کاهش در ضریب همبستگی -که به خاطر خطای اندازه‌گیری روی می‌دهد- را تصحیح کنیم و بر اساس این تصحیح شاخص اعتبار را به دست آوریم. به این ترتیب، توضیحات و

۱- استادیار، روان‌شناسی سلامت، پژوهشکده خانواده دانشگاه شهید بهشتی، تهران، ایران

۲- دانشیار، سازمان سنجش آموزش کشور، تهران، ایران

۳- دانشجوی دکتری، گروه سنجش آموزش، دانشگاه تهران، بورسیه دانشکده روان‌شناسی، دانشگاه خوارزمی، تهران، ایران (نویسنده مسؤول)

Email: b.ezanloo@gmail.com

اثبات‌های Spearman آغاز نظریه کلاسیک آزمون را مشخص کرد (۴).

از آن جا که نظریه آماری رایج در آن زمان آمار Pearson بود و بر مفهوم همبستگی تأکید داشت، نظریه کلاسیک آزمون نیز به طور عمده بر این مفهوم قرار گرفت (۷-۵). چارچوب این نظریه ۲۵ سال بعد به وسیله Spearman، Udny Yule، Kelley و سایر افراد در سال ۱۹۰۴ گسترش یافت و پالایش گردید. نقاط عطف در تاریخ این نظریه با انتشار فرمول‌های Kuder-Richardson در سال ۱۹۳۷ و به مدت کوتاهی بعد از آن با ایده حد پایین برای اعتبار و چارچوب مربوط به درک پیشرفت‌های به وجود آمده در کار Guttman دنبال شد. اوج نظریه کلاسیک آزمون در کار Novick تحقق یافت (به نقل از Traub) (۴).

مفصل‌ترین تبیین از این نظریه در کتاب Gulliksen که در سال ۱۹۵۰ منتشر گردید، آمده است. بعدها نیز Lord و Novick در کتاب نظریه آماری آزمون‌های ذهنی Statistical theory of mental test scores در سال ۱۹۶۸ مفاهیم اصلی این نظریه را با استفاده از رویکرد جدید آمار ریاضی مورد تجدید نظر قرار دادند. عنصر اصلی در این نظریه نمره آزمون است و سؤال‌ها و ویژگی‌های روان‌سنجی آن‌ها نقش کمی در ساختار این نظریه دارند (۵). مهم‌ترین فرمول‌های این نظریه عبارتند از:

الف. فرمول اسپیرمن براون (Spearman- brown formula)

ب. فرمول خطای استاندارد اندازه‌گیری (Standard error of measurement formula)

ج. فرمول Kuder-Richardson

د. فرمول‌های تصحیح برای کاهش (Attenuation formula) (۸).

مانند هر نظریه علمی، به تدریج و با گذشت زمان نقطه ضعف‌های نظریه کلاسیک نیز آشکار شد. از این رو متخصصان اندازه‌گیری به تدریج به دنبال مدل‌های کارامدتری برآمدند. هدف از مقاله حاضر، توصیف مفاهیم اولیه، مزایا و معایب نظریه کلاسیک و سؤال پاسخ و مقایسه آن‌ها بود.

نظریه کلاسیک آزمون

نظریه کلاسیک آزمون سه مفهوم اساسی دارد که عبارت از: ۱- نمره آزمون که اغلب نمره مشاهده شده (Observed score) نامیده می‌شود، ۲- نمره واقعی (True score) و ۳- نمره خطا (Error score) می‌باشد. در این نظریه نمره واقعی و نمره خطا، ساخت‌های فرضی و غیر قابل مشاهده هستند. از این رو از آن‌ها به عنوان متغیرهای پنهان (Latent variable) یاد می‌شود. در چارچوب این نظریه مدل‌های متعددی صورت‌بندی شده است که مشهورترین آن‌ها عبارت از: $X = T \pm E$ می‌باشد. از آن جا که در این معادله برای هر آزمون دهنده (هر نمره مشاهده شده) دو مجهول وجود دارد، پس این معادله حل‌نشدنی است. برای کنار آمدن با این مشکل از توزیع‌های آماری و پیش فرض‌های مربوط به آن‌ها استفاده کرده و نمرات غیر قابل مشاهده خطا و واقعی را مدل‌سازی می‌کنند. توزیع نمرات مشاهده شده از طریق آزمون مشخص شده و به دست می‌آید. پس می‌توان با مدل‌سازی نمره خطا یا واقعی، وضعیت یکی از مجهول‌ها را مشخص کرد و سپس با مشخص شدن دو مجهول از سه مجهول می‌توان سومی را پیدا نمود. به طور مثال در جامعه‌ای از افراد توزیع نمرات خطا را نرمال با میانگین صفر و واریانس معینی در نظر می‌گیرند. سپس فرض می‌کنند که الف) نمرات واقعی و خطا ناهمبسته هستند، ب) متوسط نمرات خطا در جامعه آزمون دهندگان صفر است و ج) نمرات خطا در آزمون‌های موازی ناهمبسته است. تحت این پیش فرض‌ها نمرات واقعی از تفاوت بین نمرات آزمون و نمرات خطا به دست می‌آیند. در این فرمول‌بندی نمره واقعی برابر است با نمره مورد انتظار بین فرم‌های موازی.

فرم‌های موازی به عنوان آزمون‌هایی تعریف شده‌اند که محتوای مشابهی را اندازه‌گیری می‌کنند. نمرات واقعی آزمودنی‌ها در این فرم‌ها یکسان است و مقادیر خطای اندازه‌گیری بین فرم‌های موازی مساوی است، هر چند که بر

اساس این مدل ساخت آزمون‌های موازی فقط یک ایده فرضی است (۹). نظریه کلاسیک آزمون بر مدل‌هایی متمرکز است که نمره آزمون را مورد توجه قرار می‌دهند. به عبارتی این نظریه نمرات آزمون را به نمرات واقعی ربط می‌دهد. با این وجود، ارتباط مشخصه‌های آماری سؤال با مشخصه‌های آماری آزمون مانند میانگین، انحراف استاندارد و اعتبار آزمون به خوبی شناخته شده است و از آن‌ها در ساخت آزمون‌ها استفاده می‌شود (۱۰)، ولی همان طور که می‌دانیم این مشخصه‌ها وابسته به نمونه هستند و همین امر ارزش آن‌ها را کاهش می‌دهد. این مشخصه‌ها زمانی مفید می‌باشند که نمونه معرف جامعه‌ای باشد که آزمون برای آن ساخته می‌شود. اگر نمونه معرف جامعه نباشد، ارزش این مشخصه‌ها کاهش می‌یابد. یکی از راه‌حل‌های ارایه شده برای حل این مشکل، استفاده از سؤالات لنگر Anchor است. به این ترتیب مشکلات نمونه‌گیری کاهش یافته، ولی چون روابط غیر خطی است، هر گونه تحلیلی را پیچیده می‌سازد.

تعریف نمره واقعی به عنوان امید ریاضی توزیع نمرات مشاهده شده با استفاده از آزمون‌های موازی همگی بر اساس مفهوم توزیع گرایشی (Propensity distribution) صورت می‌گیرد. در این رویکرد فرض می‌شود که رفتار خاصی، تحت شرایط معین از یک توزیع تبعیت می‌کند. این توزیع حاصل اندازه‌گیری یک رفتار تحت شرایط معین با آزمون‌های موازی و با فرض استقلال مشاهدات از هم است. دستیابی به پارامترهای این توزیع امکان پیش‌بینی همان رفتار تحت شرایط مختلف را فراهم می‌کنند. این ایده زیربنای قضایا و اصول نظریه کلاسیک است (۱۲، ۱۱).

