مقاله مروری

مروری بر نظریه تعمیم‌پذیری و کاربردهای آن در آموزش پزشکی

سارا مرتاض‌هجری[1]، لیلا جانانی[2]، محمد جلیلی[3*]

1- گروه آموزش پزشکی، دانشگاه علوم پزشکی تهران 2- گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی ایران 3- گروه طب اورژانس، گروه آموزش پزشکی، دانشگاه علوم پزشکی تهران

* نویسنده مسؤول: تهران، بلوار کشاورز، خیابان نادری، خیابان حجت دوست، پلاک 57، طبقه سوم، گروه آموزش پزشکی. تلفن: 02188955846 نمابر: 021889550160

پست الکترونیک: mjalili@tums.ac.ir

دریافت: 94/1/17 پذیرش: 94/3/31

چکیده

مقدمه: *تمام اندازه‌گیری‌ها از جمله ارزیابی دانشجویان پزشکی در ذات خود واجد خطا هستند. با توجه به ضرورت حفظ کیفیت ارزیابی در علوم پزشکی، یافتن روش‌هایی برای برآورد میزان خطا و کاهش آن حائز اهمیت است. طی سال‌های اخیر، پیشرفت‌های چشمگیری در زمینه نظریه‌های اندازه‌گیری صورت گرفته است که در این راه کمک‌کننده است.*

روش کار: *در این مقاله، پس از مروری بر نظریه کلاسیک آزمون و محدودیت‌های آن، با استفاده از مثال‌هایی در خصوص آزمون عینی ساختارمند بالینی، نظریه تعمیم‌پذیری را بررسی می‌کنیم.*

یافته‌ها: *طبق نظریه کلاسیک آزمون، نمره‌ای که دانشجو در امتحان کسب کرده است، معادل نمره واقعی او نیست و با خطا همراه است. با تعیین پایایی آزمون می‌توانیم بگوییم میزان تأثیر خطای تصادفی در نمره چقدر بوده است اما در این نظریه نمی‌توان در آن واحد، تمام منابع خطا را در نظر گرفت. برای رفع این محدودیت از نظریه تعمیم‌پذیری استفاده می‌شود که بر اساس مدل‌های آماری آنالیز واریانس، سهم منابع مختلف خطا را در اندازه‌گیری مشخص می‌کند و علاوه بر برآورد پایایی آزمون، پیشنهادهایی برای بهبود تعمیم‌پذیری نتایج ارایه می‌دهد.*

نتیجه‌گیری: *در نظریه کلاسیک آزمون، شناسایی میزان اثر منابع مختلف خطاهای احتمالی به تفکیک امکان‌پذیر نیست. در حالی که با استفاده از نظریه تعمیم‌پذیری می‌توان سهم هریک از منابع خطا را شناسایی نمود و برای آزمونی پایاتر برنامه‌ریزی کرد.*

گل‌واژگان: *اندازه‌گیری، نظریه کلاسیک آزمون، نظریه تعمیم‌پذیری، پایایی*

مقدمه

تمام شاخه‌های علم با هدف جمع‌آوری اطلاعات در مورد یک پدیده خاص نیازمند اندازه‌گیری هستند و تمام اندازه‌گیری‌ها در ذات خود واجد خطا هستند و تحت تأثیر شرایط گوناگون، مقادیر مختلفی به دست می‌دهند (1). خطاهای اندازه‌گیری به صورت کلی به دو نوع خطای سیستماتیک و تصادفی تقسیم می‌شوند. خطای سیستماتیک که نام دیگر آن سوگرایی[1] است، به طور مستمر، منظم و مکرر مقادیر را تحت تأثیر قرار می‌دهد و معمولاً الگو و جهت خاصی دارد. در حالی که خطای تصادفی کاملاً شانسی و بدون الگو و جهت مشخص، اندازه‌ها را تحت تأثیر قرار می‌دهد و میانگین آن در تکرار زیاد اندازه‌گیری‌ها نزدیک به صفر است. خطای سیستماتیک را می‌توان با شناسایی الگوهای مشخص آن حذف کرد و خطای تصادفی را می‌توان با تکرار دفعات اندازه‌گیری کاهش داد (2). آزمون‌های تحصیلی نیز که به نوعی دانش یا مهارت فراگیران را اندازه‌گیری می‌کنند، از این امر مستثنا نیستند و منابع مختلفی ممکن است باعث بروز خطا

