فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
دوره ششم، شماره ۲۴، زمستان ۹۵، ص ۴۳ تا ۸۱

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 6, No. 24, Winter 2017

معرفی نظریه تعمیم‌پذیری و تشریح فرآیند یک مطالعه اندازه‌گیری جهت سنجش اعتبار

نورعلی فرخی[۱]، لیلا بهرامی[۲]

تاریخ دریافت: ۹۵/۰۳/۲۱
تاریخ پذیرش: ۹۵/۱۰/۱۵

چکیده

شناسایی و جداسازی منابع چندگانه خطای اندازه‌گیری و برآورد هر یک از آن‌ها، تمایز گذاشتن میان تصمیم‌های نسبی و مطلق، تمایز گذاشتن میان رویه‌های اندازه‌گیری ثابت و تصادفی و همچنین پرداختن به طرح‌های مختلف مطالعه D را می‌توان ازجمله نقاط قوت نظریه تعمیم‌پذیری ذکر کرد که نظیر آن‌ها در نظریه کلاسیک آزمون وجود ندارد. باوجوداین نقاط قوت، نظریه تعمیم‌پذیری برای محققان کشورمان ناشناخته است و تحقیقات انگشت‌شماری در این زمینه صورت گرفته است. هدف از این مقاله، معرفی نظریه تعمیم‌پذیری و نشان دادن قابلیت کاربرد عملی آن در سنجش اعتبار اندازه‌ها بود. علاوه بر مقایسه میان دو نظریه کلاسیک آزمون و تعمیم‌پذیری، چارچوب مفهومی نظریه تعمیم‌پذیری به‌سادگی بیان گردید. همچنین، در این مقاله فرآیند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری در قالب یک مثال به همراه محاسبات و معادلات مربوطه، در ۱۵ گام اصلی با جزئیات تشریح شد تا برای محققان و آزمون سازان یک راهنمای کاربردی جهت سنجش اعتبار باشد. سودمندی GT نسبت به CTT جهت برآورد اعتبار به‌خصوص در وضعیت‌های اندازه‌گیری پیچیده نمایان گردید. همچنین، GT محققان را قادر می‌سازد که با اقدامات بهینه‌سازی، سهم خطا را در طرح اندازه‌گیری‌شان کاهش دهند که این کار، افزایش دقت در تعمیم نتایج را به دنبال خواهد داشت.

۱. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران (نویسنده مسئول)
farrokhinoorali@yahoo.com
۲. کارشناسی ارشد گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی

Archive of SID
www.SID.ir

واژگان کلیدی: اعتبار، نظریه تعمیم‌پذیری، نظریه کلاسیک اندازه‌گیری

مقدمه

عوامل مختلفی، نمرات مشاهده‌شده آزمون را تحت تأثیر قرار می‌دهد. در نظریه کلاسیک آزمون[1] (CTT)، همه این عوامل تحت یک عبارت خطای تصادفی نامتمایز در مدل آماری گنجانیده می‌شود. روان سنجان برای لحاظ کردن این عوامل و تفکیک آن‌ها در مدل‌های آماری‌شان، نظریه تعمیم‌پذیری[2] (GT) را گسترش دادند.

GT، یک نظریه‌ی آماری درباره‌ی اتکاپذیری اندازه‌های رفتاری است (شولسون و وب[3]، ۱۹۹۱). بنا به گفته‌ی برنان[4] (۱۹۹۲)، بااینکه ریشه‌های GT را می‌توان در CTT و تحلیل واریانس (آنوا)[5] یافت، باوجوداین، چارچوب مفهومی GT منحصربه‌فرد است. GT محقق را قادر می‌سازد که منابع چندگانه‌ی خطا را در یک روش اندازه‌گیری شناسایی، تفکیک و برآورد کند و بدین طریق، CTT را به نحو چشمگیری گسترش داده است (شولسون، وب و رولی[6]، ۱۹۸۹؛ برنان، ۲۰۰۱). GT با در نظر گرفتن همزمان منابع چندگانه‌ی خطا، اجازه می‌دهد مدل دقیق‌تری از وضعیت اندازه‌گیری ساخته شود که در این صورت، نتایج به‌دست‌آمده را با دقت بیشتری می‌توان به سایر موقعیت‌های اندازه‌گیری تعمیم داد (شولسون و وب، ۱۹۹۱؛ ون لیون[7]،۱۹۹۷؛ برنان،۲۰۰۱). این در حالی است که CTT در برآورد اعتبار[8] وضعیت‌های اندازه‌گیری که شامل منابع چندگانه‌ای از خطاست، کارایی ندارد. بااین‌حال، سهولت کاربرد و درک روشن از CTT و همچنین، پیچیدگی مفهومی GTدر مقایسه با CTT، باعث شده است که CTT تا به امروز در سنجش اعتبار همچنان به قوت خود باقی بماند. بررسی مطالب مجلات علمی معتبر و پایان‌نامه‌های دانشجویان

1. Classical Test Theory (CTT)
2. Generalizability Theory (GT)
3. Shavelson & Webb
4. Brennan
5. analysis of variance (ANOVA)
6. Rowley
7. VanLeeuwen
8. reliability

Archive of SID

www.SID.ir

تحصیلات تکمیلی در کشورمان، نشان می‌دهد که محققان و دانشجویان حوزه‌های مختلف ازجمله آموزش‌وپرورش، روان‌شناسی، پزشکی و ... برای برآورد اعتبار داده‌هایشان از رویکردهای کلاسیک (مانند آلفای کرونباخ، باز آزمون، فرم‌های موازی، دونیمه کردن و ...) استفاده می‌کنند و بندرت می‌توان به مقاله‌ای اشاره کرد که از طرح‌های اندازه‌گیری[1] GT جهت سنجش اعتبار استفاده کرده باشد؛ بنابراین، به دلیل اهمیت موضوع، در مقاله‌ی حاضر سعی شده است ضمن معرفی GT به محققان، قابلیت این نظریه در سنجش اعتبار اندازه‌ها نشان داده شود. بدین منظور، ابتدا تاریخچه‌ی GT بیان می‌گردد و بعدازآن، خلاصه-ی مفیدی از شباهت‌ها و تفاوت‌های موجود میان CTT و GT مطرح می‌شود که به محققان کمک خواهد کرد، درک مفهومی خود را از دو نظریه مذکور گسترش دهند و علاوه بر این، به مزیت‌های GT پی ببرند. سپس، مفاهیم و رویکردهای GT تک متغیری[2] تشریح می‌شود تا محققان بتوانند با چارچوب مفهومی GT بهتر ارتباط برقرار کنند. در پایان، به‌منظور آشنایی با چگونگی انجام یک مطالعه اندازه‌گیری جهت سنجش اعتبار، فرآیند طراحی، تحلیل و تفسیر این مطالعه اندازه‌گیری در قالب یک مثال در بافت GT تک متغیری ارائه می‌گردد.

تاریخچه GT. سال‌ها قبل از آن‌که GT به‌طور رسمی توسط کرونباخ[3] و همکارانش معرفی شود، نویسندگان مختلفی در این زمینه تلاش کرده بودند. به‌طوری‌که طبق گفته‌ی کرونباخ (۱۹۹۱، به نقل از برنان، ۱۹۹۷)، GT را می‌توان یک اثر ترکیبی به تألیف حداقل ۲۴ نویسنده پنداشت. با استفاده از مقاله برنان (۱۹۹۷) که در خصوص تاریخچه‌ی GT نگاشته شده است، شرحی از مهم‌ترین کارها و آثار نویسندگان مختلف در جدول زیر ارائه گردیده است.

