الگوها و روش‌های سنجش و اندازه‌گیری در برنامه درسی

تاریخ دریافت: ۰۲ /۱۳۹۳/۰۸ تاریخ پذیرش و انتشار: ۱۳۹۴/۰۳/۲۳

دکتر بلال ایزانلو [۱]

مفاهیم اندازه‌گیری، ارزشیابی و سنجش

در منابع ارزشیابی برنامه درسی اغلب به تعریف استیونز[۲] (۱۹۵۱) از اندازه‌گیری[۳] به عنوان اختصاص عدد به ویژگی‌های ذهنی (هوش، توانایی ریاضی و ...) استناد می‌شود. اما این تعریف تنها بر فرایند تجربی اندازه‌گیری تأکید دارد و جنبه‌های نظری آن را نادیده می‌گیرد، یعنی حتی با استفاده از نشانگرهای نامربوط به یک سازه نیز می‌توان به محقق شدن این تعریف دست یافت. این تعریف در واقع مبنایی برای سنجش پایایی[۴] اندازه‌گیری فراهم می‌کند و روایی[۵] آن را نادیده می‌گیرد، چرا که پایایی بر جنبه‌های تجربی استوار است، ولی روایی بر جنبه‌های نظری فرایند اندازه‌گیری و ادغام آنها با جنبه‌های تجربی تأکید دارد. از این رو، می‌توان اندازه‌گیری را به عنوان فرایند ایجاد ارتباط میان مفاهیم انتزاعی با نشانگرهای تجربی تعریف کرد. در این صورت اندازه‌گیری از گام‌های زیر تبعیت می‌کند (الف) تعریف مفهوم یا سازه (ب) انتخاب نشانگرهای بازنمایی‌کننده تجربی سازه مورد (ج) جمع‌آوری اطلاعات تجربی برای نشانگرها (د) ارزیابی میزان بازنمایی نشانگرها از سازه مورد نظر (بلالک[۶]، ۱۹۶۸). بر این اساس، مفهوم سنجش[۷] عبارتست از توصیف منظم عملکرد فرد در یک محیط خاص با استفاده از داده‌های گوناگون کمی و کیفی و سپس تصمیم‌گیری در مورد کارایی فرایند آموزش. با این توصیف، مفهوم ارزشیابی[۸] به فرایند پردازش داده‌های کمی و کیفی حاصل از سنجش برای دستیابی به قضاوت ارزشی درباره میزان موفقیت یا مطلوبیت برنامه درسی اشاره دارد (پین[۹]، ۲۰۰۳).

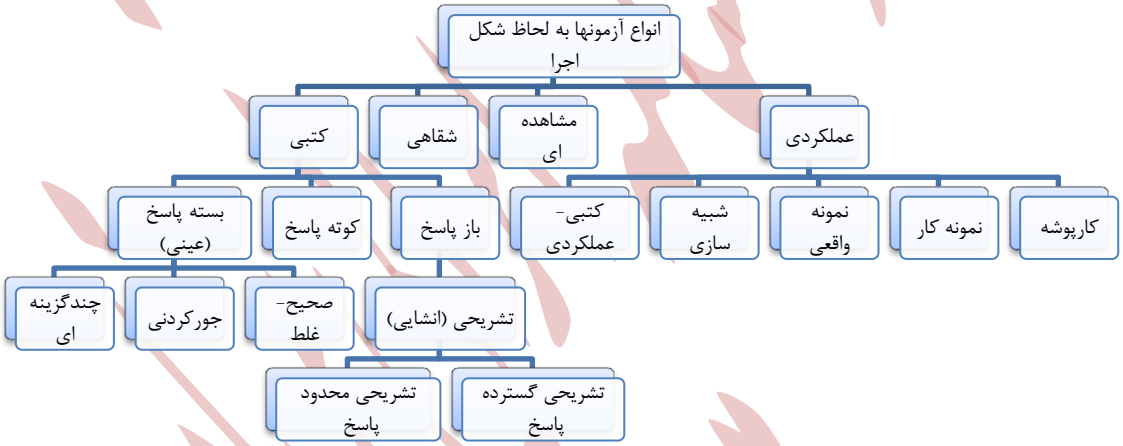
همانطورکه اشاره شد، سنجش از طریق تحلیل داده‌های کمی و کیفی گردآوری شده صورت می‌گیرد. این داده‌ها به وسیله آزمون‌های شفاهی و کتبی (مانند آزمون‌های صحیح-غلط، چندگزینه‌ای، باز پاسخ، محدود پاسخ و کوتاه پاسخ) و آزمون‌های عملکردی (مانند کارپوشه، نمونه‌کار، شبیه‌سازی و نمونه واقعی) و پرسشنامه‌های مختلف سنجش نگرش گردآوری می‌شوند. آزمون‌ها به لحاظ ویژگی مورد اندازه‌گیری نیز در دسته‌های شناختی، عاطفی و مهارتی دسته‌بندی می‌شوند (نک. به شکل‌های ۱ و ۲). اغلب صاحب‌نظران حوزه سنجش از تأکید بر آزمون‌ها و روش‌های سنجش سنتی انتقاد کرده و بر رویکرد سنجش اصیل (مبتنی بر زندگی واقعی)[۱۰] تأکید می‌کنند که با استفاده از آزمون‌های عملکردی میسر می‌شود. در این رویکرد هدف اساسی آموزش و متناسب با آن اندازه‌گیری و سنجش آماده‌کردن فرد برای مواجه با زندگی واقعی است (پین، ۲۰۰۳). آزمون‌ها البته به لحاظ نحوه ساخت (معلم ساخته در برابر استاندارد)، به لحاظ معیار (هنجار-مرجع در برابر ملاک-مرجع) و به لحاظ زمان و هدف (آغازین (ورودی)، تکوینی، تشخیصی و پایانی) هم دسته‌بندی می‌شوند.