مزایای نظریه کلاسیک آزمون

الف) آشنایی کاربران روان‌سنجی با مفاهیم پایه‌ای آن. پژوهشگرانی که درگیر مسایل اندازه‌گیری هستند آشنایی و انس بیشتری با نظریه کلاسیک اندازه‌گیری دارند. به طور تقریبی اکثر مقیاس‌هایی که در حال حاضر مورد استفاده قرار می‌گیرند بر اساس اصول و قوانین این نظریه طراحی شده‌اند. ب) روش‌های مورد استفاده در آن از نظر اجرایی قابل اجرا هستند. برای مثال بسیاری از برنامه‌های آماری، روش‌های مربوط به اجرای تحلیل‌های این نظریه را دارند. ج) مدل زیربنایی، با انواع ابزارها تناسب دارد. برای مثال، نمرات حاصل از سؤال‌های موجود در یک مقیاس را به عنوان شاخص‌هایی از یک متغیر نهفته مشترک با هم جمع کرده و به یک نمره کل می‌رسیم.

مقیاسی که تحت نظریه کلاسیک آزمون طراحی شده است باید شامل سؤالاتی باشد که همگی واسط بین نمره واقعی و متغیر مورد اندازه‌گیری باشند. به عبارتی، مقدار عددی یک نمره بر روی یک سؤال باید به معنی و مفهوم مشابه مقدار عددی سایر سؤال‌های همان مقیاس باشد. با افزودن سؤال‌های مختلف به هم، اکثر خطاهای مربوط به تک‌تک سؤال‌ها تقلیل یافته و یا خنثی می‌شود. این مجموعه ویژگی‌های یک مقیاس، یک مفهوم به نسبت مشترک در همه ابزارهای اندازه‌گیری در مدل کلاسیک آزمون است. د) الزامی بر بهینه بودن تک‌تک سؤالات وجود ندارد. سؤالاتی که در حالت متوسطی با متغیر زیربنایی مورد اندازه‌گیری ارتباط داشته باشند را می‌توان در مقیاس به شکل موفقیت‌آمیزی مورد استفاده قرار داد. گاهی اوقات طراحی سؤالاتی که قادر باشند به تنهایی و به شکل بهینه متغیر زیربنایی را اندازه‌گیری کنند کار بسیار دشواری است. اگر همبستگی بین سؤال‌ها ضعیف باشد، افزودن سؤال‌های اضافی به مقیاس می‌تواند به رفع این مشکل کمک کند و به لحاظ نظری این کار باعث افزایش مقدار اعتبار آزمون می‌گردد (۱۳).

نقطه ضعف‌های نظریه کلاسیک آزمون

۱- از آن جا که وجود تعداد زیادی سؤال در یک مقیاس ریشه اصلی افزایش دقت در این مدل است، مقیاس‌های ساخته شده با این مدل دارای سؤال‌های زیاد و گاهی مشابه هستند. در برخی موارد، تلاش برای طراحی سؤال‌هایی که با سایر سؤال‌ها همبستگی به نسبت قوی داشته باشند، می‌تواند منجر به شباهت‌های صوری و ظاهری بین سؤال‌های یک مقیاس گردد. وقتی که این مسأله اتفاق بیافتد، نه تنها متغیر مورد

اندازه‌گیری، بلکه ویژگی‌های نامناسبی از سؤال‌ها مثل مشکلات مربوط به ساختار و گرامر زبانی نیز می‌تواند به طور مشترکی در همه سؤال‌ها روی دهد. در اصل، نمره واقعی به عنوان آمیزه‌ای غیر افتراقی (Un-differential mixture) از همه ویژگی‌های سؤال‌ها در حالت مشترک بوده و ویژگی‌های زیربنایی متغیر مورد اندازه‌گیری و ویژگی‌های صوری که مدنظر و هدف مقیاس اندازه‌گیری نبوده را در بر می‌گیرد (۱۴). روش‌های نظریه کلاسیک آزمون به سختی بین موضوعات مشترک سؤالاتی که مقیاس و متغیر مورد اندازه‌گیری و حالت صوری و ظاهری دارند، تمایز قایل می‌شوند.

۲- روش‌های کلاسیک آزمون شامل بررسی دقیق و موشکافانه ویژگی‌های سؤال که در سایر روش‌های جدید اندازه‌گیری وجود دارند، نمی‌گردد. این مسأله می‌تواند به عنوان یک نقطه ضعف قلمداد شود. برای مثال، مقیاس‌های مبتنی بر نظریه کلاسیک آزمون در نقاط مرکزی مقیاس اندازه‌گیری دارای حساسیت و دقت افتراقی بیشتر و در دو انتهای مقیاس حساسیت اندازه‌گیری کمتری دارند. بنابراین، دو نمره متفاوت در مرکز دامنه نمره‌ها می‌توانند نشانگر تفاوت کوچک‌تری در نمره واقعی در مقایسه با دو نمره متفاوت در دو کناره دامنه نمره‌ها باشند (۱۳).

۳- برآورد پارامترها در مدل کلاسیک آزمون وابسته به تک‌تک افراد مورد مطالعه است. ضرایب دشواری، تشخیص و آلفا، مبتنی بر همبستگی‌های به دست آمده از نمونه پژوهش است. نمونه‌های مختلف با واریانس‌های متفاوت نمی‌توانند منجر به نتایج معادل یا نتایج قابل مقایسه‌ای بین نمونه‌های مختلف شوند. آماره‌های برآورد شده با استفاده از این مدل به گروه آزمودنی که آزمون و سؤال‌ها بر روی آن‌ها اجرا می‌شود، وابسته است. ضریب دشواری و تشخیص سؤال‌ها در این مدل از یک گروه نمونه به گروه دیگر تغییر می‌کند. به طور مثال زمانی که آزمودنی‌های گروه نمونه به لحاظ توانایی بالاتر از متوسط باشند، ضریب دشواری به یک نزدیک می‌شود و زمانی که گروه مورد نظر به لحاظ توانایی پایین‌تر از حد متوسط باشند مقدار این ضریب به صفر نزدیک می‌گردد. شاخص تشخیص سؤال‌ها و اعتبار آزمون نیز از این قاعده مستثنی نیستند. به عنوان مثال اگر گروه نمونه مورد نظر نامتجانس باشند، مقدار این دو شاخص افزایش می‌یابد و زمانی که گروه نمونه مورد نظر متجانس باشند مقدار این دو شاخص کاهش می‌یابد.

۴- نمرات مشاهده شده و واقعی به سطح دشواری آزمون وابسته‌اند. به عبارتی، نمرات مشاهده شده و واقعی آزمون با تغییر در دشواری آزمون افزایش و کاهش می‌یابند. اهمیت این نقطه ضعف زمانی روشن می‌شود که بدانیم آزمودنی‌هایی که مقایسه می‌شوند اغلب فرم‌های متفاوتی از یک آزمون و یا حتی بخش‌های متفاوتی در درون یک آزمون دریافت می‌کنند. چگونه می‌توان افرادی که فرم‌های مختلفی از یک آزمون دریافت می‌کنند یا حتی در درون یک آزمون به بخش‌های متفاوتی پاسخ می‌دهند را به طور عادلانه با هم مقایسه کرد؛ در حالی که دشواری آزمون‌ها و خرده مقیاس‌ها متفاوت است و این که نمره واقعی و مشاهده شده افراد بر حسب دشواری یا سادگی آزمون افت و خیز دارد.