1. measurement designs
2. univariate
3. Cronbach

Archive of SID

www.SID.ir

جدول ۱. خلاصه‌ای از تاریخچه‌ی GT در خلال سال‌های ۱۹۹۷-۱۹۳۶

۱۹۳۶	بارت[1] با مقاله‌ای با عنوان «تحلیل نمرات امتحانی» اولین کسی بود که رویکرد آنوا را برای مسائل اندازه‌گیری بکار گرفت.
۱۹۴۱	هایت[2] نشان داد از طریق آنوا می‌توان اعتبار آزمون را برآورد کرد.
۱۹۴۷	کرونباخ بیان کرد؛ برای مرتفع کردن ناهمسانی‌ها در برخی از برآوردهای اعتبار انواعی از تحلیل‌های چند رویه‌ای[3] لازم است.
۱۹۵۰	- پیش از سال ۱۹۵۰، آثار غنی در مورد اعتبار از چشم‌انداز CTT وجود داشت که اکثر آن‌ها به طرز فوق‌العاده‌ای در کتاب گالیکسن[4] (۱۹۵۰) منتشر شد. - در دهه‌ی ۱۹۵۰ محققان در کارهایشان از این موضوع که آنوا می‌تواند به‌طور همزمان رویه‌های چندگانه را بررسی کند، استفاده کردند.
۱۹۵۱	- احتمالاً اولین بررسی اعتبار برحسب مؤلفه‌های واریانس[5]، مطالعه‌ی فینلایسون[6] درباره‌ی نمرات تخصیص داده‌شده به مقالات بود. - ایبل[7] مقاله‌ای را در مورد اعتبار درجه‌بندی‌ها منتشر کرد که در آن دو نوع واریانس خطا را در نظر گرفته بود؛ یکی شامل اثرات اصلی ارزیاب بود و دیگری بدون آن. همچنین، طرح‌های متقاطع و آشیانه‌ای[8] یک‌رویه‌ای را در نظر گرفت.
۱۹۵۲	پلینر[9] ارتباط نظری میان همبستگی‌های بین طبقه‌ای و آنوا را به دست آورد.
۱۹۵۳	لیندکوئیست[10] گزارش جامعی از نظریه چند رویه‌ای ارائه داد که تمرکز آن روی برآورد مؤلفه‌های واریانس در مطالعات اعتبار بود. لیندکوئیست نشان داد که تحلیل‌های چند رویه‌ای منجر به تعاریف متفاوتی از خطا و ضرایب اعتبار نسبت به قبل شده است. بخش لیندکوئیست به‌وضوح بر قسمت‌های مهمی از GT دلالت می‌کرد.
۱۹۵۹-۱۹۵۵	- درواقع، اثرات اصلی ارزیاب در مقاله‌ی ایبل نقش اثرات اصلی سؤال را در مقالات لرد[11] (۱۹۵۵،۱۹۵۷،۱۹۵۹) که درباره‌ی خطاهای استاندارد اندازه‌گیری شرطی[12] (SEMs) و اعتبار تحت

1. Burt
2. Hoyt
3. multifacet
4. Gulliksen
5. variance components
6. Finlayson
7. Ebel
8. crossed & nested
9. Pilliner
10. Lindquist
11. Lord
12. conditional standard errors of measurement

Archive of SID

www.SID.ir

	مفروضات مدل خطای دوجمله‌ای بود، ایفا می‌کرد. علاوه بر این، لرد در مقالاتش با صراحت از آزمون-های تصادفی موازی[1] نام برد. موضوعاتی که لرد با آن دست‌وپنجه نرم می‌کرد، تأثیر واضحی بر گسترش GT داشت. - کارهای ایبل و لرد سرانجام به تمایز میان خطای نسبی و مطلق[2] انجامید. - در مقاله‌ی بارت (۱۹۵۵) با عنوان «برآورد اعتبار آزمون به شیوه‌ی تحلیل واریانس» و اثر مدلی، میتزل و دوی[3] (۱۹۵۶) در مورد مشاهدات کلاسی، رویه‌های چندگانه به‌طور همزمان بررسی گردید. - کورنفیلد و توکی[4] (۱۹۵۶) قواعدشان را در خصوص معادلات میانگین مجذورات مورد انتظار (EMS)[5] برحسب مؤلفه‌های واریانس منتشر کردند.
۱۹۶۱-۱۹۶۰	- دستاورد کرونباخ و همکارانش، یک چارچوب مفهومی و استفاده از یک متدولوژی (تحلیل مؤلفه‌های واریانس) بود که مقالات بسیاری از محققان را، باوجوداینکه برخی از آن‌ها مغایر هم به نظر می‌رسیدند، یکپارچه کرده بود. - خطوط اصلی GT تک‌متغیری با گزارش‌های فنی کرونباخ، راجاراتنام و گلیسر[6] (در قالب سه مقاله هرکدام با نویسنده‌ی اول متفاوت) تا حد زیادی کامل شد.
۱۹۶۶-۱۹۶۴	- بررسی‌های ناندا[7] درباره‌ی اعتبار درون مجموعه‌ای، انگیزشی برای توسعه‌ی GT چندمتغیری[8] گردید. - تیم کرونباخ، توسعه GT چند متغیری را آغاز کردند که در آن هر یک از سطوح یک‌رویه ثابت[9] یا بیشتر، با یک نمره جهان[10] متمایز مرتبط می‌شد.
۱۹۷۲	کرونباخ، گلیسر، ناندا و راجاراتنام، کتابی با عنوان «اتکاپذیری اندازه‌های رفتاری» منتشر کردند که هنوز کامل‌ترین و گسترده‌ترین بررسی از GT است.
۱۹۷۶	کاردینت، تورنر و الل[11]، اصل تقارن[12] را مطرح کردند.
۱۹۷۷	برنان و کانه[13] در مقاله‌های مشترکشان، در پرداختن به موضوعات پیرامون اعتبار نمرات ملاک مرجع، GT را بکار گرفتند.
۱۹۸۱	شولسون و وب، مروری از GT در طی سال‌های ۱۹۸۰- ۱۹۷۳ را منتشر کردند.

1. randomly parallel
2. relative & absolute error
3. Medley, Mitzel & Doi
4. Cornfield & Tukey
5. expected mean squares
6. Rajaratnam & Gleser
7. Nanda
8. multivariate
9. fixed facet
10. universe score
11. Cardinet, Tourneur & Allal
12. symmetry
13. Kane

Archive of SID

www.SID.ir

۱۹۸۲	مقاله کانه با عنوان «یک مدل نمونه‌گیری برای روایی»، یکی از مقاله‌های نظری مهم در ادبیات GT در ۲۵ سال گذشته است.
۱۹۸۳	– نوشتن یک مونوگراف دربارهٔ GT با عنوان «کلیات GT» توسط برنآن که نسبت به اثر کرونباخ و همکارانش (۱۹۷۲) از جامعیت کمتری برخوردار بود. – برنامه کامپیوتری بنام GENOVA توسط کریک[1] و برنان جهت اجرای تحلیل GT طراحی شد.
۱۹۹۶–۱۹۸۹	– علاقه به انجام سنجش در اواخر دههٔ ۱۹۸۰ شهرت نسبتاً زیادی را برای GT رقم زد. – کاربرد GT در سنجش‌های عملکردی به‌خصوص در مجموعه‌ای از سخنرانی‌ها و مقالات شولسون و همکارانش به‌خوبی نشان داده شد. همچنین، برنان و جانسون[2] (۱۹۹۵) و برنان (۱۹۹۶) برخی از مسائل نظری و کاربردی را در این خصوص مطرح کردند. – گزارش‌های خلاصه و ساده از GT از قبیل؛ اثر فلت[3] و برنان (۱۹۸۹)، مقاله‌ی شولسون و همکاران (۱۹۸۹)، کتاب درسی مقدماتی شولسون و وب (۱۹۹۱) و مقاله‌ی برنان (۱۹۹۲) با عنوان «GT» منتشر شد.

در سال ۲۰۰۱، کتاب جامعی در مورد GT توسط برنان نگاشته و منتشر شد. همچنین، می‌توان از کتابی که توسط کاردینت، جانسون و پینی[4] ۱۶(۲۰۱۰) در این زمینه نگاشته شده است، نام برد.