[1] Bias

در آن‌ها شود. یکی از روش‌های ارزیابی فراگیران در علوم پزشکی، آزمون عینی ساختارمند بالینی[2] است. در این آزمون، دانشجو به ترتیب وارد ایستگاه‌های مختلف می‌شود و در هر یک از آن‌ها در زمان مشخصی یک وظیفه بالینی را در مواجهه با بیمارنما یا مولاژ انجام می‌دهد و آزمون‌گر مستقر در ایستگاه با مشاهده مستقیم عملکرد دانشجو بر اساس چک‌لیست‌های از پیش تدوین شده به ارزیابی وی می پردازد (3). تأثیر هر عاملی به غیر از سطح علمی دانشجویان روی نمرات OSCE به عنوان خطای اندازه‌گیری محسوب می‌شود. به عنوان مثالی از خطای سیستماتیک در این آزمون، آزمونگری است که سخت‌گیر است و احتمالاً برای همه دانشجویان سخت‌گیر است و نمره تمام آن‌ها را کمتر از حد واقعی برآورد می‌کند. در اینجا آموزش دادن و توجیه وی می‌تواند کمک‌کننده باشد. در حالی که برای کاهش خطای تصادفی OSCE، افزایش تعداد آزمونگران یا افزایش تعداد ایستگاه‌های آزمون کمک‌کننده است.

با توجه به ضرورت حفظ کیفیت روش‌های ارزیابی در علوم پزشکی و اطمینان از عملکرد قابل قبول دانشجویان و فارغ‌التحصیلان، یافتن روش‌هایی برای برآورد میزان خطا و کاهش آن در اندازه‌گیری حائز اهمیت است (4). به موازات تغییرات صورت گرفته در سایر علوم، پیشرفت‌های چشمگیری در زمینه نظریه‌های اندازه‌گیری[3] صورت گرفته است که در ساخت و تحلیل آزمون‌ها و پرسشنامه‌ها کاربرد زیادی دارند. اولین نظریه منسجم در این زمینه تحت عنوان نظریه کلاسیک آزمون[4] در اوایل قرن بیستم شکل گرفت (5) و توسعه ابزارهای ارزیابی طبق اصول آن به سرعت گسترش یافت. در عین حال، برخی مطالعات، محدودیت‌هایی را در این نظریه نشان دادند و زمینه‌ساز ظهور نظریه‌های جدید از جمله نظریه تعمیم‌پذیری[5] شدند که امکان شناسایی میزان خطا را در ابعاد مختلف اندازه‌گیری فراهم می‌سازد (1 و 2).

در این مقاله قصد داریم به صورت خلاصه مروری بر این دو نظریه اندازه‌گیری انجام دهیم. با توجه به منابع متعدد خطا در OSCE، در این مقاله با ذکر نمونه‌هایی از این آزمون، ابتدا به بیان ویژگی‌ها و محدودیت‌های نظریه کلاسیک می‌پردازیم و سپس نظریه تعمیم‌پذیری را مورد بررسی قرار می‌دهیم.

نظریه کلاسیک آزمون: اساس و پایه نظریه کلاسیک بر این فرض استوار است که نمره مشاهده شده[6] (X) دانشجو از دو جزء یعنی نمره واقعی[7] وی (T) و خطای اندازه‌گیری[8] (E) تشکیل شده است. به عبارت دیگر نمره‌ای که دانشجو در امتحان کسب کرده است، معادل نمره واقعی او نیست و با درجاتی از خطا همراه است (1 و 5). این مسأله به این صورت قابل نمایش است: $X= T+ E$

بر اساس معادله فوق، خطای اندازه‌گیری حاصل اختلاف بین نمره مشاهده شده و نمره واقعی فرد است. مفهوم نمره واقعی در معادله فوق، یک مفهوم فرضی است که از میانگین نمرات مشاهده شده در بی‌نهایت تکرار اندازه‌گیری‌ها به دست می‌آید. فرمول زیر در خصوص واریانس این نمرات صدق می‌کند:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