۵- نقطه ضعف دیگر نظریه کلاسیک با فرض خطای یکسان اندازه‌گیری برای همه آزمودنی‌ها در ارتباط است. این فرض که یکی از مشهورترین و رایج‌ترین جنبه‌های کاربرد نظریه کلاسیک است غیر موجه به نظر می‌رسد. به ویژه خطای اندازه‌گیری نمرات مقیاس در یک آزمون دشوار برای آزمودنی‌هایی با توانایی کم در مقایسه با آزمودنی‌هایی با توانایی متوسط یا بالا بیشتر است. موارد تخطی از فرض واریانس خطای یکسان رایج است و مدل‌هایی که در آن‌ها این فرض نقض نشود، قابل ترجیح هستند. با این حال مواردی که در آن این مفروضه نقض می‌شود، برای سودمندی کلی مدل کلاسیک آزمون که گاهی اوقات از آن به عنوان مدل ضعیف نمره واقعی یاد می‌شود، تهدید کننده نمی‌باشد.

۶- آخرین نقطه ضعف نظریه کلاسیک آزمون به تعریف آزمون‌های موازی بر می‌گردد. در ارتباط با این تعریف باید گفت که در عمل ساخت دو آزمون به طور کامل موازی مشکل است و تخطی از این مفروضه نتایج مربوط به نظریه

کلاسیک را تضعیف می‌کند. در مدل کلاسیک، مفروضه دو آزمون موازی به عنوان مبنایی در تعریف این مدل از اعتبار آزمون به کار گرفته می‌شود. در عمل نه تنها آزمون‌های موازی به دست نمی‌آیند، بلکه در حقیقت، ساخت آزمون‌های به طور کامل موازی در عمل ناموفق است. در واقع آزمون‌های ناموازی که موازی فرض می‌شوند منجر به برآورد نادرستی از اعتبار آزمون، خطای استاندارد اندازه‌گیری و طول آزمون مورد نیاز برای رسیدن به اعتبار مطلوب می‌شود (۸). نقطه ضعف‌های نظریه کلاسیک آزمون می‌تواند در برخی از زمینه‌ها در مقایسه با سایر زمینه‌ها مشکلات بیشتری به وجود آورد. مانند مقایسه پارامترهای سؤال و آزمون در بین جوامع پژوهشی متفاوت. به همین دلیل بررسی کارکرد افتراقی سؤال (Differential item functioning یا DIF) و آزمون‌ها (۱۶، ۱۵) که در بافت نظریه کلاسیک با روش‌هایی مثل نمودار دلتا (Delta plot) و Mantel hanzel صورت می‌گیرد، قابل اطمینان نیستند (۷).

توسعه نظریه سؤال پاسخ در بطن نظریه کلاسیک آزمون

با مرور زمان مشخص شد که به جای نظریه کلاسیک آزمون که به نمره آزمون متکی است و سؤالات و ویژگی‌های آن‌ها نقش چندانی در آن ندارند باید نظریه‌ای طراحی کرد که به جای نمره آزمون بر ویژگی‌های سؤال‌های تشکیل دهنده آزمون متکی بوده و دارای ثبات در برآورد پارامترها نیز باشد. نخستین گام‌ها در همین رابطه استفاده از ضریب همبستگی دو رشته‌ای به جای دو رشته‌ای نقطه‌ای بود. بررسی‌ها نشان داد که پایداری ضریب دو رشته‌ای از دو رشته‌ای نقطه‌ای در نمونه‌های مختلف بیشتر است. متخصصان اندازه‌گیری در حوزه نظری و عملی به شکاف موجود در نظریه کلاسیک واقف بودند. در واقع آن‌ها به دنبال نظریه‌ای در اندازه‌گیری بودند که دارای ویژگی‌های مطلوبی باشد. یعنی، الف) آماره‌های برآورد شده به گروه نمونه وابسته نباشند. ب) نمرات توصیف کنند که چیرگی آزمودنی‌ها به دشواری آزمون وابسته نباشند. ج) مدل‌هایی از آزمون پایه‌ای برای منطبق کردن سؤالات آزمون با سطوح توانایی فراهم کنند. د) مدل‌هایی از آزمون بر پایه مفروضات غیر موجه نباشند. و) مدل‌هایی از آزمون برای برآورد اعتبار به آزمون‌های به طور کامل موازی نیاز نداشته باشند. این پنج ویژگی مطلوب به همراه سایر ویژگی‌های مطلوب آزمون‌سازی در چارچوب نظریه سؤال پاسخ به دست می‌آیند (۸).

به لحاظ تاریخی ریشه‌های چنین نظریه آزمونی را می‌توان در کار Binet و Simon در سال ۱۹۱۶ مشاهده کرد. اصلی که این افراد مورد استفاده قرار دادند این بود که بین نسبت پاسخ صحیح به یک سؤال و سن تقویمی افراد ارتباطی برقرار کردند؛ به طوری که با افزایش سن، نسبت پاسخ صحیح نیز افزایش یافت. بعدها افرادی مثل Terman و Merill به جای این که روش Binet و Simon را برای هر سؤال به صورت جدول عرضه کنند، از منحنی‌هایی برای این کار استفاده کردند. در واقع آن‌ها برای ترسیم منحنی ارتباط دهنده این دو متغیر، سن تقویمی را بر روی محور xها و نسبت پاسخ صحیح را بر روی محور yها قرار دادند، نمودارهایی که به این صورت به دست می‌آیند چیزی است که ما امروزه منحنی ویژگی سؤال می‌نامیم (۵).

Thurston نیز در سال ۱۹۲۵ از چنین رویکردی استفاده کرد و توانست منحنی‌های S شکل را با استفاده از تابع توزیع نرمال تراکمی بازنمایی کند. این منحنی سال‌های متمادی به عنوان یک روش متمایز برای تحلیل سؤالات مورد استفاده قرار گرفت (۱۷). نشانه‌های آغاز نظریه‌ای که مبتنی بر سؤالات باشد در پژوهش Lawly در سال ۱۹۴۳ دیده می‌شود. وی پارامترهای منحنی ویژگی سؤال را بر اساس روش بیشینه درست‌نمایی (Maximum likelihood) به دست آورد و نشان داد که بسیاری از شاخص‌های نظریه کلاسیک را می‌توان بر اساس تابعی از پارامترهای سؤال‌ها بازگو نمود.

بعد از Lawly، Lord تلاش عظیمی برای گسترش کار وی انجام داد. تلاش‌های این دو نفر مفاهیم بنیادی روان‌سنجی- آن چه که امروزه به نظریه سؤال پاسخ معروف

است- را پایه‌ریزی کرد. فرایندی که بعدها به وسیله افرادی مانند George Rash، Birnbam و ... ادامه یافت (۵). اولین بار Lord در رساله دکترای خود در سال ۱۹۵۲ و چندین مقاله بعد از آن بر کاربرد مدل‌ها و مفاهیم نظریه سؤال پاسخ برای رسیدن به ثبات در شاخص‌های روان‌سنجی سؤال تأکید کرد. از نظر تاریخی نخستین بار Birnbam در سال‌های ۱۹۵۷ و ۱۹۵۸ مجموعه مقالات فنی در این زمینه را نوشت که به معرفی مدل‌های منطقی آزمون و روش‌های برآورد پارامترهای این مدل پرداخت. سپس George Rash در سال ۱۹۶۰ کتابی در این خصوص منتشر کرد و در نهایت در دهه ۱۹۹۰ با کارهای Lord و Right توجه قابل ملاحظه‌ای به نظریه سؤال پاسخ معطوف شد (۱۸). نظریه سؤال پاسخ همان نظریه صفت پنهان برای متغیرهای دو ارزشی است (۱۹).

