رویآوردهای نوین در روان‌سنجی

قسمت یکم : مبانی و محدودیت‌های نظریه کلاسیک اندازه‌گیری

New Approaches to Psychometrics

Part One : Basis and Limitations of Classic Test Theory

علی عسگری
دانشجوی دکتری
دانشگاه تهران

Ali Asgari
PhD Student
Tehran University

از پیدایش علم روان‌سنجی تقریباً یک قرن می‌گذرد. از آن زمان تاکنون، تغییرات بنیادی زیادی در نظریه‌های آن به وجود آمده‌اند و تداوم این سیر تحولی، دانش روان‌سنجی را به سوی تکامل رانده است. با آنکه تهیه و تفسیر تستها بر پایه نظریه‌های جدید و پیشرفته از فعالیتهای عمده و مهم روان‌شناسان و متخصصان تعلیم و تربیت به‌شمار می‌آید، متأسفانه اصول آنها هنوز برای کاربران و حتی بسیاری از متخصصان اندازه‌گیری در کشور ما به درستی شناخته‌شده نیست و بیشتر کتابهای اندازه‌گیری و سنجش و نیز راهنمای اجرا و تفسیر تستها هنوز بر پایه نظریه کلاسیک تست قرار دارند. بنابراین، تلاش شده است تا با استناد به آثار مختلف ـ در سلسله مقاله‌هایی که از شماره حاضر آغاز می‌شوند ـ به معرفی و توصیف این نظریه‌ها پرداخته شود. برای درک بیشتر نظریه‌های جدید و علل پیدایش آنها، در این مقاله مبانی نظری، مشکلات و محدودیتهای مدلهای کلاسیک اندازه‌گیری توصیف شده است.

مقدمه

با توجه به اینکه آزمونهای روانی و تربیتی در زندگی امروز انسان نقش مهمی ایفا می‌کنند، درک این مطلب که نمره آنها چگونه به دست می‌آید و دربردارنده چه اطلاعاتی هستند، اهمیت فراوانی دارد. برای پاسخ به این گونه پرسشها، دانش تهیه و توسعه آزمون و ابزار اندازه‌گیری که روان‌سنجی نامیده می‌شود گسترش یافته و نظریه‌ها و مدلهای جدید در این حوزه، اصول اندازه‌گیری سازه‌های زیربنایی آزمونها را دگرگون ساخته‌اند. به گونه‌ای که مدلهای جدید اندازه‌گیری را می‌توان برای ارزیابی دامنه گسترده‌ای از حوزه‌های علمی مانند روان‌شناسی، تعلیم و تربیت، اقتصاد و علوم اجتماعی و نیز برای اندازه‌گیری مفاهیم و سازه‌های مختلفی مانند تواناییها، بازخوردها، خصیصه‌ها و رگه‌های شخصیتی به کار برد.

بسیاری از اندازه‌گیریهایی که در این حوزه‌ها انجام می‌شوند شامل یک گروه نمونه معرف است که به یک مجموعه سؤال پاسخ می‌دهند. اغلب این پرسشها نیز به اندازه‌هایی منجر می‌شوند که از طریق جمع نمره‌های مواد آزمون به دست می‌آیند. این نمره‌ها بیشتر به گونه دو ارزشی[1] یا در مقیاسی با طبقه‌های مرتب شده[2] بیان می‌شوند. برای نمونه، اغلب تستهای تشخیصی، بالینی و تربیتی به گونه درست ـ نادرست، موافق ـ مخالف، بلی ـ نه، یا در سطح مقیاسهای چند ارزشی[3] (مانند کاملاً مخالف، مخالف، بی‌تفاوت، موافق و کاملاً موافق) نمره‌گذاری و نمره کل آنها محاسبه می‌شود. نمره کل در واقع

1. dichotomous
2. ordered categories
3. polytomous

پاسخ به همه سؤالها را خلاصه می‌کند. بنابراین، فردی که نمره بالاتری نسبت به دیگران به دست می‌آورد، دارای مقدار بیشتری از متغیر مورد اندازه‌گیری است.

خلاصه‌کردن نمره همه سؤالها برای به دست آوردن نمره کل، که مبنای اصلی نظریه کلاسیک تست[1] را تشکیل می‌دهد، بدین معناست که همه سؤالها برای اندازه‌گیری یک متغیر واحد به یک اندازه نقش دارند. اما به اعتقاد بسیاری از صاحبنظران بین اندازه‌های حقیقی و نمره‌های حاصل از این گونه مقیاسها و اندازه‌گیریها تفاوتهای چشمگیری وجود دارند (رایت و لاینرس، ۱۹۸۹) و نتایج محاسبات حاصل از آنها کاملاً تورش‌دار[2] و ناقص‌اند (والر، ۲۰۰۰).