---

۱. استادیار دانشگاه خوارزمی b.ezanloo@gmail.com

2. Stevens
3. Measurement
4. Reliability
5. Validity
6. Blalock
7. Assessment
8. Evaluation
9. Payne
10. Authentic Assessment

شکل (۱) طبقه‌بندی انواع آزمون‌ها به لحاظ ویژگی مورد اندازه‌گیری



شکل (۲) طبقه‌بندی انواع آزمون به لحاظ نحوه اجرا

الگوهای اندازه‌گیری و سنجش

اندازه‌گیری صفات ذهنی، دانشی ترکیبی است که نیازمند کسب دانش نسبی در دو حوزه ریاضیات و آمار نیز می‌باشد (گیلفورد[1]، ۱۹۵۴). هر نوع اندازه‌گیری مستلزم تحقق دو نوع عملیات اساسی است: عملیات «رتبه‌بندی[2]» و عملیات «ترکیب[3]» (کمپل[4]، ۱۹۱۷). رتبه‌بندی به معنای آن است که در مقایسه دو شیء از نظر یک ویژگی خاص سرانجام باید مشخص شود که آن ویژگی در کدام یک از دو شیء بیشتر یا کمتر است. مثلاً با مقایسه دو شیء از نظر طول باید بتوان تعیین کرد که کدام یک از دیگری بلندتر است. تحقق این عمل مستلزم حصول اصل «انتقال‌پذیری[5]» است. بر اساس این اصل، از میان سه شیء «الف»، «ب» و «ج»، چنانچه «الف» بزرگتر از «ب» و «ب» هم بزرگتر از «ج» باشد، در نتیجه «الف» از «ج» بزرگتر است. در دنیای فیزیک، تأیید

1. *Guilford*
2. *Ordering*
3. *Combination*
4. *Compel*
5. *Transitivity*

تجربی این اصل تا اندازه‌ای ساده است. به عنوان مثال، برای ترسیم مقیاس رتبه‌بندی مواد از نظر میزان سختی، بر اساس میزان خش انداختن مواد بر روی یکدیگر قضاوت کرده و ماده‌ای را سخت‌تر می‌نامیم که قابلیت ایجاد خش بر روی ماده دیگر را داشته باشد (گیلفورد، ۱۹۵۴). عدم تحقق شرط انتقال‌پذیری ناشی از عدم ثبات ابزار استفاده شده برای اندازه‌گیری و یا ابهام در بُعد مورد مقایسه است که اشیاء بر اساس آن با هم مقایسه می‌شوند. در صورت حصول عملیات رتبه‌بندی به طور ثابت، دومین نوع عملیات، یعنی «ترکیب» نیز باید امکان‌پذیر باشد. یعنی امکان ترکیب دو یا چند شیء با هم بر اساس بُعد اندازه‌گیری شده میسر باشد، به طوری که بتوان آن ترکیب را با هر شیء دیگر یا ترکیب اشیاء در آن بُعد مقایسه نمود.