نظریه سؤال پاسخ

نظریه سؤال پاسخ یک نظریه آماری است که به عملکرد آزمون دهندگان در سؤال و آزمون مربوط است. این نظریه به بررسی این مطلب می‌پردازد که چگونه عملکرد افراد در آزمون و سؤال‌ها به توانایی‌هایی که به وسیله سؤالات آزمون اندازه‌گیری می‌شود، ارتباط پیدا می‌کند. در این نظریه روش‌ها و مدل‌های زیادی را می‌توان مشخص کرد که پاسخ به سؤالات را به توانایی یا توانایی‌های زیربنایی ارتباط می‌دهند. با آن که در چارچوب نظریه سؤال پاسخ مدل‌های زیادی وجود دارند که در مورد داده‌های واقعی به کار رفته‌اند، ولی در مقاله حاضر تنها بر مدل‌هایی تأکید می‌شود که دارای این ویژگی‌ها هستند. الف) یک توانایی زیربنای عملکرد در آزمون است، ب) در مورد داده‌های دو ارزشی به کار می‌روند و ج) رابطه بین عملکرد سؤال و توانایی را با یک تابع منطقی یک، دو یا سه پارامتری توصیف می‌کنند. به طور کل دو فرض در تعیین مدل‌های نظریه سؤال پاسخ مطرح است. یکی از این مفروضه‌ها به بعد ساختاری داده‌های آزمون و دیگری به شکل ریاضی تابع یا منحنی ویژگی سؤال (ICC یا Item characteristic curve) مربوط می‌شود. این نظریه دارای ویژگی‌هایی است که در ادامه توصیف می‌شوند.

تابع سؤال پاسخ (Item response function یا IRF)

عنصر بنیادی نظریه سؤال پاسخ، توابع سؤال پاسخ است. در مورد سؤال‌های موجود در یک مقیاس درجه‌بندی، تابع سؤال پاسخ، یک تابع ریاضی است که رابطه بین جایگاهی که فرد در پیوستار یک سازه معین (نظیر افسردگی یا هر خصیصه دیگر) و احتمال آن که آن فرد به یک سؤال مقیاس که برای اندازه‌گیری آن سازه طراحی شده است پاسخ خاصی بدهد را توصیف می‌کند. در نظریه سؤال پاسخ، سازه، خصیصه پنهان نامیده می‌شود؛ چرا که فرض می‌شود خصیصه زیربنایی اصلی است و به طور مستقیم پاسخ به سؤال‌های موجود در یک مقیاس را که برای اندازه‌گیری آن خصیصه طراحی شده، تحت تأثیر قرار می‌دهد.

هدف اصلی مدل‌بندی تابع سؤال پاسخ این است که برای هر سؤال موجود در یک ابزار یک تابع سؤال پاسخ تعیین کند. از سوی دیگر توابع سؤال پاسخ برای ارزیابی کیفیت سؤال‌ها مورد استفاده قرار می‌گیرند و به عنوان عناصر اصلی برای به دست آوردن سایر ویژگی‌های روان‌شناختی مهم به کار برده می‌شوند. در اصطلاح تابع سؤال پاسخ، نقطه‌ای در امتداد محور خصیصه پنهان که منحنی تابع سؤال پاسخ تغییر جهت می‌دهد (نقطه عطف تابع) دشواری سؤال است. یک فرد برای پاسخ دادن به یک سؤال دشوارتر نیاز به سطح بالاتری از خصیصه دارد. پارامتر دشواری سؤال در تابع سؤال پاسخ مشابه میانگین سؤال در نظریه کلاسیک است. شیب تابع سؤال پاسخ در نقطه عطف منحنی (یعنی نقطه دشواری سؤال) قدرت تشخیص سؤال نامیده می‌شود. سؤالاتی که پارامتر تشخیص آن‌ها بیشتر است قادر هستند افراد موجود در اطراف نقطه دشواری سؤال در خصیصه را بهتر متمایز کنند. پارامتر تشخیص در مدل‌های نظریه سؤال پاسخ مشابه همبستگی سؤال با آزمون در نظریه کلاسیک یا بار عاملی (همبستگی سؤال- عامل) در تحلیل عاملی است (۲۰).

تابع ویژگی آزمون (Test characteristic function یا TCF)

یک ویژگی جالب نظریه سؤال پاسخ، تابع ویژگی آزمون

است. این تابع برابر با مجموع همه توابع پاسخ یک آزمون است. می‌توان از تابع ویژگی آزمون برای پیش‌بینی نمرات آزمودنی‌ها در سطوح مختلف توانایی استفاده کرد. اگر آزمونی از سؤالات دشوار تشکیل شده باشد، تابع ویژگی آزمون به سمت راست حرکت کرده و انتظار می‌رود که آزمون دهندگان نمرات مورد انتظار پایینی بگیرند. در مقابل اگر سؤالات آزمون ساده باشند، منحنی ویژگی آزمون به سمت چپ کشیده می‌شود و نمره مورد انتظار افراد بالا خواهد بود. از این رو این امکان وجود دارد که از طریق تابع ویژگی آزمون تبیین کنیم که صرف نظر از نمرات خطا، چگونه آزمودنی‌هایی با توانایی ثابت در دو آزمون که توانایی یکسانی را اندازه‌گیری می‌کنند، به طور متفاوت عمل خواهند کرد. تابع ویژگی آزمون، نمره‌های توانایی موجود در نظریه سؤال پاسخ را به نمره‌های واقعی نظریه کلاسیک ارتباط می‌دهد؛ چرا که نمره مورد انتظار آزمون دهنده در سطح خاصی از توانایی همان تعریف نمره واقعی آزمون دهنده در آن مجموعه از سؤال‌ها است (۲۱).