نظریه تعمیم‌پذیری در مقابل نظریه کلاسیک آزمون. طبق گفته‌ی شولسون و همکارانش (۱۹۸۹)، ارتباط بین GT و CTT همانند ارتباط بین آنوای عاملی و ساده[5] است. محققی که آنوای ساده را بکار می‌گیرد، واریانس را به دو مؤلفه‌ی بین گروهی (واریانس منظم) و درون‌گروهی (واریانس تصادفی) تقسیم می‌کند. به شیوه‌ای مشابه، CTT نیز، واریانس را به دو مؤلفه، واریانس واقعی (واریانس منظم) و خطا (واریانس تصادفی) تقسیم می‌کند. با به‌کارگیری آنوای عاملی بجای ساده، محقق می‌پذیرد که عوامل چندگانه‌ای در واریانس کل داده‌ها سهیم است. آنوای ساده، سؤالات محدودتری را مطرح می‌کند و نسبت به آنوای عاملی با کارایی کمتری به آن‌ها می‌پردازد. به همین ترتیب، بسط CTT به GT اثرات

1. Crick
2. Johnson
3. Feldt
4. Pini
5. factorial & simple ANOVA

Archive of SID

www.SID.ir

چندگانه در واریانس اندازه‌گیری را تأیید می‌کند. GT همانند آنوای عاملی، واریانس را به چندین منبع نظیر؛ واریانس منظم در بین اهداف اندازه‌گیری[1]، منابع چندگانه‌ی خطا و به تعاملاتشان تقسیم می‌کند.

مفروضات زیربنایی GT اساساً مشابه مفروضات CTT است. در هر دو نظریه، خطاها مستقل از نمره‌ی جهان / واقعی و نا همبسته‌اند. همان‌طور که در CTT، SEM در همه‌ی سطوح نمرات (برای همه آزمودنی‌ها) یکسان است در GT نیز، SEM یکسانی برای همه‌ی اهداف اندازه‌گیری بدون در نظر گرفتن نمره‌ی جهان زیربنایی به کار می‌رود (استراپ[2]، ۲۰۰۲، به نقل از الخاروسی[3]، ۲۰۱۲). از دیگر شباهت‌های دو نظریه‌ی مذکور می‌توان به موارد زیر اشاره کرد.

از CTT و GT می‌توان به‌عنوان نظریه‌های ارزش مورد انتظار[4] نام برد که نمره واقعی[5] یا نمره‌ی جهان را به‌عنوان یک ارزش مورد انتظار از نمرات مشاهده‌شده تعریف می‌کنند. نمره موردنظر در CTT و GT، نمره‌ی مشاهده‌شده آزمون است و هدف اصلی این دو نظریه، ارزیابی کیفیت نمره مشاهده‌شده آزمون است که از طریق ضرایب اعتبار و خطاهای استاندارد برآورد می‌شود، بدون اینکه تلاشی جهت برآورد نمرات در صفت مکنون صورت گیرد. هر دو نظریه، به‌وضوح خطاهای اندازه‌گیری تصادفی را شامل می‌شوند و مفاهیم اعتبار (تعمیم‌پذیری) در هر دو، به‌خوبی تعریف شده است. واریانس نمره‌ی جهان در GT مشابه واریانس نمره‌ی واقعی در CTT است. همچنین، واریانس خطای نسبی[6] (σ_δ^2) و ضریب تعمیم‌پذیری[7] ($E\rho^2$) در GT به ترتیب مشابه واریانس خطا و ضریب اعتبار در CTT است. از دیگر شباهت‌ها، می‌توان به خطی بودن مدل‌های هر دو نظریه اشاره کرد و اینکه، در برابر نقض مفروضات مدل‌هایشان مقاوم هستند. به بیانی دیگر، مفروضات نسبتاً ضعیفی دارند.

1. objects of measurement
2. Strube
3. Alkharusi
4. expected value theories
5. true score
6. generalizability coefficient
7. generalizability coefficient

Archive of SID

www.SID.ir

علاوه بر این مستلزم نمونه‌هایی با حجم بزرگ نیستند و واحد تحلیل در آن‌ها، نمرات آزمون است (شولسون و وب، ۱۹۸۱؛ بولس، هنوفتس و بیلی[1]، ۱۹۸۲؛ سوئن و لی[2]، ۲۰۰۷؛ برنان، ۲۰۱۱a). دو نظریه مذکور، به‌رغم شباهت‌ها، تفاوت‌های بسیار مهمی نیز دارند که در زیر به شرح مواردی از آن پرداخته می‌شود.

چارچوب مفهومی: GT نسبت به CTT چارچوب مفهومی قدرتمندتری دارد که منجر به برطرف کردن تعدادی از تناقضات آشکار در چند بحث CTT از اعتبار شده است. دو ویژگی مهم GT که به حل تناقضات کمک می‌کند، عبارت‌اند از: تمایز گذاشتن GT میان رویه‌های[3] اندازه‌گیری ثابت و تصادفی و همچنین قابلیت این نظریه در پرداختن به طرح‌های مختلف مطالعه‌ی تصمیم[4] (مطالعه D) (برنان، ۲۰۱۱a).

پیچیدگی مفهومی: چارچوب مفهومی قدرتمند GT و قابلیت آن در جداسازی منابع خطا منجر به پیچیدگی مفهومی آن شده است که این موضوع، شناخت محققان را می‌طلبد (برنان، ۲۰۱۱a). همان‌طور که شولسون و همکاران (۱۹۸۹) مطرح کرده‌اند، پیچیدگی GT می‌تواند روشن کند که چرا CTT روش ارجحی برای برآورد اعتبار باقی مانده است.

تعاریف فرم‌های موازی: در CTT تعاریف آزمون‌های موازی و آزمون‌های اساساً تائو معادل، اغلب غیرقابل‌دفاع هستند. درحالی‌که GT فرض می‌کند که آزمون‌ها تصادفی موازی هستند و محتوای آزمون یک نمونه تصادفی از حیطه یا جهان تعریف‌شده در نظر گرفته می‌شود (سوئن و لی،۲۰۰۷). برنان (۲۰۱۱a) بیان می‌کند که GT مفهوم ساده‌ای از موازی بودن را در خود گنجانده است که کاملاً متفاوت با مفهوم کلاسیک فرم‌های موازی در CTT است. همچنین مطرح می‌کند، هر دو نوع موازی بودن ایده‌آل هستند و هیچ‌گاه احتمال اینکه کاملاً واقعیت داشته باشد، نیست. اگرچه یکی یا دیگری ممکن است در زمینه‌ای خاص مناسب‌تر باشد.

1. Bolus, Hinofotis & Bailey
2. Suen & Lei
3. facets
4. decision study

Archive of SID

www.SID.ir

مدل‌سازی نمرات مشاهده‌شده: در CTT نمره‌ی مشاهده‌شده (X) یک فرد در آزمون از دو متغیر غیرقابل مشاهده یعنی؛ یک نمره واقعی (T) و یک نمره خطا (E) تشکیل شده است و به‌صورت مدل خطی ($X = T + E$) نشان داده می‌شود. در GT نمره مشاهده‌شده در یک آزمون از رویه‌های مختلف مورداستفاده در آزمون تأثیر می‌پذیرد و با توجه به رویه‌های مورداستفاده در آزمون معرف عملکرد فرد در همان رویه‌هاست (کیامنش، ۱۳۷۴). طبق بیان شولسون و وب (۱۹۹۱)؛ نمره‌ی مشاهده‌شده‌ی یک فرد به یک نمره جهان و یک یا بیشتر از یک منبع خطا تجزیه می‌شود. در صفحه‌ی ۱۷، مدل‌سازی نمره مشاهده‌شده در GT برای طرح s×i×r نشان داده شده است.

منابع چندگانه‌ی خطا و برآوردهای چندگانه از اعتبار: در وضعیت‌های اندازه‌گیری پیچیده که شامل منابع چندگانه‌ای از خطای اندازه‌گیری (رویه‌ها) است، CTT قادر به برآورد اعتبار نیست؛ زیرا شیوه‌های سنتی اعتبار تنها برای یک منبع خطا طراحی شده‌اند. روش معمول CTT برای برآورد اعتبار این است که از روش‌های مختلفی (همچون باز آزمایی، بین ارزیابان، همسانی درونی،...) استفاده می‌کند. درنتیجه، برآوردهای خطا و برآوردهای اعتبار (نسبت واریانس نمره واقعی به مشاهده‌شده) مطابق با طرح جمع‌آوری داده‌ها تغییر می‌کند. اینکه روش‌های مختلف، ضرایب اعتبار مختلفی را به دنبال دارند، به‌نوبه‌ی خود منجر به خطاهای استاندارد اندازه‌گیری متفاوتی می‌شود. سؤالی که اینجا پیش می‌آید این است که در چنین وضعیتی، دقیق‌ترین برآورد ضریب اعتبار کدام است؟ و به‌منظور ساخت فاصله‌های اطمینان حول نمرات مشاهده، کدام خطای استاندارد اندازه‌گیری را باید به کار برد؟ متأسفانه CTT قادر به پاسخ‌گویی به این سؤالات نیست و تعاریف متغیر نمرات واقعی و خطا را مسکوت می‌گذارد. درحالی‌که GT اذعان می‌دارد که تعاریف چندگانه‌ای از نمرات جهان و خطا می‌تواند وجود داشته باشد. برخلاف CTT، در GT می‌توان منابع چندگانه خطا را همزمان در ترکیب‌های متفاوتی از تصادفی یا ثابت در نظر گرفت. با تعیین اینکه آیا یک‌رویه تصادفی یا ثابت باشد، امکان برآورد اعتبار و خطای استاندارد ناشی از منابع معین خطا در GT وجود دارد. به بیانی دیگر، GT سهم هر منبع خطا را در واریانس نمرات آزمون تعیین می‌کند و فرصت محاسبه‌ی برآوردهای متفاوتی از اعتبار