شرط محقق شدن ترکیب، اصل «جمع‌پذیری[1]» است که به موجب آن عدد مربوط به شیء «الف» به اضافه عدد مربوط به شیء «ب» باید برابر با عدد حاصل از ترکیب آنها، یعنی «الف+ب» باشد. بر خلاف اصل انتقال‌پذیری، تأیید تجربی اصل جمع‌پذیری حتی در دنیای فیزیک هم راحت نیست. با اینکه استفاده از اعداد برای اندازه‌گیری مستلزم حصول دو اصل انتقال‌پذیری و جمع-پذیری است، گزارش فِرگُسِن و همکاران[2] (۱۹۴۰) حاکی از آن بود که در آزمون‌های تربیتی و روان‌شناسی تنها شرط اول (انتقال-پذیری) با استفاده از برخی روش‌های اندازه‌گیری قابل دست‌یابی است. به دنبال این چالش، برخی افراد همچون گالیکسن[3](۱۹۴۶) در جستجوی حل این مسئله برآمدند و بر جمع‌پذیر بودن تفاوت‌های بین فردی و ایجاد مقیاس فاصله‌ای تأکید کردند. در همین راستا استیونز (۱۹۵۱) یک طبقه‌بندی ارائه داد که امکان اجرای تحلیل‌های آماری بر روی داده‌های حاصل از اجرای آزمون‌های ذهنی را بر اساس نوع مقیاس موجه نشان می‌دهد. این طبقه‌بندی مورد قبول همه نیست (از جمله لرد، ۱۹۵۳و بونئو[4]، ۱۹۶۱)، چرا که تلاش دارد یک رابطه دقیق بین سطح اندازه‌گیری متغیرها و استفاده از روش‌های آماری برقرار کند. با این حال تحلیل تفاوت‌های بین فردی و ایجاد مقیاس فاصله‌ای به حل مشکل اصلی منجر نمی‌شود. سرانجام مسئله ترکیب و جمع‌پذیری در اندازه-گیری‌های حوزه علوم تربیتی و روان‌شناسی توسط افرادی مثل لوئیس و توکی[5] (۱۹۶۴) و راش[6] (۱۹۶۰) برطرف شد.

اندازه‌گیری صفات شناختی با ارایه نظریه کلاسیک اندازه‌گیری از سوی اسپیرمن در اواخر قرن نوزدهم و اوایل قرن بیستم آغاز شد. مشخصات عمده این نظریه عبارتند از: ۱) مدلی که این نظریه بر آن استوار است مدلی خطی است (X=T+E) که درآن X،T و E به ترتیب نشان دهنده نمره مشاهده شده[7]، واقعی[8] و خطای[9] یک آزمودنی است که به صورت تصادفی از یک جامعه انتخاب شده است. ۲) بر تکرار شیوه اندازه‌گیری استوار است. یعنی برای برآورد نمره واقعی فرد فرایند اندازه‌گیری باید تکرار شود؛ هرچه تعداد تکرارها بیشتر شود به نمره واقعی نزدیک‌تر می‌شویم. ۳) برای موجه ساختن تکرار اندازه‌گیری در حالت اولیه از تعریف فرم-های همتا[10] استفاده کرد و به تدریج تا فرم‌های اساسا تائو معادل[11] و هم‌جنس[12] پیش رفت (لرد و نویک[13]، ۱۹۶۸). ۴) سپس با

1. *Additivity*
2. *Ferguson*
3. *Gulliksen*
4. *Boneau*
5. *Luce&Tukey*
6. *Rasch*
7. *Observed Score*
8. *True Score*
9. *Error Score*
10. *Parallel Forms*
11. *Essential Tau Equivalent*
12. *Congeneric Forms*
13. *Lord&Novick*