تابع آگاهی سؤال (Item information function یا IIF)

برای قضاوت کردن در مورد دقت اندازه‌گیری یک سؤال، می‌توان تابع سؤال پاسخ هر سؤال را به یک تابع آگاهی تبدیل کرد که نشان می‌دهد سؤال مورد نظر در هر سطح خصیصه چه مقدار آگاهی روان‌سنجی را فراهم می‌کند. تابع آگاهی سؤال، سهم یک سؤال خاص در اندازه‌گیری توانایی را نشان می‌دهد. سؤال‌های گوناگون می‌توانند مقادیر آگاهی متفاوتی در نقاط مختلف یک خصیصه پنهان خاص را فراهم کنند. سؤال‌های به نسبت آسان برای تفکیک کردن افراد ضعیف در خصیصه بهتر هستند؛ در حالی که سؤال‌های به نسبت دشوار برای تفکیک افراد قوی در خصیصه مناسب‌تر هستند. از این رو، سؤال‌های دارای دشواری متفاوت در نقاط مختلف خصیصه پنهان آگاهی متفاوتی ارایه می‌دهند و سؤال‌هایی که قدرت تشخیص بیشتری دارند در مقایسه با سؤال‌هایی که قدرت تشخیص کمتری دارند آگاهی بیشتری فراهم می‌کنند. در کل سؤال‌های با قدرت تشخیص بالا در مقایسه با سؤال‌های دارای قدرت تشخیص پایین، سهم بیشتری در دقت اندازه‌گیری دارند. بیشترین سهم یک سؤال در دقت اندازه‌گیری در اطراف پارامتر دشواری (b) آن‌ها به دست می‌آید. در واقع پارامتر دشواری، موقعیت تابع آگاهی را در مقیاس توانایی مشخص می‌کند و قدرت تشخیص سؤال میزان آگاهی را تعیین می‌کند (۱۸). در مدل یک پارامتری چون همه سؤال‌ها شیب یکسانی دارند پس آگاهی همه سؤال‌ها نیز یکسان است و فقط بر حسب پارامتر دشواری موقعیت این تابع بر روی مقیاس توانایی تغییر می‌کند. در مدل دو پارامتری چون پارامتر شیب در سؤال‌های مختلف فرق می‌کند، پس مقدار آگاهی سؤال‌ها نیز متفاوت است. در مدل سه پارامتری –که عامل حدس نیز در آن وجود دارد– مقدار آگاهی زمانی به حداکثر می‌رسد که مقدار حدس مساوی صفر باشد. هر چه حدس زیاد شود مقدار آگاهی کاهش می‌یابد (۲۲).

تابع آگاهی آزمون (Test information function یا TIF)

برای تشکیل تابع آگاهی آزمون، می‌توان توابع آگاهی مربوط به سؤال‌های مختلف را با هم جمع کرد. از آن جا که آگاهی به طور مستقیم به دقت اندازه‌گیری مربوط می‌شود (آگاهی بیشتر معادل با دقت اندازه‌گیری بیشتر است) تابع آگاهی مقیاس، میزان دقت توابع یک ابزار در نقاط مختلف خصیصه به عنوان یک کل را برآورد می‌کند. تابع آگاهی با علامت $I(\theta)$ نشان داده می‌شود (۲۳). تابع آگاهی آزمون در واقع خطای مربوط به برآورد توانایی را نشان می‌دهد. بین خطای برآورد توانایی و تابع آگاهی رابطه $SE(\theta)=\frac{1}{\sqrt{I(\theta)}}$ برقرار است. با توجه به این رابطه می‌توان گفت بیشترین دقت اندازه‌گیری یک آزمون در نقطه خاصی از مقیاس توانایی فراهم می‌شود، نقطه‌ای که خطای اندازه‌گیری در آن به حداقل می‌رسد. وجود توابع آگاهی سؤال و آزمون، اساس روش‌هایی را که آزمون‌ها در چارچوب نظریه سؤال پاسخ ساخته می‌شوند را تغییر داده است. تابع ویژگی سؤال و آزمون

و نیز تابع آگاهی سؤال و آزمون ویژگی‌های مکمل مدل‌های نظریه سؤال پاسخ بوده و بسیار مفید هستند (۲۴). هر چند که ویژگی اصلی این توابع چیزی است که از آن به عنوان ثبات یا تغییرناپذیری (Invariance) پارامترهای مدل نام می‌برند.

تغییرناپذیری

ثبات پارامترها در نظریه سؤال پاسخ هم برای پارامترهای سؤال و هم برای پارامترهای افراد مطرح است و تنها با مدل‌هایی به دست می‌آید که با داده‌های آزمونی که مدل برای آن‌ها به کار برده شده برازش داشته باشد. اگر مدلی با داده‌ها برازش نداشته باشد بدون شک ثبات پارامترها نیز وجود نخواهد داشت. این واقعیت که توابع آگاهی سؤال‌ها را می‌توان با هم جمع کرد، مبنایی برای ساخت مقیاس در نظریه سؤال پاسخ است. آگاهی سؤال و مقیاس در نظریه سؤال پاسخ مشابه با اعتبار سؤال و آزمون در نظریه کلاسیک است. هر چند که یک تفاوت مهم بین آن‌ها این است که در چارچوب نظریه سؤال پاسخ آگاهی (دقت) می‌تواند بر حسب این که یک فرد در کجای دامنه خصیصه قرار گرفته باشد تغییر کند؛ در حالی که در نظریه کلاسیک، اعتبار مقیاس (دقت) برای همه افراد صرف نظر از سطوح نمرات خامشان یکسان است (۲۵).

در مدل‌های نظریه سؤال پاسخ تغییرناپذیری دو معنی دارد، اول موقعیت یک فرد در پیوستار خصیصه پنهان را می‌توان از روی پاسخ‌هایش به هر مجموعه سؤال با توابع سؤال پاسخ مشخص برآورد کرد، حتی اگر سؤال‌ها از ابزارهای متفاوتی باشند. در مقابل در نظریه کلاسیک پاسخ سؤال‌ها جمع می‌شوند تا نمره واقعی که تنها به آن مقیاس تعلق دارد را برآورد کنند. دوم، ویژگی‌های سؤالات، همان طور که به وسیله تابع سؤال پاسخ مشخص شد، به ویژگی‌های جامعه خاصی وابسته نیستند. همچنین مقیاس خصیصه به هیچ مجموعه خاصی از سؤالات وابسته نیست، بلکه به طور مستقل وجود دارد (۲۶). در نظریه کلاسیک میانگین سؤال‌ها و همبستگی‌های سؤال- آزمون، همین طور ضرایب اعتبار و خطاهای استاندارد به ویژگی‌های جوامع خاص وابسته‌اند. در نظریه کلاسیک، مقیاس نمره خام و واقعی به وسیله مجموعه سؤال‌های خاص موجود در یک ابزار منفرد تعریف می‌شوند. به عنوان مثال، در سنجش آموزش در سطح گسترده، مقیاس تغییرناپذیری پارامترهای سؤال پیوند دادن مقیاس‌های مربوط به ابزارهای مختلف (یعنی قرار دادن مقیاس‌ها در یک مقیاس منفرد مشترک) بین دانش‌آموزان در سطوح تحصیلی متفاوت (به طور مثال پایه ۳ تا ۶ در یک مدرسه) و در داخل یک سطح تحصیلی (به طور مثال پایه‌های چهارم در مدارس مختلف) را تسهیل می‌کند. همین طور استفاده از روش‌های نظریه سؤال پاسخ برای مقایسه افرادی که به ابزارهای متفاوتی پاسخ داده‌اند در تحقیقات بین فرهنگی و رشدی آشکار است (۲۷).

انتخاب مدل

اگر سؤالات تشخیص برابری نداشته باشند بهتر است که از مدل‌های ۲ و ۳ پارامتری به جای مدل یک پارامتری استفاده شود. برای بررسی تأثیر پارامتر حدس در پاسخ دادن به سؤالات بهتر آن است که مشخص نماییم آزمون دهندگان با کمترین توانایی تا چه میزان به سؤالات بسیار دشوار در آزمون پاسخ داده‌اند. اگر این افراد به سؤالات بسیار دشوار پاسخ نداده باشند می‌توان از این پارامتر چشم‌پوشی کرد، اما اگر این گونه افراد به سؤالات بسیار دشوار بیشتر از مقدار مورد انتظار پاسخ داده باشند، باید پارامتر حدس در مدل گنجانده شود که در این صورت مدل سه پارامتری بر مدل‌های یک و دو پارامتری ارجحیت دارد (۲۸). روش دیگر برای انتخاب بهترین مدل استفاده از تابع آگاهی است. بر اساس این روش مدلی بهترین برآورد از توانایی افراد را ارایه خواهد نمود که در نقطه میانگین توزیع توانایی افراد بیشترین آگاهی و کمترین خطا را داشته باشد (۸). برای انتخاب بهترین مدل متناسب با داده‌های مشاهده شده می‌توان از مقایسه سؤالات نیز استفاده کرد (۲۹).