Archive of SID

www.SID.ir

را می‌دهد که بستگی به این دارد که کدام منبع خطا برای هر استفاده‌ی خاص از آزمون مهم در نظر گرفته می‌شود (شولسون و همکاران، ۱۹۸۹؛ ون لیون، ۱۹۹۷؛ سوئن و لی،۲۰۰۷؛ فن و سان، ۲۰۱۳).

سنجش‌های هنجار مرجع و ملاک مرجع[1] (تمایز گذاشتن میان تصمیم‌های نسبی و مطلق[2]): CTT چون نمی‌تواند خطای اندازه‌گیری منظم را در خود جای دهد، تنها برای سنجش اعتبار آزمون‌های هنجار مرجع مناسب است. درحالی‌که GT به دلیل انعطاف‌پذیری که دارد هر دو خطای اندازه‌گیری نسبی و مطلق را در خود جای می‌دهد و میان تصمیم‌های نسبی و مطلق تمایز می‌گذارد؛ بنابراین، هم برای سنجش هنجار مرجع مناسب است و هم نسبت به CTT، برآورد مناسبی از اعتبار را برای آزمون‌های ملاک مرجع فراهم می‌آورد (شولسون و وب، ۱۹۹۱؛ سوئن و لی، ۲۰۰۷؛ کمازاوا[3]، ۲۰۰۹).

طراحی و بهینه‌سازی پروتکل اندازه‌گیری: در طرح‌های اندازه‌گیری چند رویه‌ای، با به‌کارگیری GT می‌توان هر یک از منابع واریانس تشکیل‌دهنده‌ی خطای کل را شناسایی و با انجام یک مطالعه‌ی تعمیم‌پذیری[4] (مطالعه‌ی G)، سهم جداگانه‌ی هر یک از آن‌ها را برآورد کرد. به دنبال آن، برای بهبود روند اندازه‌گیری و از طریق انجام مطالعه‌ی D، با دست‌کاری رویه یا رویه‌هایی که سهم بیشتری در خطای کل دارند، می‌توان خطای اندازه‌گیری برآورد شده را کاهش داد. قابلیت GT در پرداختن به طرح‌های مختلف مطالعه D به محقق اجازه می‌دهد که با در نظر گرفتن اندازه اعتبار دلخواه، هزینه، زمان و دیگر محدودیت‌های عملی، پروتکل اندازه‌گیری بهینه‌ای را برای مطالعات واقعی و کاربردهای عملی طراحی و ارزشیابی کند. در مقابل، در طرح‌های چند رویه‌ای، نظریه کلاسیک از طریق روش شناخته‌شده‌ای مثل فرمول پیشگویی اسپیرمن- براون[5] یک برآورد تک‌بعدی ازآنچه خطا را کاهش خواهد داد فراهم می‌کند. به‌بیان‌دیگر، این فرمول برخلاف مطالعات

1. norm-referenced & criterion-referenced
2. relative & absolute decisions
3. Kumazawa
4. generalizability study
5. Spearman–Brown prophecy formula

Archive of SID

www.SID.ir

D، نمی‌تواند اثر تغییر تعداد سطوح دورویه یا بیشتر را به‌طور همزمان ارزیابی کند (شولسون و وب، ۱۹۹۱؛ بولس و همکاران، ۱۹۸۲؛ فن و سان، ۲۰۱۳).

نظریه تعمیم‌پذیری چندمتغیری: GT چندمتغیری، مسائل اعتبار را در راستای جهان‌های تعمیم چندگانه گسترش داده است که وضعیت متناظر آن در CTT وجود ندارد (برنان، ۲۰۱۱a).

اصل تقارن: در CTT، اشخاص به‌عنوان هدف اندازه‌گیری در نظر گرفته می‌شوند و راجع به مناسب بودن یک اندازه برحسب اینکه چقدر خوب ما را برای تمایز قائل شدن میان افراد قادر می‌سازد، قضاوت می‌شود. همه رویه‌های دیگر (سؤالات آزمون، موقعیت‌ها، ارزیابان، غیره) به‌عنوان خطا در نظر گرفته می‌شوند (شولسون و همکاران، ۱۹۸۹). کاردینت، تورنر و الل (۱۹۷۶) با مطرح کردن اصل تقارن در GT، بیان کردند؛ برخلاف تمرکز سنتی روی افراد، هدف اندازه‌گیری ممکن است بسته به هدف خاص تصمیم‌گیرنده تغییر کند و تفاوت‌های فردی ممکن است به‌عنوان منبع خطا در نظر گرفته شوند.

مفاهیم و اصطلاحات در GT. بنا به گفته‌ی برنان (۲۰۱۰)، مهم‌ترین وجه و ویژگی منحصربه‌فرد GT، چارچوب مفهومی آن است. لذا، برای ارتباط برقرار کردن با این نظریه و درک چارچوب مفهومی آن، مفاهیم و اصطلاحات رایج در آن جداگانه شرح داده می‌شود.

هدف اندازه‌گیری: هدف اندازه‌گیری عاملی است که محقق عمدتاً روی آن تمرکز می‌کند و تغییرپذیری میان آن‌ها مطلوب است. در بسیاری از وضعیت‌های اندازه‌گیری، افراد هدف اندازه‌گیری هستند. بااین‌حال، طبق اصل تقارن، هر یک از رویه‌های موجود در یک طرح را می‌توان به‌عنوان هدف اندازه‌گیری در نظر گرفت. کاردینت و همکاران (۲۰۱۰) مطرح کردند که واریانس حاصل از هدف اندازه‌گیری مترادف با مفهوم واریانس نمره‌ی واقعی در CTT است.

Archive of SID

www.SID.ir

رویه‌ها و موقعیت‌ها[1]: منابع بالقوه‌ی خطا در تعمیم دهی را رویه‌ها و سطوح[2] رویه‌ها را، موقعیت‌ها یا حالت‌ها می‌نامند. اصطلاحات رویه‌ها و موقعیت‌ها مشابه با عامل‌ها[3] و سطوح در ادبیات طرح‌های آزمایشی است (وب، رولی و شولسون، ۱۹۸۸؛ شولسون و وب، ۱۹۹۱). می‌توان گفت، یک‌رویه مجموعه‌ای از سطوح[4] مشابه‌اندازه‌گیری است (برنان، ۲۰۰۱) که محقق تصمیم دارد عملکرد آزمودنی‌ها را در این سطوح موردسنجش قرار دهد و سطوح یک اصطلاح کلی است که به فرم‌ها یا محرک‌های آزمونی خاص، مشاهده‌کننده‌ها، وضعیت‌ها یا شرایط مشاهده و غیره اشاره می‌کند (کرونباخ، راجاراتنام و گلیسر، ۱۹۶۳). جهان و جامعه[5]: در GT؛ واژه‌ی جهان به سطوح اندازه‌گیری اختصاص داده شده است، درحالی‌که واژه‌ی جامعه برای اهداف اندازه‌گیری استفاده می‌شود (برنان، ۲۰۰۱).