محدودیتهای اندازه‌گیری کلاسیک

اصطلاح کلاسیک علاوه براینکه به زمان شکل‌گیری مدلهای این نظریه اشاره دارد، به تقابل آن با نظریه‌های جدید روان‌سنجی، یعنی مدلهای نظریه سؤال پاسخ[3] نیز مربوط می‌شود. هدف اصلی نظریه کلاسیک تست درک و بهبود اعتبار آزمونهای روانی است (آلن و یان، ۲۰۰۲). با آنکه ساختار و معادله‌های نظریه کلاسیک تست بیانگر آن است که مفروضه‌های آن در سطح نمره انفرادی آزمودنی تدوین شده‌اند، اما هرگز در مدلهای آن نمره‌های انفرادی تحلیل نمی‌شوند بلکه تمرکز اصلی بیشتر بر ویژگیهای نمره‌های آزمون در ارتباط با گروه نمونه (مجموعه‌ای از افراد) است.

به‌گونه کلی، مفروضه‌ها و اندازه‌هایی که بر پایه نظریه کلاسیک اندازه‌گیری به دست می‌آیند از محدودیتها و مشکلات متعددی برخوردارند. زیرا این نظریه‌ها بر جمع جبری نمره‌ها متکی هستند و پذیرش و عدم پذیرش مفروضه‌های آنها به اراده شخص روان‌سنج بستگی دارد و چون این نمره‌ها خارج از کنترل وی نیستند، نمی‌توانند پیش‌بینی کننده چگونگی پاسخ افراد به سؤالها باشند. با آنکه اندازه‌گیریهای سنتی دشواریها و محدودیتهای متعددی دارند، اما در اینجا به بیان چند مورد اساسی اکتفا می‌شود.

۱) یکی از محدودیتهای عمده مربوط به وجود فواصل برابر در طول مقیاس است. اندازه‌گیری هر خصیصه یا کیفیتی که به گونه فزاینده به مقدار آن افزوده می‌شود، باید بیانگر فواصل برابر در طول پیوستار آن باشد. برای نمونه، در اندازه‌گیری طول، فاصله بین ۳ تا ۴ سانتیمتر، کاملاً با فاصله بین ۱۳ تا ۱۴ سانتیمتر برابر است. اما وقتی افراد به سؤالهای یک آزمون به گونه خود گزارش‌دهی[4] پاسخ می‌دهند به هیچ وجه نمی‌توان اطمینان داشت که فاصله بین نمره‌ها برابر است.

اختصاص اعداد به این گونه خود ارزیابیها[5] که بر پایه عبارتهای کیفی صورت می‌گیرد، در واقع یک جهش مفهومی است که نمی‌توان آن را بر پایه تعاریفی که به نمره‌ها و درجه‌های مختلف داده می‌شود، توجیه کرد (اسمیت، ۲۰۰۱). این گونه نمره‌ها و اعداد، که به هیچ وجه شبیه اندازه‌های مقیاس فاصله‌ای نیستند، در روان‌سنجی از اهمیت بسیار زیادی برخوردارند. زیرا ساده‌ترین عملیات ریاضی نیز به داده‌های فاصله‌ای نیاز دارند که در آن فواصل بین نمره‌های متوالی، برابر باشند. بنابراین فاصله‌ای دانستن داده‌هایی که از طریق مقیاسهای دو ارزشی و چند ارزشی به دست می‌آیند به استنباطهای نادرست درباره آزمودنیها منجر می‌شوند، به ویژه وقتی که داده‌ها از مقایسه گروهی از افراد به دست آمده باشند (مربیتز، موریس و گریپ، ۱۹۸۹؛ رایت و لاینرس، ۱۹۸۹).

1. Classic Test Theory (CTT)
2. biased
3. Item Response Theory (IRT)
4. self-report
5. self-assessment

۲) با تغییر آزمودنیها و انتخاب نمونه دیگری از جامعه، مشخصه‌های‌آماری سؤال تغییر می‌کنند. یک سؤال واحد برای گروه نمونه قوی‌تر ساده‌تر، و برای گروه نمونه ضعیف‌تر دشوارتر خواهد بود. در مدل کلاسیک، شاخص دشواری سؤال علاوه بر خود سؤال، گروه آزمودنی را نیز توصیف می‌کند. شاخص قدرت تشخیص سؤال نیز از گروههای مختلف نمونه تأثیرات متفاوتی می‌پذیرد. بدین معنا که نمونه‌های همگون‌تر ضرایب تشخیص پایین‌تری برای سؤال به دست می‌دهند چون ضرایب دو رشته‌ای و دو رشته‌ای نقطه‌ای نوعی ضریب همبستگی است و همبستگی نیز با افزایش ناهمگونی گروه افزایش می‌یابد (هومن، ۱۳۸۲).