بحث و نتیجه‌گیری

مقایسه مدل‌های موجود در نظریه کلاسیک و سؤال پاسخ به

وسیله افراد زیادی مورد بررسی قرار گرفته است. رابطه بین قدرت تشخیص و دشواری در نظریه کلاسیک و مدل دو پارامتری لوجستیک در نظریه سؤال پاسخ به وسیله Lord (۲۹) مورد بررسی قرار گرفته است. وی نشان داد که تحت شرایط خاصی (مثل زمانی که عملکرد افراد به وسیله حدس تحت تأثیر قرار نمی‌گیرد) همبستگی دو رشته‌ای بین سؤال و آزمون در چارچوب نظریه کلاسیک و پارامتر شیب سؤال در نظریه سؤال پاسخ به طور تقریبی یکنواخت و به صورت تابعی از یکدیگر افزایش می‌یابند که این رابطه را می‌توان به این صورت نشان داد.

$$a_i \cong \frac{r_i}{\sqrt{1 - r_i^2}}$$

در این رابطه a_i شیب سؤال در نظریه سؤال پاسخ r_i و همبستگی دو رشته‌ای در مدل کلاسیک است. به دلیل آن که توزیع‌ها و نمره‌های اختصاص یافته در هر دو مدل متفاوت می‌باشد، این رابطه تقریبی است. توزیع نمره (X) در نظریه کلاسیک و نمره توانایی (θ) در نظریه سؤال پاسخ شکل‌های متفاوتی دارند و رابطه بین X و θ غیر خطی است. به علاوه نمره کل آزمون (X) در نظریه کلاسیک تابعی از خطاهای اندازه‌گیری است؛ در حالی که نمره توانایی (θ) این طور نیست. Lord مشابه همین رابطه را بین p و b، هنگامی که همه سؤال‌ها قدرت تشخیص یکسانی دارند (نظیر مدل George Rash) بیان کرده است؛ به طوری که با افزایش p، b کاهش می‌یابد. اگر سؤالات مقادیر تشخیص نامساوی داشته باشند، پس رابطه بین p و b به r وابسته خواهد بود که می‌توان به این صورت نشان داد:

$$b_i \cong \frac{y_i}{r_i}$$

در این رابطه b_i پارامتر دشواری برای سؤال i در مدل سؤال پاسخ و y_i نمره استاندارد توزیع نرمال متناظر با نمره توانایی در جایی که مقدار p قرار می‌گیرد. شاید مهم‌ترین تمایز بین نظریه‌های کلاسیک و نظریه‌های جدید آزمون، ثبات پارامترهای سؤال و توانایی در نظریه سؤال پاسخ باشد. این ویژگی پیامدهایی دارد که در اکثر منابع به آن‌ها اشاره می‌شود. یعنی، وابسته نبودن پارامترهای افراد به سؤال‌های موجود در آزمون و وابسته نبودن پارامترهای سؤال‌ها به توزیع توانایی آزمودنی‌ها. البته این نتایج تحت برازش مدل با داده‌ها به دست خواهند آمد.

در نظریه کلاسیک آزمون نه تنها پارامترهای افراد و سؤال‌ها ثبات ندارند، بلکه تشخیص نیز در نمونه‌های نامتجانس زیاد و در نمونه‌های متجانس کاهش می‌یابد. ضریب دشواری نزدیک یک نیز از نمونه آزمودنی‌هایی که سطح توانایی آن‌ها بالاتر از میانگین است به دست می‌آید و مقادیر نزدیک صفر این شاخص از نمونه آزمودنی‌هایی به دست می‌آید که توانایی آن‌ها پایین‌تر از میانگین قرار دارد. نمره‌های به دست آمده برای افراد نیز در مدل وابسته به آزمون هستند؛ چرا که دشواری آزمون، به طور مستقیم نمره‌ها را تحت تأثیر قرار می‌دهد (۳۰). مدل نمره واقعی که قسمت اعظم نظریه کلاسیک آزمون بر آن مبتنی است، بررسی پاسخ آزمودنی‌ها به یک سؤال خاص را جایز نمی‌داند، از این رو هیچ مبنایی برای پیش‌بینی این که یک آزمودنی یا یک گروه از آزمودنی‌ها چه طور در یک سؤال خاص عمل می‌کنند، وجود ندارد. در حقیقت نظریه کلاسیک آزمون به عنوان یک نظریه آزمون مرجع و نظریه سؤال پاسخ به عنوان یک نظریه سؤال مرجع توصیف می‌شود (۳۱).

بر عکس، نظریه سؤال پاسخ به متخصصان اندازه‌گیری اجازه انعطاف‌پذیری بیشتری می‌دهد و طیف وسیعی از تعبیر و تفسیرها را در سطح سؤال ممکن می‌سازد. بنابراین نظریه سؤال پاسخ به متخصصان اندازه‌گیری اجازه می‌دهد احتمال آن که یک آزمودنی خاص به یک سؤال معین پاسخ صحیح بدهد را تعیین کنند. اگر سازنده آزمون نیاز به دانستن مشخصات نمره‌های آزمون یک یا چند تن از آزمودنی‌های یک جامعه آماری داشته باشد این ویژگی امتیاز بارزی است. به طور مشابه اگر لازم باشد آزمونی با ویژگی‌های خاصی برای جامعه معینی از آزمودنی‌ها طرح شود، مدل‌های سؤال پاسخ به سازنده آزمون اجازه می‌دهد که این کار را دقیق انجام دهد. این ویژگی نظریه سؤال پاسخ در مقایسه با کاربردهای جدید خاص، نظیر سنجش انطباقی، زیاد ارزشمند

نیست. با این وجود، مدل‌های سؤال پاسخ نیز دارای معایب فنی و عملی هستند. از جنبه فنی مدل‌های نظریه سؤال پاسخ پیچیده هستند و ممکن است در عمل مشکلات مربوط به برآورد پارامترهای مدل به وجود آیند.