جهان مشاهدات قابل‌قبول[6]: جهان مشاهدات قابل‌قبول به‌وسیله همه ترکیبات ممکن از سطوح رویه‌ها تعریف می‌شود. یک اندازه رفتاری (مثل نمره آزمون) به‌عنوان نمونه‌ای از یک جهان مشاهدات قابل‌قبول تصور می‌شود که این جهان، شامل همه مشاهدات ممکن است که تصمیم‌گیرندگان جایگزین‌های قابل‌قبولی از آن برای مشاهده موجود (تحت مطالعه) در نظر می‌گیرند (وب و شولسون، ۲۰۰۵). جهان مشاهدات قابل‌قبول می‌تواند بزرگ‌تر از جهانی باشد که تصمیم‌گیرنده می‌خواهد به آن تعمیم دهد (شولسون و وب، ۱۹۸۱). طبق گفته‌ی کرونباخ و همکاران (۱۹۶۳)، جهان شامل تعداد محدود یا نامحدودی[7] از سطوح است که لازم است آن را به‌وضوح تعریف کرد، به‌گونه‌ای که سطوحی که در جهان قرار می‌گیرند مشخص باشد؛ اما ضروری نیست که جهان همگن باشد.

جهان تعمیم: عمدتاً محقق علاقه‌مند به تعمیم از یک مشاهده به دیگر مشاهداتی است که عضو جهان همانندی هستند. ازآنجاکه یک اندازه معین می‌تواند به‌طور قابل‌قبولی به

1. conditions
2. levels
3. factors

۴. در این مقاله، از واژه‌ی سطوح برای conditions استفاده کرده‌ایم.

5. universe & population
6. universe of admissible observations
7. universe of generalization

Archive of SID

www.SID.ir

جهان‌های متفاوت بسیاری تعمیم داده شود، لذا قبل از اینکه محقق مطالعه تعمیم‌پذیری را انجام دهد، باید جهانی که علاقه‌مند به تعمیم دهی به آن است را تعیین کند (کرونباخ و همکاران، ۱۹۶۳). جهان تعمیم، یعنی؛ جهانی که محقق قصد دارد نتایج یک روش اندازه‌گیری خاص را به آن تعمیم دهد (برنان، ۲۰۰۱) و می‌تواند با جهان مشاهدات قابل‌قبول یکسان و یا زیرمجموعه‌ای از رویه‌ها و سطوحشان در جهان مشاهدات قابل‌قبول باشد. جهان تعمیم را می‌توان از طریق؛ کاهش دادن جهان مشاهدات قابل‌قبول، یعنی کاهش سطوح یک‌رویه (برای مثال، تبدیل به یک‌رویه ثابت)، انتخاب کردن و درنتیجه کنترل کردن سطحی از رویه‌ای، یا با نادیده گرفتن رویه‌ای تعیین کرد (شولسون و وب، ۱۹۸۱).

نمره جهان: مارکولیدس[1] (۱۹۹۶) بیان می‌کند، مفهوم نمره‌ی جهان را می‌توان قلب GT در نظر گرفت. متأسفانه به سبب آنکه نمره‌ی جهان را فقط می‌توان برآورد کرد، انتخاب سؤال، فرم آزمون، یا موقعیت سنجش خاصی به‌طور اجتناب‌ناپذیری خطا را وارد روند اندازه‌گیری می‌کند. طبق بیان برنان (۲۰۱۰)، در اصل برای هر فرد (فرد هدف اندازه‌گیری در نظر گرفته شده است)، برای هر تکرار از روش اندازه‌گیری در جهان تعمیم می‌توان نمره‌ی میانگینی تصور کرد. برای چنین فردی، ارزش مورد انتظار از این نمرات میانگین به‌عنوان نمره جهان فرد تعریف می‌شود و واریانس نمرات جهان همه افراد در جامعه، واریانس نمره جهان نامیده می‌شود. در هر اندازه‌گیری سعی می‌شود شرایطی مهیا شود که واریانس نمره جهان زیاد و اندازه‌ی سایر مؤلفه‌های واریانس (واریانس‌های خطا) کم باشد. یک شخص می‌تواند نمرات جهان متفاوتی داشته باشد که بستگی به مشخصات جهان دارد. به بیان دقیق‌تر، نمره‌ی جهان تعریف نمی‌شود تا اینکه هدف اندازه‌گیری در مطالعه‌ی D مشخص شود (وب، شولسون و هرتل[2]، ۲۰۰۷).

رویه تصادفی[3]: GT اساساً نظریه‌ی اثرات تصادفی است. یک‌رویه تصادفی از طریق نمونه‌گیری سطوح یک‌رویه به‌طور تصادفی ایجاد می‌شود. حتی اگر سطوح یک‌رویه به‌طور

1. Marcoulides
2. Haertel
3. random facet

Archive of SID

www.SID.ir

تصادفی نمونه‌گیری نشده باشند، رویه ممکن است تصادفی در نظر گرفته شود اگر سطوح مشاهده نشده در مطالعه G قابل‌جابه‌جایی با سطوح مشاهده شده باشد (وب و همکاران، ۲۰۰۷).

رویه ثابت: سطوح یک‌رویه‌ی ثابت در مطالعه G تمام سطوح ممکن موردنظر است. رویه ثابت زمانی رخ می‌دهد که:

- تصمیم گیرنده با قصد قبلی سطوح خاصی را انتخاب کند و تمایلی به تعمیم دهی فراتر از آن‌ها را نداشته باشد.
- تعمیم دادن به ورای سطوح مشاهده‌شده غیرمنطقی باشد.
- هنگامی که کل جهان سطوح، کوچک باشد و طرح اندازه‌گیری همه سطوح را در برگیرد (شولسون و همکاران، ۱۹۸۹؛ وب و همکاران، ۲۰۰۷).

مدل‌های تصادفی و آمیخته[1] با جهان‌های تعمیم نامحدود و محدود: بنا به گفته‌ی شولسون و وب (۱۹۸۱، ۱۹۹۱)، برای معنی‌داری تحلیل‌ها، در هر تحلیل تعمیم‌پذیری حداقل باید یک‌رویه تصادفی باشد. ازاین‌رو، در GT مدل ثابت (مدلی که تمام رویه‌های آن ثابت باشد) وجود ندارد. در مدل تصادفی، همه‌ی رویه‌ها تصادفی هستند و این مدل‌ها با جهان‌های تعمیم نامحدود مرتبط‌اند. در مقابل در مدل آمیخته، ترکیبی از رویه‌های ثابت و تصادفی وجود دارد. تثبیت یک‌رویه، جهان تعمیم محدود را به دنبال خواهد داشت؛ بنابراین جهان تعمیم برای رویه ثابت محدودتر از جهان تعمیم برای رویه تصادفی است. جهان تعمیم محدود ازآنجاکه نسبت به جهان نامحدود کمتر مستعد خطاست، لذا واریانس نمره جهان و به‌تبع، ضریب تعمیم‌پذیری بزرگ‌تری دارد. همان‌طور که برنان (۲۰۱۰) متذکر می‌شود، نمی‌توان گفت که جهان محدود رجحان دارد، زیرا محدود کردن یک جهان همچنین گسترده‌ای که یک محقق می‌تواند به آن تعمیم دهد را محدود می‌کند. در اکثر زمینه‌های ارزیابی آموزشی و روان‌شناختی، رویه‌ها را باید به‌عنوان تصادفی در نظر گرفت تا بتوان نتایج

1. random & mixed models

Archive of SID

www.SID.ir

را به بیش از یک وضعیت اندازه‌گیری خاص تعمیم داد (بریچ، سوامیناتان، ولش و چفولز[1]، ۲۰۱۴).

رویه‌های متقاطع، آشیانه‌ای و درآمیخته[2]: دو شرط زیر را در نظر بگیرید؛ (الف) دو سطح یا بیشتر (سطوح چندگانه) از A با هر یک از سطوح B مشاهده شود و (ب) سطوح متفاوتی از A با هر یک از سطوح B مربوط باشد. اگر فقط شرط اول برقرار باشد اما سطوح یکسانی از A با هریک از سطوح B مربوط باشد، گفته می‌شود؛ دورویه متقاطع هستند. اگر هر دو شرط برقرار باشد، گفته می‌شود؛ A درون B آشیانه کرده است. درصورتی‌که فقط شرط دوم برقرار باشد، یعنی؛ یک سطح متفاوت از A برای هر یک از سطوح B رخ دهد، گفته می‌شود؛ A با B درآمیخته است (شولسون و وب، ۱۹۹۱). به‌عنوان‌مثال، هر یک از ارزیابان (r) مجموعه‌ی یکسانی از سؤالات (i) را نمره‌گذاری می‌کنند. در این صورت، ارتباط بین r و i متقاطع بوده و به‌صورت r × i نشان داده می‌شود. علامت ×«متقاطع با» خوانده می‌شود. اگر هر یک از ارزیابان (r) مجموعه‌ی متفاوتی از سؤالات (i) را نمره‌گذاری کنند، در این حالت گفته می‌شود؛ سؤالات در درون ارزیابان آشیانه کرده‌اند و به‌صورت r: i نمایش داده می‌شود. علامت: «آشیانه کرده درون» خوانده می‌شود. اگر هر یک از ارزیابان (r)، یک سؤال (i) متفاوت را نمره‌گذاری کنند، گفته می‌شود؛ رویه سؤالات با رویه ارزیابان درآمیخته است.