۳) اندازه‌های توصیف‌کننده افراد تحت تأثیر ویژگیهای سؤالها قرار دارند و نمونه‌های متفاوت سؤال به برآوردهای متفاوتی از توانایی افراد مورد سنجش منجر می‌شوند. در مدل کلاسیک، چه نمره مشاهده شده و چه نمره حقیقی (که مقدار مورد انتظار نمره مشاهده شده است)، هر دو به آزمون مورد استفاده وابسته‌اند. افزون بر این، نمره حقیقی از طریق مجموعه مشخصی از سؤالها به دست می‌آید. به بیان‌دیگر، برای هر آزمون تنها یک مقیاس روان‌سنجی[1] وجود دارد. بنابراین، اضافه یا حذف کردن یک سؤال به مقیاس روان‌سنجی متفاوتی منجر خواهد شد (رایس و هنسون، ۲۰۰۳).

۴) مفهوم اعتبار در نظریه کلاسیک در واقع ویژگی ثابت آزمون به حساب نمی‌آید، بلکه ویژگی نمره‌های آن در ارتباط با یک جامعه معین است. زیرا نمره‌های آزمون در هر جامعه اعتبار یکسانی ندارند و اعتبار، مانند هر نوع ضریب همبستگی دیگر، با محدود شدن دامنه نمره‌ها کاهش پیدا می‌کند. بنابراین، نمره یک آزمون هوش که در کل جامعه اعتبار بالایی دارد، ممکن است در جامعه دانشجویان از اعتبار کمی برخوردار باشد. افزون براین، نمره‌های آزمون برای هر آزمودنی نیز کاملاً نامعتبراست. زیرا نمره‌های حقیقی افراد ثابت و واریانس آن در واقع برابر با صفر است. بنابراین، نسبت واریانس نمره حقیقی به واریانس مشاهده شده، یعنی اعتبار، نیز برابر با صفر خواهد بود (آلن و یان، ۲۰۰۲).

۵) در نظریه کلاسیک تست، خطای تصادفی اندازه‌گیری برابر با تفاوت بین نمره حقیقی و نمره مشاهده‌شده‌است. از لحاظ نظری فرض می‌شود همه خطاهای موجود در یک موقعیت سنجش، تصادفی هستند، پس نه‌تنها با یکدیگر همبسته نیستند، بلکه بین خطای تصادفی و نمره‌های حقیقی نیز رابطه نظامدار و پیش‌بینی‌پذیر وجود ندارد. اما این مفروضه با شواهد تجربی همخوانی ندارد. افزون بر این، خطای استاندارد اندازه‌گیری نیز براساس برآورد اعتبار آزمون و نمره‌های مشاهده شده محاسبه می‌شود، در نتیجه مقدار آن برای آزمون و همه افراد یکسان خواهد بود.

۶) یکی از عمده‌ترین محدودیتهای این نظریه مربوط به چگونگی مشارکت مواد آزمون در اندازه‌گیری سازه مورد نظر است. وقتی از یک ابزار برای اندازه‌گیری یک خصیصه یا رگه استفاده می‌شود در حقیقت با این سؤال اساسی روبه‌رو هستیم که آیا در هر یک از واحدهای[2] آن یک نشانگر[3] وجود دارد. برای نمونه، چگونه می‌توان مطمئن بود که مواد آزمون واقعاً معرف خصیصه زیربنایی مورد اندازه‌گیری هستند. بی‌تردید، چنانچه مقیاس دارای سؤالهایی باشد که سازه مورد نظر را به گونه مؤثر اندازه‌گیری نکنند، روایی آن تهدید می‌شود. در نظریه کلاسیک تست، روایی هر مقیاس اساساً بر پایه همپراشی[4] بررسی می‌شود (نانالی، ۱۹۷۸؛ همبلتون و جونز، ۱۹۹۳). به بیان دیگر، بر اساس مقایسه نمره‌های کل حاصل

1. psychometric scale
2. units
3. indicator
4. covariation

از یک اندازه با نمره‌های کل سایر اندازه‌ها، میزان مطابقت روابط مشاهده شده با آنچه انتظار می‌رود مشخص شود (مانند روایی همزمان، پیش‌بین و سازه). با آنکه این روشهای سنتی شواهد مفیدی در باره روایی فراهم می‌آورند، اما نمی‌توانند نشان دهند که سؤالهای یک آزمون معرف سازه مورد اندازه‌گیری است. وقتی نمره‌های کل از جمع سؤالهایی به دست آیند که برازش ضعیفی با سازه مورد نظر دارد روایی آن کاملاً به خطر می‌افتد.