مشکل دیگر، برازش مدل است. هنوز به طور کامل روشن نیست که چگونه باید مشکلات برازش مدل را نشان داد، به ویژه مشکلاتی که مربوط به ابعاد آزمون است. چالش‌های عمده‌ای که نظریه سؤال پاسخ (به ویژه در حوزه اندازه‌گیری نگرش و شخصیت) با آن مواجه است عبارتند از: انتخاب مدل مناسب، تعداد آزمودنی‌های مورد نیاز برای برآورد خوب پارامترها و برازش مدل با داده‌ها. هیچ پاسخ روشن و قطعی در خصوص هیچ یک از این موارد وجود ندارد. همان طور که ذکر شد، این مشکلات در حوزه اندازه‌گیری شخصیت و نگرش با این نظریه بارزتر هستند تا در حوزه آزمون‌های پیشرفت تحصیلی. از سوی دیگر از نظر نمره‌گذاری نیز تاکنون هیچ کس در داده‌های واقعی یک یافته روان‌شناختی که ترجیح نمره‌های سؤال پاسخ بر نمره‌های به دست آمده از مقیاس خام را نشان دهد به دست نیاورده است (این موضوع متفاوت از مزیت یکسان بودن مقیاس نمره‌ها و پارامترها است). همبستگی بین نمره‌های به دست آمده از مقیاس نمره‌های خام و نمره‌های به دست آمده از نظریه سؤال پاسخ، در اکثر مواقع بیش از ۰/۹۵ است. از سوی دیگر، تمام مزایای نظریه سؤال پاسخ در صورتی حاصل می‌شوند که مدل با داده‌ها برازش داشته باشد، مسأله‌ای که هنوز جواب روشنی برای آن وجود ندارد. در نهایت، به طور دقیق معلوم نیست که نقض شدن مفروضه‌ها چه تأثیری بر روی برآورد پارامترها دارد (۳۲).

از نظر عملی، به طور تقریبی بدون توجه به کاربرد، سؤال‌های فنی پیچیده‌تری در این نظریه در مقایسه با مدل‌های کلاسیک مطرح می‌شود. به طور یقین کاربرد مدل تک پارامتری در مقایسه با سایر مدل‌های نظریه سؤال پاسخ ساده‌تر است. از سوی دیگر به دلیل محدودیت‌های مربوط به مفروضه‌های این مدل، سؤال‌هایی نیز در مورد برازندگی مدل تک پارامتری مطرح است. وجود محدودیت‌های موجود در نظریه کلاسیک و فواید بالقوه‌ای که در نظریه سؤال پاسخ مطرح می‌شود، باعث شد که بعضی از کاربران اندازه‌گیری تنها نظریه سؤال پاسخ را برای کار انتخاب کنند. به طور کلی تأکید بر کاربرد مدل‌های سؤال پاسخ به شرط برقراری مفروضه‌ها این مزایا را به همراه دارد. الف) مستقل بودن پارامترهای سؤال از نمونه ب) مستقل بودن پارامتر توانایی از آزمون و سؤالات آن ج) مدل‌های آزمون پایه‌ای برای کنار هم گذاشتن سؤال‌های آزمون و سطوح توانایی فراهم می‌کنند د) مدل‌های آزمون برای سنجش اعتبار نیازی به آزمون‌های موازی ندارند.

مزایایی که از کاربرد مدل‌های کلاسیک آزمون در مسایل اندازه‌گیری حاصل می‌شوند عبارتند از: الف) برای تحلیل‌ها، حجم نمونه کمتری مورد نیاز است. ب) در مقایسه با نظریه سؤال پاسخ تحلیل‌های ریاضی ساده‌تری دارد. ج) برآورد پارامتر مدل به لحاظ مفهومی ساده است. د) تحلیل‌ها نیازی به مطالعه خوبی برازش، برای اطمینان از برازش مدل با داده‌های آزمون ندارند. تفاوت‌های عمده بین دو نظریه کلاسیک آزمون و نظریه سؤال پاسخ (۳۳) در جدول ۱ آمده است. برخی از محدودیت‌های نظریه کلاسیک آزمون تنها در حالت نظری اثبات شده است تا در عمل. در برخی از پژوهش‌ها همبستگی بین نمره‌های حاصل از نظریه کلاسیک و نظریه‌های جدید اندازه‌گیری بیشتر از ۰/۸۵ گزارش شده است (۱۳). همبستگی‌هایی در این حد حاکی از آن هستند که این دو نظریه در واقع مکمل یکدیگر می‌باشند. در واقع می‌توان گفت که امروزه به شکل صریح و دقیق مقدار و میزان ارتباط بین نظریه کلاسیک آزمون و نظریه‌های جدید اندازه‌گیری مشخص شده است. برخی از ویژگی‌های نظریه سؤال پاسخ که در نظریه کلاسیک وجود ندارد و به انعطاف‌پذیری آن کمک می‌کند این است که توانایی و پارامترهای سؤال بر روی یک مقیاس مشترک قرار دارند. این مطلب که می‌دانیم یک سؤال کجای مقیاس توانایی بهترین اندازه‌گیری را انجام می‌دهد و دانستن رابطه دقیق بین

عملکرد سؤال و توانایی مواردی هستند که به این انعطاف‌پذیری کمک می‌کنند.

نظریه کلاسیک به خوبی و به سهولت در موقعیت‌های پژوهشی به عنوان یکی از روش‌های پژوهشی در سطح کلان و خرد مورد استفاده پژوهشگران است. پژوهشگران نباید نظریه کلاسیک آزمون را در موقعیت‌های نامتناسب مثل بررسی وضعیت گروه‌های سنی مختلف و اقلیت‌های قومی و مذهبی به کار برند. با وجود نقطه ضعف‌های نظریه کلاسیک هنوز شواهدی مستند و قوی برای کنار گذاشتن کامل این روش ارایه نشده است. نظریه کلاسیک آزمون به دلیل ویژگی‌هایی مانند، آشنایی عمیق پژوهشگران با مفاهیم آن، درک گسترده و راحت مفروضه‌های آن، کاربرد راحت و تجارب موفقیت‌آمیز قبلی مورد استفاده قرار می‌گیرد. هر چند که مسایلی مثل بررسی وضعیت گروه‌های مختلف سنی و اقلیت‌ها با مدل کلاسیک آزمون تناسب ندارند، اما استفاده از روش‌های کلاسیک آزمون در سایر حوزه‌ها همچنان به قوت خود باقی است و نظریه سؤال پاسخ به عنوان نیروی مکمل وارد میدان شده نه رقیبی که بتواند به کلی آن را کنار بگذارد (۳۴، ۱۳). به طور کلی نمی‌توان به دلیل نقطه ضعف‌های نظریه کلاسیک، از آن در سنجش انطباقی و تشکیل بانک سؤال استفاده کرد. این موضوعات در چارچوب نظریه سؤال پاسخ قابل انجام هستند.

جدول ۱. تفاوت‌های عمده بین نظریه کلاسیک و سؤال پاسخ

حیطه	CTT*	نظریه سؤال پاسخ
۱. مدل	خطی $X = T \pm E$	غیر خطی $P_i(\theta) = C_i + \frac{(1+C_i)e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$
۲. سطح مورد بررسی	آزمون	سؤال
۳. مفروضه‌ها	ضعیف (به راحتی با داده‌های آزمون برازش می‌کنند) *$E_{(e)} = 0$ *$P_{(TE)} = 0$ *$P_{e\text{ر}er} = 0$	نیرومند (به سختی با داده‌ها برازش پیدا می‌کنند) * تک بعدی بودن * استقلال موضعی
۴. خطای اندازه‌گیری	خطا = نمره مشاهده شده– غیر واقعی	خطا = نمره مشاهده شده– پاسخ پیش‌بینی شده George Rash = لوجیت (Logit) + باقی‌مانده نظریه سؤال پاسخ = توانایی ± خطا
۵. تفسیر نمره	$X = SEM$	که در این جا نمره بیانگر احتمال پاسخ درست به یک سؤال در سطح مشخصی از صفت پنهان است.
۶. رابطه توانایی با سؤال	نامشخص	منحنی ویژگی سؤال
۷. آماره‌های سؤال	r و P	a و b به اضافه توابع آگاهی و منحنی ویژگی سؤال‌ها
۸. توانایی	نمرات آزمون یا نمرات واقعی برآورد شده روی مقیاس نمره آزمون گزارش می‌شوند (به یک مقیاس مشترک تبدیل می‌شوند)	نمرات توانایی در مقیاس ∞- تا ∞+ یا بر روی یک مقیاس تبدیل شده گزارش می‌شوند.
۹. ثبات آماره‌های سؤال و شخص	بی‌ثباتی پارامترهای سؤال و افراد وابسته به نمونه است.	باثبات– اگر مدل با داده‌ها برازش کند پارامترهای سؤال و افراد مستقل از نمونه هستند.
۱۰. حجم نمونه	۲۰۰ تا ۵۰۰ آزمودنی	بر اساس مدل اندازه‌گیری متفاوت است، اما در مجموع نمونه‌ای بزرگ‌تر از ۵۰۰ مناسب هستند.