توجه به دشواری‌های همراه با تعریف نمره واقعی و خطا بر اساس تکرار شیوه اندازه‌گیری، این نظریه در تحلیل عاملی کلاسیک که مبتنی بر واریانس مشترک و اختصاصی است ادغام شد (مک‌دونالد[1]، ۱۹۸۱) و در نهایت با تعریف فرم‌های هم‌جنس، نظریه کلاسیک به مدل‌یابی معادله ساختاری[2] پیوست ۵) بر گشتاورهای مرتبه اول (میانگین) و دوم (واریانس) در توزیع نمره‌های مشاهده شده استوار است، بدون آنکه برای نمره‌های مشاهده شده توزیعی مشخص کند. ۶) مهم‌ترین نتایج کاربردی این نظریه پایایی، روایی و شاخص‌های دشواری و تشخیص[3] پرسش است. ۷) معمولاً برای آزمون‌های هنجار مرجع[4] کاربرد دارد ۸) مهم‌ترین نقطه ضعف‌های آن این است که پارامتری پرسش و فرد در این نظریه وابسته به نمونه است. به خصوص زمانی که بین پرسش‌ها و توانایی نمونه ناهمخوانی زیادی وجود داشته باشد (مثلاً وقتی پرسش‌های دشوار برای یک نمونه دارای نمره واقعی پایین استفاده شوند یا برعکس). به علاوه برای تمام سطوح توانایی یک خطای استاندارد اندازه‌گیری[5] برآورد می‌شود (لرد و نویک، ۱۹۶۸) (برای مرور بیشتر نظریه کلاسیک نک. به آلن و ین[6]، ۱۹۷۹؛ همچنین برای مرور بیشتر مدل‌یابی معادله ساختاری نک. به ریکاف و مارکولایدز، ۲۰۰۶).

با توجه به نقطه ضعف‌های مختلف نظریه کلاسیک نظریه‌های گوناگونی شکل گرفت. به عنوان مثال، برای لحاظ کردن سهم منابع مختلف خطا در میزان پایایی نمره‌ها نظریه تعمیم‌پذیری[7] شکل گرفت. با استفاده از این نظریه می‌توان سهم وجوه[8] مختلف در اجرای فرایند اندازه‌گیری را مشخص و براساس نتایج حاصل برای کاهش منابع مختلف خطا فرایند اندازه‌گیری را بهتر پی‌ریزی کرد (کرانباخ، راجراتنام و گلاسر[9]، ۱۹۶۳) (برای مرور بیشتر نک. به شی‌ولسُن و وب[10]، ۱۹۹۱).

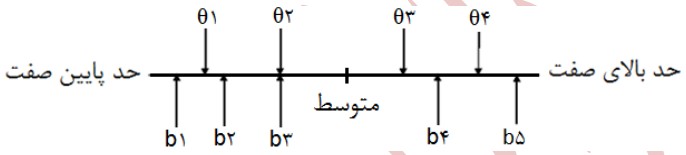
همان طور که ذکر شد چون نظریه کلاسیک بر اساس تفاوت‌های بین فردی پایه‌ریزی شده فقط برای آزمون‌های هنجار مرجع کاربرد دارد. برای جبران این کاستی نظریه تصمیم‌گیری طبقه‌بندی[11] شکل گرفت که علاوه بر سایر کاربردهای آن می‌توان از آن برای بررسی ویژگی‌های روان‌سنجی آزمون‌های ملاک مرجع[12] نیز استفاده کرد (کتس و لرد[13]، ۱۹۶۲؛ لرد، ۱۹۶۵). بر خلاف نظریه کلاسیک، این نظریه برای نمره‌های مشاهده شده افراد توزیع مشخص می‌کند و به همین خاطر به نظریه قوی نمره واقعی[14] نیز مشهور است (برای مرور بیشتر نک. به آلن و ین، ۱۹۷۹).