طرح‌های تعمیم‌پذیری[3]: عناصر یک طرح (طرح‌های مطالعه G یا (D، عبارت‌اند از: تعداد رویه‌ها، روش نمونه‌گیری که برای انتخاب سطوح رویه‌ها بکار گرفته می‌شود (ثابت در مقابل تصادفی) و ارتباط میان رویه‌ها (متقاطع در مقابل آشیانه‌ای). ترکیب این عناصر می‌تواند طرح‌های مختلف بسیاری را به وجود بیاورد. برای مثال، دورویه‌ای کاملاً متقاطع با رویه‌های تصادفی، سه رویه‌ای آشیانه‌ای با رویه‌های ثابت (براکنر، یودر و مک ویلیام[4]، ۲۰۰۶). طرح‌ها را می‌توان بر اساس تعداد رویه‌هایشان نام‌گذاری کرد. اگر فقط یک‌رویه در

1. Briesch, Swaminathan, Welsh & Chafouleas
2. confounded
3. generalizability designs
4. Bruckner,Yoder & McWilliam

Archive of SID

www.SID.ir

آزمون به کار گرفته شود، طرح آزمون را می‌توان طرح یک‌رویه‌ای[1] نامید. بنا به گفته‌ی برنان (۲۰۰۳)، برای طرح‌های یک‌رویه‌ای، جهان مشاهدات قابل‌قبول و جهان تعمیم شامل سطوحی از همان یک‌رویه می‌شود. اگر دورویه به‌صورت همزمان در آزمون به کار گرفته شود طرح آزمون را می‌توان طرح دورویه‌ای[2] دانست. همچنان که تعداد رویه‌ها (منابع بالقوه خطای اندازه‌گیری) افزایش می‌یابد، تعداد طرح‌های انتخابی در دسترس برای استفاده از نمره آزمون نیز افزایش می‌یابد. همچنین، طرح‌ها را می‌توان بر اساس روابط میان رویه‌ها نام‌گذاری کرد. طبق بیان شولسون و وب (۱۹۹۱)؛ نظریه تعمیم‌پذیری، طرح‌های کاملاً متقاطع یا متقاطع (همه‌ی رویه‌ها متقاطع هستند)، کاملاً آشیانه‌ای (همه‌ی رویه‌ها آشیانه‌ای هستند)، نسبتاً آشیانه‌ای (هر دورویه‌ی متقاطع و آشیانه‌ای وجود دارد) و طرح‌های درآمیخته را در خود جای داده است. در تقسیم‌بندی دیگر، طرح‌های تعمیم‌پذیری را می‌توان در دو گروه متعادل و نامتعادل[3] و نیز، تک متغیری و چندمتغیری جا داد.

انواع مطالعات در GT: GT دو نوع مطالعه را شامل می‌شود؛ مطالعه‌ی تعمیم‌پذیری (G) و مطالعه‌ی تصمیم (D). یک مطالعه G به‌منظور برآورد مؤلفه‌های واریانس زیربنایی یک فرآیند اندازه‌گیری از طریق تعریف کردن جهان مشاهدات قابل‌قبول طراحی می‌شود. در مطالعه D، تصمیم‌گیرنده از اطلاعات به‌دست‌آمده‌ی مطالعه G استفاده می‌کند تا یک اندازه‌گیری که خطا برای هدف خاصی در آن به حداقل برسد را طراحی کند. در طراحی یک مطالعه D، تصمیم‌گیرنده باید: (الف) جهان تعمیم را تعریف کند؛ یعنی، تعداد و گستره‌ی رویه‌هایی را که مایل است به آن تعمیم دهد. (ب) نوع تفسیر در نظر گرفته‌شده برای اندازه‌گیری را بیان کند؛ زیرا از نمره آزمون یکسانی به شیوه‌های متفاوتی می‌توان استفاده کرد. برای مثال، برخی از تفسیرها ممکن است روی تفاوت‌های فردی تمرکز کنند (تصمیمات نسبی)، برخی ممکن است نمره مشاهده‌شده را به‌عنوان برآوردی از نمره جهان شخص بکار گیرند (تصمیمات مطلق). درحالی‌که سایرین، نمره مشاهده‌شده را در یک

1. one facet design
2. two facet design
3. balanced & unbalanced

Archive of SID

www.SID.ir

برآورد رگرسیون از نمره جهان بکار ببرند. هریک از این تفاسیر مطرح‌شده با خطای اندازه‌گیری متفاوتی مرتبط هستند. (ج) از اطلاعات مطالعه G که در مورد اندازه‌ی منابع مختلف خطای اندازه‌گیری به دست آمده است، جهت بهینه‌سازی طرح اندازه‌گیری (کمینه‌سازی خطا و بیشینه‌سازی اعتبار) مطالعه D استفاده کند. این ارزشیابی تا حدی مشابه با فرمول پیشگویی اسپیرمن- بروان در CTT انجام می‌گیرد. به‌طوری‌که با افزایش تعداد سطوح یک‌رویه در یک اندازه‌گیری می‌توان سهم خطای متعلق به آن رویه را کاهش داد. البته در مطالعه D، علاوه بر تغییر در تعداد سطوح رویه‌ها، دست‌کاری رویه‌ها شامل تغییر در ماهیت (تصادفی/ ثابت) و روابط آن‌ها (متقاطع/ آشیانه‌ای) نیز می‌شود (شولسون و وب، ۱۹۸۱؛ شولسون و همکاران، ۱۹۸۹؛ شولسون و وب، ۱۹۹۱).

ملاحظات مطالعات G و D: تمایز قائل شدن میان مطالعات G و D مهم است، زیرا داده‌های تصمیم‌اند که اتکاپذیری‌شان مدنظر است. این دو نوع مطالعه می‌توانند متفاوت از هم باشند و طرح‌های تجربی متفاوتی داشته باشند (کرونباخ و همکاران، ۱۹۶۳). در مطالعات G می‌بایست تا جایی که امکان‌پذیر است از طرح‌های متقاطع استفاده کرد؛ زیرا با این طرح‌ها، امکان برآورد جداگانه‌ی همه‌ی مؤلفه‌های واریانس وجود دارد. درحالی‌که در دیگر طرح‌ها، اثر مستقیم مؤلفه واریانس مربوط به رویه‌ی آشیانه‌ای به‌صورت جداگانه برآورد نمی‌شود. در مواردی، رویه‌ها ذاتاً آشیانه‌ای هستند و محقق حق انتخاب ندارد؛ مانند؛ سؤالاتی که درون هر یک از فصل‌های یک کتاب درسی آشیانه کردند. بااین‌حال، اغلب در مطالعه D برای سهولت، افزایش اندازه نمونه، یا هر دو، طرح‌های آشیانه‌ای به کار گرفته می‌شود. اگر رویه‌های ثابت به کار روند و سطوح این رویه‌ها قابل تعویض نباشد، باید برای هر سطح، مطالعات G جداگانه‌ای انجام داد. حجم نمونه در مطالعه D ضرورت ندارد همانی باشد که در مطالعه G به کار رفته است. در این مطالعه تصمیم‌ها معمولاً مبتنی بر میانگین همه‌ی مشاهدات چندگانه (مانند سؤالات آزمون) است، درحالی‌که مطالعه G بر روی مشاهدات منفرد (مانند یک تک سؤالی) تمرکز می‌کند. انتخاب تعداد سطوح هر رویه در مطالعه D و همچنین انتخاب طرح (آشیانه‌ای در مقابل متقاطع، رویه ثابت در مقابل تصادفی)، ملاحظات منطقی و عملی و همچنین مسائل اتکاپذیری را در برمی‌گیرد (شولسون و وب،

Archive of SID

www.SID.ir

۱۹۸۱؛ شولسون و همکاران، ۱۹۸۹؛ وب و شولسون، ۲۰۰۵). در پاسخ به سؤال «چه هنگام مطالعات G و مطالعات D به کار می‌روند؟» براون (۲۰۰۵) بیان می‌کند که ابتدا مطالعه G باید انجام شود؛ سپس و تنها بعدازآن، یک مطالعه D کاربردی را دنبال کرد. انجام هر یک بدون دیگری کمتر معقول است و این دو مطالعه را باید با هم و به‌طور متوالی به کار بست.