در واقع تعریف رایج پایایی، یعنی تعیین میزان ثبات در اندازه‌های به دست آمده، از همین مفهوم گرفته شده است. مفهوم پایایی به این امر دلالت دارد که با تکرار آزمون می‌توان امیدوار بود که به نمره واقعی فراگیر نزدیک شویم. به عبارت دیگر، هر چند هرگز نمی‌توانیم نمره واقعی دانشجو را دقیقاً به دست آوریم، با تعیین پایایی آزمون می‌توانیم بگوییم میزان تأثیر خطای تصادفی در نمره او چقدر بوده است و سپس با تخمین خطای معیار اندازه‌گیری[9] می‌توانیم محدوده نمره واقعی دانشجو را برآورد کنیم (1). ضریب پایایی در نظریه کلاسیک، نسبت واریانس نمره واقعی به واریانس نمره مشاهده شده (به شرط مستقل بودن خطا و نمره مشاهده شده) است و به صورت فرمول زیر نمایش داده می‌شود:

$$r = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

روش های آماری تعیین پایایی متنوع هستند و هر چند اصول کلی آن‌ها در حوزه ارزیابی فراگیر همانند سایر حیطه‌های اندازه‌گیری است اما به علت شرایط خاص حاکم بر آزمون‌ها برخی از روش‌ها کمتر قابل استفاده هستند. به صورت مشخص‌تر دو روش بازآزمایی[10] و آزمون‌های هم‌ارز[11] علی‌رغم استفاده زیاد در تعیین پایایی پرسشنامه‌ها، کاربرد کمتری در

[2] Objective Structured clinical examination (OSCE)
[3] Measurement Theories
[4] Classical Test Theory (CTT)
[5] Generalizability Theory

[6] Observed score
[7] True score
[8] Measurement error
[9] Standard Error of Measurement (SEM)
[10] Test-retest
[11] Parallel tests

حوزه آموزش دارند. در مقابل، روش‌هایی که مبتنی بر یک بار اجرای آزمون هستند و تحت عنوان همسانی درونی[12] شناخته می‌شوند، مانند روش‌های دو نیمه کردن[13]، کودر-ریچاردسون[14] و آلفای کرونباخ برای سنجش پایایی ابزارهای ارزیابی فراگیر بیشتر استفاده می‌شوند (4). اگرچه نظریه کلاسیک به دلیل سادگی مفاهیم ابتدایی، قابلیت اجرای بالا و داشتن نرم‌افزارهای آماری متنوع، متداول‌ترین نظریه اندازه‌گیری مورد استفاده است، محدویت‌هایی نیز برای آن ذکر شده است. همان‌طور که قبلاً ذکر شد، منابع متنوع خطا مانند خود سؤالات، شرایط امتحان و آزمونگر با دور کردن نمرات مشاهده شده از نمرات واقعی دانشجویان، بر پایایی آزمون تأثیر می‌گذارند. نکته اینجاست که برای محاسبه پایایی آزمون در نظریه کلاسیک نمی‌توان در آن واحد تمام منابع خطا را در نظر گرفت و در هر لحظه تنها اثر یکی از این منابع قابل بررسی و برآورد است. در واقع نظریه تعمیم‌پذیری برای رفع این محدودیت ارایه شد و برتری اصلی آن نسبت به نظریه کلاسیک این است که سهم منابع مختلف خطا را در اندازه‌گیری مشخص می‌کند (1).

نظریه تعمیم‌پذیری: با توجه به محدودیت نظریه کلاسیک، نظریه تعمیم‌پذیری، توسط کرونباخ[15] معرفی شد و بعدتر توسط برنان[16] گسترش یافت (1 و 6). این نظریه که عموماً به عنوان جی-تئوری[17] شناخته می‌شود، در واقع بسط نظریه کلاسیک است و در شناخت منابع خطا بسیار کارایی دارد (7). در نظریه کلاسیک می‌توان تنها اثر یک منبع خطا را مشخص نمود اما در نظریه تعمیم‌پذیری، می‌توان منابع مختلف خطا را شناسایی نمود و مقدار اثر هر یک از آن‌ها را برآورد کرد تا تصویر واضح‌تری از خطای اندازه‌گیری به دست آید و تفسیر دقیق‌تری از نمرات قابل ارایه باشد. به این ترتیب متولیان آزمون قادر خواهند بود منابع اصلی خطا را شناسایی کنند و برای برگزاری آزمونی پایاتر برنامه‌ریزی کنند (1 و 8).