نظریه پاسخ پرسش[15] با تأکید بر مدل‌سازی احتمال پاسخ درست به پرسش‌ها بر اساس میزان توانایی برخی از آرمان‌های سنجش و اندازه‌گیری از جمله تهیه بانک پرسش[16] و سنجش انطباقی[17] را تحقق بخشیده است (امبرستون و رایز، ۲۰۰۰). برای توصیف

---

1. *Mcdonald*
2. *Structural Equation Modeling*
3. *Difficulty And Discrimination*
4. *Norm Reference*
5. *Standard Error Of Measurement*
6. *Allen & Yen*
7. *Generalized Theory*
8. *Facet*
9. *Cronbach, Rajaratnum&Gleser*
10. *Shavelson& Webb*
11. *Decision Classification*
12. *Criterion Reference*
13. *Keats& Lord*
14. *Strong True Score Theory*
15. *Item Response Theory*
16. *Item Bank*
17. *Adaptive Testing*

رویکرد نظریه پاسخ پرسش از رویکرد راش[1] استفاده می‌کنیم. متغیرهای مورد علاقه پژوهش‌گران حوزه برنامه‌ریزی درسی را می‌توان به وسیله آموزش تحت تأثیر قرار داد تا جایگاه[2] دانش‌آموزان در صفت پنهان مربوطه ارتقاء یابد. تحت پیش فرض بنیادی تک‌بعدی بودن[3]، متغیر آموزشی را می‌توان به عنوان یک صفت پنهان در نظر گرفت که هم دانش‌آموزان و هم پرسش‌ها (یا فعالیت‌های موجود در آزمون) را می‌توان بر روی پیوستار آن به ترتیب جایگاه نشان داد (شکل شماره ۳ را ببینید). در این شکل خط افقی نشانگر صفت پنهان و توانایی i امین آزمودنی ($i=1,2, \ldots, n$) با علامت θ_i بر روی این خط نشان داده شده است. مثلا θ_1 و θ_2 به ترتیب توانایی اولین و دومین آزمودنی در مقیاس صفت پنهان هستند. در پایین خط افقی دشواری j امین پرسش ($j=1,2, \ldots, k$) با علامت b_j نشان داده شده است. مثلا b_1 و b_2 به ترتیب دشواری اولین و دومین پرسش هستند.



شکل (۳) صفت پنهان فرضی

انتخاب صفر به عنوان مبدا و مقدار متوسط با توجه به مقیاس صفت پنهان بر روی توزیع طبیعی ارائه شده است. پیش فرض بنیادی تک‌بعدی بودن سه فرضیه آزمون‌پذیر را فراهم می‌کند. **اول**، تنها عامل اثر بخش بر موقعیت آزمودنی‌ها در صفت مشاهده-ناپذیر، توانایی مورد اندازه‌گیری. **دوم**، تنها دشواری پرسش‌ها بر موقعیت آنها در مقیاس صفت پنهان تاثیر می‌گذارد. **سوم**، خصیصه‌ای که آزمودنی‌ها بر اساس توانایی و پرسش‌ها بر اساس دشواری برروی آن مرتب شده‌اند مشابه و مقیاس هر دو برروی توزیع طبیعی استاندارد با میانگین صفر و واریانس یک قرار دارد. برآوردهای توانایی و دشواری پرسش بر روی این توزیع قرار دارد که در عمل مقادیر آنها در دامنه ۴± قرار دارند (لرد ۱۹۸۰؛ امبرستون و رایز ، ۲۰۰۰). تحت این شرایط بحث در مورد احتمال این که یک آزمودنی با توانایی معین به پرسشی با دشواری مشخص جواب صحیح دهد به لحاظ نظری و عملی بسیار آگاهی‌دهنده است. مثلاً، زمانی که یک آزمودنی با توانائی θ_2 به پرسشی با دشواری b_2 جواب دهد، محتمل‌ترین نتیجه جواب درست است (شکل ۳ را ببینید). زمانی که یک آزمودنی با توانائی θ_2 به یک پرسش با دشواری b_4 جواب دهد، محتمل‌ترین نتیجه جواب نادرست است. سرانجام اگر یک آزمودنی با توانایی θ_2 به پرسشی با دشواری b_3 جواب دهد، محتمل‌ترین نتیجه ۰/۵ است. یکی از راه‌های رسیدن به یک الگوی احتمالی استفاده از تساوی $\frac{P_{ij}}{1-P_{ij}} = \theta_i - b_j$ است. بر اساس این رابطه تفاوت توانایی فرد i از دشواری پرسش j برابر با نسبت پاسخ درست به پرسش P_{ij} بر پاسخ نادرست به آن $1 - P_{ij}$است (P_{ij} نسبت کلیه افراد دارای توانایی مساوی با θ_i است که به پرسش j پاسخ درست داده‌اند). عبارت سمت چپ تناسب فوق را شانس[4] می‌نامند. شانس به عنوان شاخص احتمال وقوع یک حادثه با مشکل ناقرینگی[5] همراه است. در حالی که دامنه شانس بین صفر تا مثبت بی‌نهایت است