برآورد مؤلفه‌های واریانس: یکی از اهداف GT، ارزیابی کردن منابع عمده‌ی تغییرپذیری است، به‌طوری‌که تغییرپذیری نامطلوب را بتوان در جمع‌آوری داده‌های آینده کاهش داد. GT، اندازه‌ی تغییرپذیری را برحسب مؤلفه‌های واریانس بیان می‌کند (شولسون و وب، ۱۹۹۱). برنان (۲۰۰۱)، هدف از مطالعه‌ی G را برآورد مؤلفه‌های واریانس برای جامعه و جهان مشاهدات قابل‌قبول عنوان می‌کند که این مؤلفه‌های واریانس برآورد شده را می‌توان در مطالعات مختلف D به‌منظور طراحی روش‌های اندازه‌گیری کارآمد برای استفاده‌ی عملیاتی و فراهم کردن اطلاعات برای گرفتن تصمیم‌های اساسی در مورد اهداف اندازه‌گیری به کار برد. هنگامی‌که فرض می‌شود جامعه و جهان مشاهدات قابل‌قبول به‌طور نامحدودی بزرگ هستند، مؤلفه‌های واریانس؛ مؤلفه‌های واریانس اثرات تصادفی[1] نامیده می‌شوند. در هر مطالعه تعداد متفاوتی از مؤلفه‌های واریانس به کار گرفته می‌شود که بستگی به طرح مطالعه و نوع تفسیر اندازه‌ها دارد (فن و سان، ۲۰۱۳). طبق بیان مارکولیدس (۱۹۹۶)؛ تمرکز GT روی این مؤلفه‌های واریانس است. اگرچه تحلیل واریانس سنتی رایج‌ترین شیوه‌ی استفاده‌شده برای برآورد مؤلفه‌های واریانس مطالعه‌ی G است، با روش‌های دیگری نیز می‌توان مؤلفه‌های واریانس را برآورد کرد؛ ازجمله، روش‌های بیزین، روش‌های واریانس حداقل[2]، برآورد بیشینه درستنمایی محدود[3] (RMLE) و رویکرد تحلیل ساختار کوواریانس[4]. موضوع مهمی که باید به آن توجه داشت، منفی بودن مؤلفه (های) واریانس برآورد شده است. این موضوع گرچه ازلحاظ نظری ناممکن است، اما در بعضی از وضعیت‌های اندازه‌گیری رخ می‌دهد. هنگامی‌که اندازه‌ی برآوردهای منفی بزرگ باشد،

1. random effects variance components
2. minimum variance methods
3. restricted maximum likelihood estimation
4. covariance structure analysis approach

Archive of SID

www.SID.ir

ممکن است که آن‌ها، نامعین بودن مدل اندازه‌گیری را منعکس کنند. در این صورت مدل مطالعه‌ی G باید مجدداً تعیین شود و مؤلفه‌های واریانس دوباره برآورد شوند. هنگامی‌که اندازه‌ی برآوردهای منفی کوچک (نزدیک به صفر) باشد، احتمالاً به سبب خطای نمونه‌گیری است که در این حالت می‌توان از روش‌های پیشنهادی استفاده کرد. کرونباخ و همکارانش، توصیه کردند که صفر جایگزین برآوردهای منفی شود. درصورتی‌که مؤلفه واریانس منفی در معادلات میانگین مجذورات مورد انتظار (EMS) مربوط به دیگر مؤلفه‌های واریانس ظاهر شد، مقدار صفر به‌جای برآورد منفی گذاشته شود و برآورد مؤلفه‌های واریانس جدید محاسبه شود. این روش برآوردهای همراه با سوگیری را پدید می‌آورد. برنان، توصیه کرد که صفر جایگزین برآوردهای منفی شود، اما هر جا که مؤلفه واریانس مذکور در معادلات میانگین مجذورات مورد انتظار دیگر مؤلفه‌ها ظاهر شد، برآورد منفی بکار گرفته شود. این فرآیند برآوردهای غیر سوگیری را فراهم می‌کند. بااین‌حال، به لحاظ آماری تغییر برآوردها به صفر رضایت‌بخش نیست. از دیگر توصیه‌ها، استفاده از دیگر رویکردهای برآورد مؤلفه‌های واریانس ازجمله، رویکرد بیزین و یا RMLE است (شولسون و وب، ۱۹۸۱؛ ۱۹۹۱).

انواع تصمیم و واریانس‌های خطا: دو نوع تصمیم وجود دارد که واریانس خطا و درنتیجه ضرایب، به‌طور متفاوتی برای هر یک از آن‌ها تعریف می‌شود. اولی، تصمیم نسبی است اگر تصمیم درباره افراد مبتنی بر جایگاهشان در ارتباط با دیگران باشد. این نوع تصمیم بر روی تفسیر هنجار مرجع نمره متمرکز است؛ یعنی نمرات اندازه‌گیری برای متمایز کردن آزمودنی‌ها به کار می‌روند. اعتبار اندازه‌گیری در این حالت مربوط به ثبات جایگاه نسبی افراد است نه در مورد ثبات نمرات واقعی. واریانس خطا برای تصمیم نسبی را با علامت σ_δ^2 نشان می‌دهند و آن را واریانس خطای نسبی می‌نامند. این نوع واریانس شامل همه‌ی مؤلفه‌های واریانس تعاملی است که هدف اندازه‌گیری را در برمی‌گیرد. واریانس خطای نسبی به‌صورت تفاوت میان نمره انحرافی مشاهده‌شده فرد[1] و نمره انحرافی جهان[2] او تعریف

1. person's observed deviation score
2. universe deviation score

Archive of SID

www.SID.ir

می‌شود. دومی، تصمیم مطلق است اگر تصمیم درباره افراد مبنی بر نمراتشان در ارتباط با یک ملاک باشد. به بیانی دیگر، تصمیم مطلق بر روی سطح عملکرد افراد بدون توجه به رتبه‌ی آن‌ها متمرکز است و در ارتباط بابام ثبات جایگاه نسبی افراد و هم ثبات نمرات واقعی است. واریانس خطا برای تصمیم مطلق را با علامت σ_{Δ}^2 نشان می‌دهند و آن را واریانس خطای مطلق[1] می‌نامند که شامل همه مؤلفه‌های واریانس مدل به‌جز هدف اندازه‌گیری است. این نوع واریانس، بیانگر تفاوت میان نمره مشاهده‌شده و نمره جهان فرد است. در کل، واریانس خطای نسبی کمتر از واریانس خطای مطلق است، زیرا شامل مؤلفه‌های واریانس کمتری است. این نشان می‌دهد که تفسیرهای نسبی در مورد نمرات افراد نسبت به تفسیرهای مطلق کمتر مستعد خطا هستند (وب و شولسون، ۲۰۰۵؛ برنان، ۲۰۱۰؛ فن و سان، ۲۰۱۳).

انواع ضرایب: ضریب اعتبار و محاسبه آن بستگی به مفهوم‌سازی خطای از پیش تعیین‌شده به‌عنوان مطلق یا نسبی دارد. به‌بیان‌دیگر، GT میان دو نوع ضریب اعتبار تمایز می‌گذارد: یکی، ضریب تعمیم‌پذیری است و زمانی به کار می‌رود که تصمیم‌ها نسبی هستند. آن را با علائم $E\rho^2$ یا ρ^2 نمایش می‌دهند و فرمول آن به‌صورت زیر است:

$$E\rho^2 = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_\delta^2}$$

دیگری، شاخص اتکاپذیری (φ) است و برای تصمیم‌های مطلق به کار می‌رود. این شاخص، بدین‌صورت فرمول‌بندی می‌شود:

$$\varphi = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_\Delta^2}$$

تفاوت ضریب تعمیم‌پذیری و شاخص اتکاپذیری در این است که اولی شامل واریانس خطای نسبی و دومی شامل واریانس خطای مطلق است؛ بنابراین شاخص اتکاپذیری عموماً کمتر از ضریب تعمیم‌پذیری است (وب و شولسون، ۲۰۰۵؛ برنان، ۲۰۰۱).