پایه محاسبات در نظریه تعمیم‌پذیری بر اساس مدل‌های آماری آنالیز واریانس[18] است که هرچند از طریق نرم‌افزارهایی مانند SPSS قابل اجرا است اما با توجه به دشواری‌های آن، مخصوصاً هنگامی که تعداد منابع خطا زیاد است، برنامه‌های اختصاصی نظریه تعمیم‌پذیری نیز تدوین شده است که از جمله آن‌ها می‌توان به GENOVA، EduG و G string اشاره کرد (7 و 9). در ابتدا به مرور چند مفهوم پایه‌ای این نظریه می‌پردازیم و سپس کاربردهای عملی آن را با ذکر چند مثال تبیین می‌کنیم:

رویه و انواع آن: نمره دانشجو تحت تأثیر خصوصیاتی از آزمون است که در نظریه تعمیم‌پذیری رویه[19] نامیده می‌شوند و هر یک دارای سطوح مشخصی هستند. به عنوان مثال، در آزمون OSCE، ایستگاه‌ها، آیتم‌های چک‌لیست‌ها، عملکرد بیمارنماها، عملکرد ارزیابان و... هر یک می‌توانند با ایجاد خطای اندازه‌گیری موجب شوند دانشجو نمره‌ای غیر از نمره واقعی خود کسب کند. رویه‌ای که مربوط به خود مورد اندازه‌گیری[20] است، مانند دانشجو در مثال آزمون OSCE، رویه تمیز[21] نامیده می‌شود زیرا تفاوتی که بین نمرات دانشجویان مختلف دیده می‌شود، ناشی از خطا نیست و منعکس‌کننده تفاوت واقعی در سطح آن‌ها است. به سایر رویه‌ها، مانند ایستگاه یا آزمونگر، رویه تعمیم[22] گفته می‌شود (7 و 8). کار نظریه تعمیم‌پذیری این است که مشخص کند نمره دانشجو تا چه حد به سایر ایستگاه‌ها یا آزمون‌گرانی که به صورت بالقوه می‌توانستند وجود داشته باشند، قابل تعمیم است. هر چه واریانس ناشی از دانشجویان بیشتر باشد، پایایی آزمون بیشتر خواهد بود و هر چه سهم رویه‌های تعمیم در پراکندگی نمرات بیشتر باشد، پایایی آزمون کمتر است (4).

جهان و نمره جهانی: در نظریه تعمیم‌پذیری به جای نمره واقعی از لفظ «نمره جهانی»[23]، برای مجموعه شرایط اندازه‌گیری از اصطلاح «جهان»[24] و برای کل مجموع افراد مورد اندازه‌گیری از اصطلاح «جمعیت»[25] استفاده می‌شود. پایه و اساس اِدراکی[26] نظریه تعمیم‌پذیری را «جهان مشاهدات قابل‌قبول»[27] تشکیل می‌دهد که طبق تعریف شامل کل مشاهدات ممکنی است که برای اندازه‌گیری در شرایط خاص از طرف محقق قابل‌قبول است (8). به عنوان مثال، به صورت بالقوه برای انجام OSCE، 50 ایستگاه قابل طراحی است که در حال حاضر تنها 10 ایستگاه از بین آن‌ها انتخاب شده است. اگر نمونه‌های دیگری انتخاب شده بودند، احتمالاً نمرات دانشجویان

[12] Internal consistency
[13] Split-halves
[14] Kuder-Richardson
[15] Cronbach
[16] Brennan
[17] G theory
[18] ANOVA

[19] Facet
[20] Object of Measurement
[21] Facet of differentiation
[22] Facet of generalization
[23] Universe Score
[24] Universe
[25] Population
[26] Conceptual
[27] Universe of admissible observations