۷) مشکل دیگر مربوط به این است که سؤالهای سازنده یک مقیاس تا چه مقدار معرف و بیانگر سازه مورد اندازه‌گیری است. در هر مقیاس سؤالهایی وجود دارد که نسبت به سایر سؤالها در بردارنده مقدار کمتر یا بیشتری از سازه هدف است. روشهای سنتی با جمع نمره سؤالهای مختلف و پذیرش این فرض که هر سؤال معرف مقدار یکسانی از سازه زیربنایی است، به بروز خطاهای غیرضروری برای دستیابی به نمره کل منجر می‌شوند (مک‌هورنی، هالی و وار، ۱۹۹۷). افزون براین، چون سؤالها معرف مقادیر مختلفی از سازه مورد نظر هستند، باید مناسب بودن سؤالها برای کسانی که مقیاس برای آنها ساخته شده مشخص شود. اگر سؤالها معرف مقدار بسیار زیادی از سازه باشد (بیشتر از آنچه تعداد زیادی از افراد دارند) نمره بسیاری از آزمودنیها در انتهای پایین مقیاس قرار می‌گیرند. این اثر کف[1]، به معنای آن است که افراد دارای مقادیر مختلفی از خصیصه مورد اندازه‌گیری، نمره یکسانی به دست می‌آورند. بالعکس، اگر سؤالها معرف مقدار بسیار کمی از سازه مورد نظر باشند. اثر سقفی[2] به‌وجود آمده موجب می‌شود افرادی که در انتهای پایین مقیاس قرار دارند به گونه مؤثر و کارآمد اندازه‌گیری نشوند.

۸) در نظریه کلاسیک، نمره‌های حقیقی دامنه محدودی دارند و محصور به تعداد سؤالهای آزمون هستند، یعنی حداقل نمره حقیقی صفر و حداکثر معادل بیشترین نمره آزمون است. بنابراین، اگر یک سؤال به مجموعه سؤالها اضافه یا از آن کم شود نمره حقیقی تغییر می‌کند (رایس و هنسون، ۲۰۰۳). این مطلب در مورد مفهوم اندازه‌های موازی، که منطق زیربنایی برآورد اعتبار را تشکیل می‌دهد نیز صادق است. زیرا میزان دقت آزمون نیز با حذف یا اضافه شدن سؤال تغییر خواهد کرد (فلدت و برنان، ۱۹۸۹).

۹) برای تعیین برازش و تناسب مدل کلاسیک با داده‌های حاصل از تست، آزمونهای آماری رسمی وجود ندارد. بنابراین، بر پایه این مدلها نمی‌توان بی‌نظمیهای[3] موجود در داده‌ها را معلوم کرد

با آنکه نظریه کلاسیک تست پر نفوذترین نظریه درباره نمره‌های حقیقی در علوم اجتماعی است، اما در روان‌سنجی و با ورود مدلهای پیچیده‌تر نظریه‌سؤال پاسخ، تقریباً از دور خارج شده است (آلن و یان، ۲۰۰۲). به همین دلیل، بسیاری از پژوهشگران (نقل از هومن، ۱۳۸۲) توصیه کرده‌اند که نظریه‌های جدید باید جایگزین نظریه کلاسیک تست در رشته‌های مختلف علمی شوند.

1. cellar effect
2. ceiling effect

منابع

هومن، حیدرعلی. (۱۳۸۲). *اندازه‌گیریهای روانی‌تربیتی (فن تهیه تست و پرسشنامه)*، (چاپ چهاردهم). تهران: نشر پارسا.

Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.

Feldt, L. S.,& Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., 105–146). New York: American Council on Education and Macmillan.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response, *Educational Measurement: Issues and Practice.; 12 (3)*:38-47.

McHorney, C. A., Haley, S. M., & Ware, J. E. (1997). "Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): 11. Comparison of Relative Precision Using Likert and Rasch Scoring Methods. *Medical Care, 50*, 451-61.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, 70, 308-313.

Nunally, J. C. (1978). *Psychometric Theory. (2nd Ed).* New York: McGraw-Hill.

Reise, S. P., & Henson, M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment , 81 (2)*, 93–103

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A rasch measurement perspective. *Journal of Applied Measurement, 2 (3)*, 281- 311.

Waller, G. N. (2000). *Micro Fact For Windows: Factor Analysis for dichotomous and polytomous response data*. Assessment System Cooperation.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-860.