1. *Rasch*
2. *Location*
3. *Unidimensionality*
4. *Odds*
5. *Asymmetry*

تفاوت بین توانایی و دشواری سؤال در دامنه $\pm\infty$ در نوسان است. ناقرینگی موجود در شانس با گرفتن لگاریتم طبیعی آن برطرف شده و مقیاس جدید آن با عنوان لگاریتم بخت (لُجیت[1]) به دست می‌آید. به این ترتیب فرم کلی معادله به صورت $ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \theta_i - b_j$ تبدیل می‌شود. با کمی دستکاری جبری می‌توان آخرین رابطه را به صورت زیر نوشت، که احتمال پاسخ درست توسط فرد i با توانایی θ_i به پرسش j با دشواری b_j را حساب می‌کند.

$$P_{ij}(X = 1|\theta_i, b_j) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}}$$

مدل‌سازی بر اساس نظریه پاسخ پرسش از دو رویکرد تبعیت می‌کند. رویکرد اول آن است که مناسب‌ترین مدل را برای برآورد پارامترهای پرسش و صفت مورد نظر به کار گرفته شود. فلسفه این رویکرد آن است که پرسش‌ها همان طور که اجرا شده‌اند اندازه-گیری می‌کنند، نه آن طور که باید عمل می‌کردند. دومین رویکرد برای مدل‌سازی بر اساس نظریه پاسخ پرسش این است که داده-ها باید با ویژگی‌های خاص اندازه‌گیری که به وسیله مدل تعریف شده‌اند متناسب باشند. اگر پرسش یا فردی با ویژگی‌های مدل اندازه‌گیری برازش نداشته باشند کنار گذاشته می‌شوند. این گروه با پیروی از رویکرد راش معتقدند که تنها مدل‌های مناسب برای اندازه‌گیری‌های ذهنی مدل‌های خانواده راش هستند، چون ویژگی‌های قوی ریاضی نظیر عینیت خاص[2] (یعنی، مستقل بودن مقایسه افراد از سوال‌ها و مقایسه پرسش‌ها از افراد) و کفایت ساده نمره کل (یعنی، عدم نیاز به اطلاع داشتن از الگوی پاسخ) را دارا هستند. به عنوان مثال، مدل تک پارامتری راش دارای چندین مزیت است. زمانی که تعداد پاسخ دهندها کم است یا نمونه‌های فوق‌العاده نامعرف مورد استفاده قرار می‌گیرند و یا توزیع جامعه در خصیصه زیربنایی مورد نظر به شدت دارای کجی باشد، این مدل از توانائی کافی برای برآوردهای پایدارتری از پارامترهای افراد و ویژگی‌های پرسش‌ها برخوردار است (تیسن و اورلند[3]،۲۰۰۱). به نظر امبرستون و رایز (۲۰۰۰) زمانی که در تعریف متغیر زیربنایی، همه پرسش‌ها وزن یکسانی دارند و زمانی که ویژگی‌های قوی مدل اندازه‌گیری (یعنی عینیت خاص و کفایت ساده نمره کل) مورد نظر است، باید از مدل‌های خانواده راش استفاده کرد. اما اگر برازش یک مدل با داده‌های موجود مورد نظر باشد یا برآوردهای دقیق‌تری از پارامترها برای شما مطلوب است، باید از مدل‌های پیچیده‌تر نظیر مدل دو پارامتری استفاده کرد (برای مرور بیشتر نک. به امبرستون و رایز، ۲۰۰۰).