متفاوت بود. بنابراین، خطایی در آزمون وجود دارد که ناشی از ایستگاه است و مقدار آن باید برآورد شود و مشخص شود نمره‌ای که دانشجو در این آزمون 10 ایستگاهی کسب کرده است، تا چه حد قابل تعمیم به تمام 50 ایستگاه فرضی است. آن حالت بالقوه، «جهان تعمیم»[28] نام دارد و به حالت عملی شده، جهان مشاهدات قابل قبول گفته می‌شود. میانگین نمره دانشجو از تمام حالات جهان تعمیم همان نمره جهانی وی است. در برخی از موارد این دو جهان با هم یکسان هستند. مثلاً هنگامی که در یک آزمون از پنج آزمونگر استفاده می‌شود و کل تعداد استادان گروه مربوطه هم پنج نفر است. در اینجا آن چه ممکن و عملی شده بر آن چه تعمیم می‌دهیم، منطبق است. تعیین این موضوع در استفاده از آنالیز مناسب اهمیت دارد. به این ترتیب که اگر تعداد رویه مورد نظر در جهان ممکن و تعمیم منطبق بر هم باشند (مثلاً همان پنج آزمونگر)، از مدل «اثرات ثابت»[29] استفاده می‌شود و اگر مانند مثال ایستگاه، تعداد رویه در جهان تعمیم بیشتر باشد، از مدل «اثرات تصادفی»[30] استفاده می‌گردد (8).

ضریب تعمیم‌پذیری: ضریب تعمیم‌پذیری[31] مقداری بین صفر تا یک دارد و معادل ضریب پایایی آزمون در نظر گرفته می‌شود. این ضریب مشخص می‌کند که نمره دانشجو را تا چه حد می‌توان به تمام رویه‌ها تعمیم داد. به عبارت دیگر مشخص می‌کند که نمره دانشجو تا چه حد به نمره واقعی او نزدیک است. نحوه محاسبه ضریب تعمیم‌پذیری به این ترتیب است که ابتدا باید رویه‌های موردنظر را تعیین کرد. هر رویه واجد مقداری به نام «جزء واریانس»[32] است که از طریق ANOVA محاسبه می‌شود. از اجزای واریانس رویه‌های مختلف در نهایت طبق فرمول مشخصی ضریب تعمیم‌پذیری به دست می‌آید. فرمول‌های متنوعی برای محاسبه ضریب تعمیم‌پذیری یک آزمون وجود دارد که بسته به این که کدام منابع خطا در نظر می‌گیریم و چگونه نمرات را تفسیر می‌کنیم، انتخاب می‌شوند. به عنوان مثال، از لحاظ مفهومی می‌توان دو نوع ضریب تعمیم‌پذیری مطلق[33] و نسبی[34] برای آزمون در نظر گرفت که به هدف برگزاری آزمون بر می‌گردد (7 و 8). اگر قرار است از نمرات برای تصمیم‌گیری به صورت معیارمحور استفاده شود (مانند آزمون پایان ترم که ردی و قبولی هر دانشجو نسبت به یک معیار مشخص می‌شود)، فرمول ضریب مطلق استفاده می‌شود اما اگر قرار است از نتایج آزمون به صورت هنجارمحور (مانند آزمون پذیرش که ردی و قبولی دانشجو بر اساس رتبه وی تعیین می‌گردد) استفاده شود، از فرمول ضریب نسبی استفاده می‌شود.

محاسبه ضریب تعمیم‌پذیری در حالت تک‌رویه: در عالم واقع آن چه در یک آزمون به عنوان منابع خطا دخیل است، محدود به یک رویه نیست. اما می‌توان در نظر گرفت که در یک آزمون فرضی تنها نقش یک رویه برای ما مهم است و می‌خواهیم تأثیر آن را بسنجیم. به عنوان مثال، یک ایستگاه با سه آزمونگر برای 10 دانشجو برگزار می‌شود. آزمونگران با مشاهده عملکرد فراگیران نمره‌ای بین 1 تا 10 به آن‌ها اختصاص می‌دهند (جدول 1).