*CTT: Classical test theory

References

1. Hambleton RK, Van der Linden WJ. Advances in item response theory and applications: An introduction. Applied Psychological Measurement 1982; 6(4): 373-8.
2. Courville TG. An empirical comparison of item response theory and classical test theory item/person statistics. [Thesis]. Texas, US: Texas A&M University. 2004.
3. Spearman C. "General intelligence," objectively determined and measured. The American Journal of Psychology 1904; 15(2): 201-91.
4. Traub RE. Classical test theory in historical perspective. Educational Measurement: 1997; 16(4): 8-14.
5. Baker FB, Kim SH. Item response theory: Parameter estimation techniques. 2nd ed. London, UK: Taylor & Francis; 2004.
6. Ezanloo B, Habibi M. An introduction to the modern approach in educational and psychological measurement. Quarterly Journal Educational Psychology 2007; 2(8): 135-65. [In Persian].
7. Ezanloo B, Habibi Asgarabadi M. Identifying differential item function (DIF) based on item-response theory: Application of the one parameter model using the BILOG MG software. Journal of Applied Psychology 2010; 4(2): 20-31. [In Persian].
8. Hambleton RK. Principles and selected application of item response theory. In: Linn RL, American Council on Education, editors. Educational measurement. 3rd ed. Washington, DC: American Council on Education; 1989.
9. Gulliksen H. Theory of mental tests. New York, NY: L. Erlbaum Associates; 1987.
10. Traub RE. Reliability for the social sciences: Theory and applications. New York, NY: SAGE; 1994.
11. Lord FM, Novick MR, Birnbaum BA. Statistical theories of mental test scores. Charlotte, US: Information Age Publishing; 2008.
12. Novick MR. The axioms and principal results of classical test theory. Journal of Mathematical Psychology 1966; 3(1): 1-18.
13. DeVellis RF. Classical test theory. Med Care 2006; 44(11 Suppl 3): S50-S59.
14. Cantrell CE. Item response theory: Understanding the one-parameter rasch model. In: Thompson B, editor. Advances in social science methodology. Stamford, CT: Jai Press; 1989. p. 171-92.
15. Predicting Gender Differences in WORD Items. Comparison of item Response Theory and Classical Test Theory. [Online]. 1999 [cited 22 Sep 2012]; Available from URL: http://www.edusci.umu.se/english/swesat/publications/
16. Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. Psychol Assess 2002; 14(1): 50-9.
17. Bock RD. A brief history of item theory response. Educational Measurement: Issues and Practice 1997; 16(4): 21-33.
18. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice 1993; 12(3): 38-47.
19. Ryan JP. Introduction to latent trait analysis and item response theory. In: Hathaway WE, editor. Testing in the schools.San Francisco, CA: Jossey-Bass; 1983. p. 49-65.
20. Partchev I. A visual guide to item response theory. [Online]. 2004 [cited 26 Feb 2004]; Available from URL: http://www.metheval.uni-jena.de/irt/VisualIRT.pdf
21. Baker FB. The basics of item response theory. 2nd ed. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation; 2001.
22. Toit MD. Irt from Ssi: Bilog-mg, multilog, parscale, testfact. Lincolnwood, IL: Scientific Software International; 2003.
23. Habibi M, Izanlo B. Comparison information function of Items and Test in 1, 2, and 3 parametric models of IRT. [Under Review].
24. McDonald RP. Test theory: A unified treatment. New York, NY: Taylor & Francis; 1999.
25. Habibi M, Moradi F, Izanlo B. Invariance of parameters in Item Response Theory (IRT) and confirmatory factor analysis. Journal of Educational Measurement 2012; 2(6): 47-70. [In Persian].
26. Hambleton RK. Traditional and modern approaches to outcomes measurement. [Online]. 2004 [cited 22 Sep 2012]; Available from: URL: http://outcomes.cancer.gov/conference/irt/hambleton.pdf
27. Reise SP, Ainsworth AT, Haviland MG. Item response theory: Fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science 2005; 14(2): 95-101.

28. Hambleton RK, Swaminathan H. Item response theory: Principles and applications. Boston, US: Kluwer-Nijhoff Pub; 1989.
29. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum Associates; 1980.
30. Wiberg M. Classical test theory Vs. Item Response theory: An evaluation of the theory test in the Swedish driving-license test. Umea, Sweden: Department of Educational Measurement, Umea University; 2004.
31. Motiee S, Izanlo B, Peyrovi R. A comparison between the "Classical Test Theory" and "Item ResponseTheory" in term of their ability to measure the test accuracy indice. Higher Education Quarterly 2008; 1(1): 29-47. [In Persian].
32. Embretson SE, Reise SP. Item response theory for psychologists. New York, NY: Routledge; 2000.
33. Hwang DY. Classical test theory and item response theory [microform]: Analytical and empirical comparisons. Washington DC: ERIC Clearinghouse; 2002.
34. Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurem 1998; 58(3): 357-81.

The classical and modern theories of measurement in behavioral science and medicine: a review on the methodological properties, benefits and Limitations

Mojtaba Habibi[1], Ebrahim Khodaei[2], Balal Izanloo[3]

Abstract

The classical test theory (CTT) and item response theory (IRT) are widely used in construction and analysis of exams and questionnaires. Because of the various shortcomings of CTT, IRT has been designed. Even though IRT has overcome most of the limitations of CTT and has made computer adaptive testing possible, it is still important to point out that not enough reasons exist to totally discard CTT. This theory is still in use in different scientific fields. Previous studies have showed that IRT is complementary to CTT, and is not a competitive model. The main difference between these two theories is that CTT is test-based, but IRT is item-based; for example, measurement error modeling in CTT and IRT were done in the test and item level, respectively. In practice, the IRT is shown to have very good performance in dealing with the problems presented in the CTT model; such as conditional standard errors, developing parallel tests, equating, and item bias and adaptive testing. In this study IRT and CTT are explored and the weaknesses and advantages of both theories have been discussed.

Keywords: Classical test theory, Item response theory, Weakness and advantage

Type of article: Original

Received: 09.05.2012 *Accepted: 10.11.2012*

1. Assistant Professor, Family Research Institute, Shahid Beheshti University, Tehran, Iran
2. Assistant Professor, National Organization of Educational Testing, Tehran, Iran
3. PhD Student, Department of Educational Measurement, Tehran University, Tehran, Iran (Corresponding Author)
Email: b.izanloo@yahoo.com