جمع‌بندی و نتیجه‌گیری

سنجش در حیطه مسائل برنامه درسی از دو جنبه قابل بررسی است. از یک نظر سنجش به عنوان ابزاری در جهت ارتقا یادگیرندگان در نظر گرفته می‌شود که تأکید عمده آن بر یادآوری و بازخوانی اطلاعات از حافظه است. در حالی که از جنبه دیگر سنجش به عنوان ابزاری در جهت کمک به رشد یادگیرنده و معلم در نظر گرفته می‌شود که تأکید عمده آن بر آماده‌سازی یادگیرندگان برای زندگی در دنیای واقعی و رویارو نمودن آنها با تکالیفی شبیه به تکالیف زندگی واقعی است. رویکرد دوم که به سنجش اصیل معروف است استفاده از آزمون‌های عینی (مثل چندگزینه‌ای یا کوتاه پاسخ) را به عنوان تنها وسیله اطلاع از یادگیری یادگیرندگان مورد نقد قرار داده و پرورش تفکر و حل مساله دانش‌آموزان را منوط به بهره‌گیری از رویه‌های سنجش اصیل

1. *Logit*
2. *Specific Objectivity*
3. *Thissen& Orlando*

همچون آزمون‌های عملکردی، کارپوشه، خودسنجی و آزمون‌های مبتنی بر تولید پاسخ می‌داند. در بین نظریه‌های اندازه‌گیری ذکر شده در قبل نیز مدل‌های نظریه پاسخ پرسش به طور گسترده مورد استفاده قرار می‌گیرند. برخی از مهم‌ترین مشکلات پیش‌روی این مدل‌ها عبارتند از: ۱) احتیاج به حجم نمونه بالایی دارند، به طوری که کاربرد آنها در همه شرایط به خصوص موقعیت‌های کلاسی ممکن نیست ۲) برازش مدل‌ها با داده‌های مشاهده شده به راحتی به دست نمی‌آید ۳) نتایج حاصل از این مدل‌ها منوط به برقراری پیش فرض‌های آنها (از جمله تک‌بُعدی بودن) در داده‌ها است. از این رو تایید واقع شدن این پیش فرض‌ها، کاربرد نتایج آنها را محدود می‌سازد.

منابع

سوزان ای. امبرتسون. و استیون پی . رایس. (۲۰۰۰). *نظریه‌های جدید روان‌سنجی برای روان‌شناسان* (به انضمام نرم افزارهای تحلیل داده‌ها). ترجمه حسن پاشاشریفی و همکاران. (۱۳۸۸). تهران: انتشارات رشد.

آلن، ام. جی. و ین، دابلیو، ام. (۱۹۷۹). *مقدمه‌ای بر نظریه‌های اندازه‌گیری (روان‌سنجی)*. ترجمه علی دلاور (۱۳۷۴). تهران: انتشارات سمت.

ریکاف، تنکو. و جرج، ای، مارکولایدز. (۲۰۰۶). *مبانی مدل‌یابی معادلات ساختاری*. ترجمه بلال ایزانلو، محسن دهقانی و مجتبی حبیبی. (۱۳۹۳). تهران: انتشارات رشد.

Blalock, H.M. (1968). The measurement problem. In H. M. Blalock & A. B. Blalock (Eds.), *Methodology in social research*. New York: McGraw-Hill.

Campbell, N.R. (1917). *Foundations of Science: The Philosophy of Theory and Experiment*. New York: Dover Publications Inc.

Cronbach, L.J., Rajaratnum, N. & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137–163.

Ferguson, A., Myers, C.S. & Bartlett, R.J. (1940). Quantitative estimation of sensory events. *Advances of Science*, *2*, 331-349.

Guilford, J.P. (1954). *Psychometric methods*. McGraw-Hill.

Gulliksen, H. (1946). Paired comparisons and the logic of measurement. *Psychological Review*, *53*(4), 199.

Keats, J.A., & Lord, F.M. (1962). A theoretical distribution for mental test scores.*Psychometrika*, *27*(1), 59–72.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *Vol 8*(12), 750-751.

Lord, F.M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*(3), 239–270.

Lord, F.M. & Novick, M.R. (1968). *statistical theories of mental test scores*. Addison Wesley.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, *1*(1), 1-27.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of mathematical and statistical psychology*, *34*(1), 100-117.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Danish Institute for Educational Research.

Shavelson, R.J. & Webb, N. (1991).*Generalizability Theory: A primer*. Thousand Oaks, CA: Sage.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics.In S. S. Stevens (Ed.), *Handbook of experimental psychology*.

Tenko, R. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*: Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Tissen & H. Wainer (Eds.), *test scoring*. Erlbaum.