جدول 1- توزیع نمرات ده دانشجو در یک ایستگاه OSCE با سه آزمونگر

شماره دانشجو	آزمونگر یک	آزمونگر دو	آزمونگر سه
1	6	7	8
2	4	5	6
3	2	2	2
4	3	4	5
5	5	4	6
6	8	9	10
7	5	7	9
8	6	7	8
9	4	6	8
10	7	9	8

برای برآورد میزان خطای این آزمون تنها یک رویه یعنی آزمونگر را در نظر می‌گیریم. اگر برای داده‌های فوق ANOVA را به صورت معمول انجام دهیم، دو جدول برای منابع تغییرات بین گروه[35] (دستیاران) و درون گروه[36] (آزمونگران و خطا) خواهد داد که خلاصه داده‌های آن‌ها در جدول 2 آمده است. باید از جدول‌ها مقادیر میانگین مربعات را انتخاب کنیم و سپس به واریانس تبدیل نماییم یا مستقیماً خروجی اجزای واریانس را دریافت کنیم. با جایگزین کردن مقادیر در فرمول ضریب تعمیم‌پذیری، مقدار آن به صورت زیر به دست می‌آید:

$$G = \frac{\sigma^2_{student}}{\sigma^2_{student} + \sigma^2_{error}}$$

$$G = \frac{4.037}{4.037 + 0.556} = 0.88$$

همان گونه که مشخص است ضریب تعمیم‌پذیری مقدار خوبی دارد و سهم آزمونگران در ایجاد خطا 17% واریانس کل بوده است که زیاد نیست. یعنی آزمونگران توانسته‌اند ارزیابی خود را با

[28] Universe of generalization
[29] Fixed effects
[30] Random effects
[31] G coefficient
[32] Variance component
[33] Absolute
[34] Relative

[35] Between-Subjects
[36] Within-Subjects

خطای اندکی انجام دهند و نمرات قابل اطمینان است. سهم بزرگی از واریانسی که در نمرات وجود دارد، یعنی بیش از 70% آن، مربوط به عملکرد فراگیران و قابل قبول است.

محاسبه ضریب تعمیم‌پذیری در حالت دو رویه: اکنون یک OSCE را تصور کنید که با 3 ایستگاه و 2 آزمونگر در هر ایستگاه برگزار شده است. برای سهولت محاسبات، مدل اثرات ثابت را در نظر می‌گیریم. به عبارت دیگر فرض بر این است که جهان تعمیم و جهان مشاهدات ممکن یکی هستند. در واقع همین سه ایستگاه و همین دو آزمونگر موجود بودند. یا این که هدف ما این نیست که نتایج آزمون را به setting دیگری تعمیم دهیم. واریانس رویه‌ها در این حالت در جدول 3 آمده است. طبق فرمول زیر، ضریب تعمیم‌پذیری آزمون دست می‌آید (ایستگاه با t نشان داده شده است) یعنی واریانس تمام منابع به همراه واریانس خطا در مخرج کسر قرار می‌گیرند:

$$G = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{s\times r}^2 + \sigma_{s\times t}^2 + \sigma_{r\times t}^2 + \sigma_{residual}^2}$$

$$G = \frac{60}{60 + 20 + 15 + 15 + 5 + 15 + 20} = \frac{60}{150} = 0.40$$

قبلاً اشاره شد که در نظر گرفتن هدف آزمون، یعنی هنجاری یا معیاری بودن آن، در محاسبه ضریب تعمیم‌پذیری مهم است. فرمول فوق مربوط به ضریب مطلق است. از آنجا که در حالت معیارمحور، باید نمره دقیق تک تک دانشجویان تعیین شود، در تعیین میزان تعمیم‌پذیری نمرات نیز تمام منابع خطا باید لحاظ شوند. در ضریب نسبی مشخص کردن جایگاه هر دانشجو نسبت به بقیه کفایت می‌کند و تعیین نمره وی مدنظر نیست. در واقع، تفاوتی که آزمونگران ایجاد می‌کنند، تأثیری در رتبه دانشجویان نسبت به یکدیگر ندارد. به همین دلیل منابع خطای کمتری وارد فرمول می‌شوند یعنی واریانس آزمونگر، واریانس تعامل آزمونگر با دانشجو و واریانس تعامل آزمونگر با ایستگاه حذف می‌شوند. این موضوع باعث می‌شود مقدار عددی ضریب تعمیم‌پذیری مطلق از مقدار عددی ضریب تعمیم‌پذیری نسبی کمتر باشد. واریانس رویه‌ها در حالت هنجاری در جدول 4 آمده‌اند. ضریب تعمیم‌پذیری نسبی که با Φ (فی) هم نشان داده می‌شود، با فرمول زیر به دست می‌آید:

$$G_{relative} = \Phi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2 + \sigma_{s\times t}^2 + \sigma_{residual}^2}$$

$$G_{relative} = \Phi = \frac{60}{60 + 15 + 5 + 20} = \frac{60}{100} = 0.60$$

مطالعه تعمیم‌پذیری و مطالعه تصمیم‌گیری: تخمین سهم منابع مختلف خطا در آزمون و تعیین میزان تعمیم‌پذیری نتایج آزمون که قبلاً شرح داده شد، به «مطالعه تعمیم‌پذیری»[37] معروف است. در عین حال، از دیگر قابلیت‌های این نظریه، امکان پیش‌بینی پایایی آزمون در شرایط مختلف فرضی است که «مطالعه تصمیم‌گیری»[38] نامیده می‌شود (8). در واقع تلاش می‌شود با استفاده از یافته‌های مطالعه تعمیم‌پذیری، طرح‌های دیگری برای آزمون پیشنهاد شود (مثلاً افزایش تعداد ایستگاه‌ها یا آزمونگران) که در آن‌ها مقدار خطای آزمون به حداقل رسیده و تعمیم‌پذیری آزمون حداکثری باشد.

مزایا و محدودیت‌های نظریه تعمیم‌پذیری: همان‌طور که قبلاً اشاره شد، مزایای نظریه تعمیم‌پذیری در طراحی و تحلیل آزمون‌های مختلف شامل مشخص کردن سهم منابع مختلف خطا، برآورد پایایی و ارایه پیشنهاد برای بهبود تعمیم‌پذیری نتایج است. از محدودیت‌های نظریه تعمیم پذیری می‌توان به اجرایی نبودن استفاده از نظریه تعمیم‌پذیری در شرایط واقعی اشاره کرد. به دلیل محدودیت‌های عملی، معمولاً از این نظریه در مطالعات پایلوت (تعداد معدود ایستگاه‌ها و ارزیابان) استفاده می‌شود و سپس نتایج حاصل به شرایط واقعی ارزیابی تعمیم داده می‌شوند (10 و 11). همچنین پیچیدگی و کمبود نرم‌افزارهای تحلیل از محدودیت‌های دیگر نظریه تعمیم‌پذیری است (12).

جدول 2- نتیجه آزمون ANOVA درون گروه و بین گروه و اجزای واریانس مربوطه

منبع خطا	جمع مربعات	درجه آزادی	میانگین مربعات	مقدار واریانس	درصد واریانس از کل
دانشجو	114	9	12/67	4/037	73
آزمونگر	20	2	10	0/944	17
خطا یا باقی‌مانده (دستیار × آزمونگر)	10	18	0/56	0/556	10

جدول 3- واریانس رویه‌های یک آزمون OSCE در حالت معیاری (مطلق)

منبع خطا	جمع مربعات	درجه آزادی	میانگین مربعات	مقدار واریانس	درصد واریانس از کل
دانشجو (10 نفر)	3915	9	435	60	40
آزمونگر (2 نفر)	815	1	815	20	13
ایستگاه (3 ایستگاه)	960	2	480	15	10
دانشجو × آزمونگر	585	9	65	15	10
دانشجو × ایستگاه	540	18	30	5	3
ایستگاه × آزمونگر	340	2	170	15	10
خطا یا باقی مانده (دانشجو × آزمونگر× ایستگاه)	360	18	20	20	13

جدول 4- واریانس رویه‌های یک آزمون OSCE در حالت هنجاری (نسبی)

منبع خطا	جمع مربعات	درجه آزادی	میانگین مربعات	مقدار واریانس	درصد واریانس از کل
دانشجو (10 نفر)	3915	9	435	60	52
ایستگاه (3 ایستگاه)	960	2	480	15	13
آزمونگر : ایستگاه	1155	3	815	20	17
دانشجو × ایستگاه	540	18	30	5	4
خطا یا باقی مانده (دانشجو × آزمونگر: ایستگاه)	945	27	170	15	13

نتیجه‌گیری

نمرات دانشجویان در هر موقعیت اندازه‌گیری تحت تأثیر ویژگی‌های مشخصی از قبیل سؤالات آزمون، شرایط آزمون و آزمونگران قرار دارد که می‌توانند به عنوان منابع خطای اندازه‌گیری موجب دور شدن نمره مشاهده‌شده دانشجو از نمره واقعی وی و کاهش پایایی آزمون شوند. در نظریه کلاسیک آزمون، شناسایی میزان اثر منابع مختلف خطاهای احتمالی به تفکیک امکان‌پذیر نیست در حالی که با استفاده از نظریه تعمیم‌پذیری می‌توان سهم هر یک از منابع خطا را شناسایی نمود و به این ترتیب ضمن برآورد ضریب پایایی امتحان برای برگزاری آزمونی پایاتر برنامه‌ریزی کرد.

[37] Generalizability study (G study)
[38] Decision study (D study)

References

1- Brennan R. *Generalizability theory*. New York: Springer Verlag; 2001
2- Raykov T, Marcoulides GA. *Introduction to Psychometric Theory*. 1st ed. New York: Routledge; 2010
3- Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Medical Teacher 2013;35(9):e1437-46.
4- Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Medical Teacher 2012;34(3):e161-75.
5- Courville TG. An empirical comparison of item response theory and classical test theory item/person statistics. Texas, US: Texas A&M University; 2004.
6- Cronbach L, Gleser GC, Harinder N, Nageswari R. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. New York: Wiley; 1972.
7- Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Medical Teacher 2012;34(11):960-92.
8- Streiner DL, Norman GR. Health Measurement Scales: A practical guide to their development and use. 4th ed. Oxford University Press: USA; 2008
9- Cardinet J, Johnson S, Pini G. Applying Generalizability Theory using EduG (Quantitative Methodology Series). 1st ed. Routledge; 2011
10- Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. J Manipulative Physiological Therapy 2006;29(6):463-7.
11- Clauser BE, Harik P, Margolis MJ, Mee J, Swygert K, Rebbecchi T. The generalizability of documentation scores from the USMLE Step 2 Clinical Skills examination. Acad Med 2008;83(10 Suppl):S41-4.
12- Webb NM, Shavelson RL. Generalizability Theory: Overview. Encyclopedia of Statistics in Behavioral Science 2005; 2:717-719.

An Overview of the Generalizability Theory and its Implications in Medical Education

Mortaz Hejri S[1] (MD, MSc, PhD candidate), Janani L[2] (PhD), Jalili M[3*] (MD)

[1] *Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran*
[2] *Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran*
[3] *Department of Medical Education, Department of Emergency Medicine, Tehran University of Medical Sciences, Tehran, Iran*

Original Article
Received: 6 Apr 2015, Accepted: 21 Jun 2015

Abstract

Introduction: All measurements, including medical students' assessment, are potentially prone to error from various sources. Considering the importance of maintaining the quality of assessment methods in medical education, finding methods to estimate the error and to decrease it, is essential. Many developments have been achieved in measurement theories during recent years, which can be used in this way.

Methods: In this study, following reviewing the classical test theory (CTT) and its limitations, we discuss the generalizability theory (GT) using some examples of the Objective Structured Clinical Examination.

Results: According to the CTT, students' scores in exams are not their true scores and are accompanied with some levels of error. Determining reliability of tests helps us to assess how much the random error has been affected the score. In the CTT, it is impossible to consider all error sources simultaneously. To cope with this limitation, the GT specifies share of various sources of error based on the ANOVA models. This theory estimates reliability of tests and offers some recommendations for improving their generalizability.

Conclusion: In CTT, specification of effect of various sources of errors is not possible, whereas GT can be used to identify share of each error source. Hence, the test reliability coefficient can be estimated and also a more reliable test may be planned.

Key words: measurement, classical test theory, generalizability theory, reliability

Please cite this article as follows:

Mortaz Hejri S, Janani L, Jalili M. An Overview of the Generalizability Theory and its Implications in Medical Education. Hakim Health Sys Res 2015; 18(2): 146- 152.

[*]Corresponding Author: No. 57, Hojatdust St., Keshavarz Blvd., Tehran, Iran. Tel: +98- 21- 88955846, Fax: +98- 21- 889550160. E-mail: mjalili@tums.ac.ir