1. absolute error variance

Archive of SID

www.SID.ir

فرآیند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری. برای آشنایی با چگونگی انجام یک مطالعه اندازه‌گیری و رعایت مسائل کلیدی موردنیاز آن، در این بخش فرایند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری در قالب یک مثال با داده‌های فرضی گام‌به‌گام تشریح می‌شود. همچنین، برای افرادی که علاقه‌مند به درک نحوه‌ی محاسبات برآورد مؤلفه‌های واریانس، واریانس‌های خطا و همچنین، ضرایب تعمیم‌پذیری هستند؛ فرمول‌ها و چگونگی انجام محاسبات (مربوط به طرح مفروض) ارائه می‌گردد. مثال: فرض می‌شود، از ۱۰ دانش‌آموز خواسته شده است که به یک آزمون تشریحی ۵ سؤالی پاسخ دهند. هر یک از سؤالات توسط ۳ ارزیاب، نمره‌گذاری شده است (حجم نمونه برای مثال کوچک در نظر گرفته شده است). هر دانش‌آموز (s) به تمام سؤالات (i) پاسخ داده و هر ارزیاب (r) نیز، هر سؤال هر دانش‌آموز را نمره‌گذاری کرده است.

۱. طرح سؤال: با توجه به هدف و تصمیم محقق از انجام یک مطالعه‌ی اندازه‌گیری، سؤالات طرح می‌شوند. ازاین‌رو یک محقق ممکن است اتکاپذیری نمرات را بسنجد، محقق دیگر، تعمیم‌پذیری نمرات را مدنظر داشته باشد یا هر دو هدف مذکور، ممکن است مطلوب محقق دیگر باشد. با توجه به اصل تقارن، می‌توان سؤالاتی در زمینه‌ی اعتبار هر یک از رویه‌های موجود (در مثال، ارزیابان و سؤالات) مطرح کرد. همچنین، سؤالاتی نیز در ارتباط با بهینه‌سازی طرح مانند؛ تعداد سطوح لازم از رویه‌های موجود برای رسیدن به ضرایب اعتبار دلخواه را می‌توان بیان نمود. در اینجا، سعی می‌شود اتکاپذیری و تعمیم‌پذیری نمرات به‌دست‌آمده برای دانش‌آموزان و همچنین، تعداد کافی ارزیابان و سؤالات، برای رسیدن به ضرایب دلخواه بررسی گردد. بعد از طرح سؤال، فرایند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری در قالب دو مرحله‌ی کلی؛ الف: مطالعه‌ی G (شامل مراحل ۲ تا ۸) و ب: مطالعه‌ی D (شامل مراحل ۹ تا ۱۵) صورت می‌پذیرد.

الف: مطالعه‌ی G. بعد از تعیین جامعه و جهان مشاهدات قابل‌قبول، مطابق با طرح اندازه‌گیری در نظر گرفته‌شده، داده‌ها جمع‌آوری می‌شوند. سپس با استفاده از نرم‌افزار یا انجام محاسبات دستی، می‌توان هدف مطالعه‌ی G را که برآورد مؤلفه‌های واریانس برای جامعه و جهان

Archive of SID

www.SID.ir

مشاهدات قابل‌قبول است، تحقق بخشید و در ادامه به تفسیر مؤلفه‌های واریانس برآورد شده پرداخت.

۲. تعیین جامعه.

۲-۱. تعیین هدف اندازه‌گیری: همان‌طور که بیان شد، واژه جامعه برای اهداف اندازه‌گیری به کار می‌رود. هدف اندازه‌گیری در مثال ذکرشده، دانش‌آموزان (s) هستند که منبع خطا نیست، درنتیجه رویه حساب نمی‌شود، لذا تغییرپذیری میان دانش‌آموزان مطلوب است.

۲-۲. وضعیت نمونه‌گیری. در این گام فرعی، به تعداد اهداف اندازه‌گیری و روش نمونه‌گیری از آن‌ها اشاره می‌شود. در مثال، فرض شده است که $n_s=$ ۱۰ (تعداد دانش‌آموزان) نفر یک نمونه‌ی تصادفی از جامعه نامحدود دانش‌آموزان هستند.

۳. تعیین جهان مشاهدات قابل‌قبول. با انجام گام‌های فرعی زیر می‌توان جهان مشاهدات قابل‌قبول که شامل همه ترکیبات ممکن از سطوح رویه‌ها است (وب و شولسون، ۲۰۰۵) را به‌طور دقیق تعریف کرد.

۳-۱. شناسایی رویه‌ها. در هر بافت اندازه‌گیری، منابع بالقوه‌ی خطا به‌عنوان رویه در نظر گرفته می‌شوند. بنا به گفته‌ی برنان (۲۰۰۱)، GT می‌تواند هر تعداد از رویه‌ها را لحاظ کند. در مثال مطرح‌شده، جهان مشاهدات قابل‌قبول، شامل دورویه‌ی سؤال و ارزیاب است.

۳-۲. وضعیت نمونه‌گیری (تعداد سطوح و تصادفی/ ثابت). در این گام، به تعداد سطوح هر یک از رویه‌های موجود در جهان مشاهدات قابل‌قبول اشاره می‌شود. همچنین، باید تعیین شود که رویه‌ها متعلق به جهان‌های محدود هستند یا نامحدود. به بیانی دیگر، رویه‌ها تصادفی‌اند یا ثابت. طبق گفته‌ی برنان (۲۰۱۱b)، باید از هر یک از رویه‌ها حداقل دو سطح وجود داشته باشد. فرض شده است که $n_i=$ ۵ سؤال، یک نمونه تصادفی از جهان سؤالات مشابه ممکن هستند. همچنین فرض شده است که $n_r=$ ۳ ارزیاب، یک نمونه تصادفی از جهان ارزیابان مشابه ممکن هستند. یا می‌توان گفت؛ رویه‌های مذکور تصادفی در نظر گرفته می‌شوند، زیرا سطوح مشاهده نشده‌ی آن‌ها قابلیت جابه‌جایی با سطوح مشاهده‌شده در مطالعه اندازه‌گیری را دارند. مثال ما، فاقد رویه‌ی ثابت است.

Archive of SID

www.SID.ir

۳-۳. تعیین روابط. در این گام، روابط بین رویه‌ها در جهان مشاهدات قابل‌قبول و همچنین، روابط جامعه و جهان مشاهدات قابل‌قبول تعیین می‌شود. هر یک از سؤالات در جهان مشاهدات قابل‌قبول می‌توانند توسط هر یک از ارزیابان در جهان ارزیابی شوند. در این صورت گفته می‌شود که دورویه در جهان مشاهدات قابل‌قبول، متقاطع (i×r) هستند. چون هر دانش‌آموز در جامعه می‌تواند به هر سؤال در جهان پاسخ دهد که توسط هر ارزیاب در جهان ارزشیابی می‌شود، گفته می‌شود که جامعه و جهان مشاهدات قابل‌قبول متقاطع هستند که به‌صورت s×i×r نشان داده می‌شود.

۴. نام طرح اندازه‌گیری. در این گام، مشخصات طرح اندازه‌گیری در مطالعه G ذکر می‌شود. طرح اندازه‌گیری برای مثالمان، طرح دورویه‌ای متقاطع با رویه‌های تصادفی نام دارد و با نماد s×i×r نشان داده می‌شود.

۵. جمع‌آوری داده‌ها. محقق می‌بایست مطابق با طرح اندازه‌گیری مشخص‌شده، به جمع‌آوری داده‌ها بپردازد. البته گاهی محقق در وضعیت اندازه‌گیری انجام‌شده‌ای قرار می‌گیرد که می‌بایست با در نظر گرفتن ملاحظات منطقی و عملی برای داده‌های موجود، طرح اندازه‌گیری مناسبی را تهیه کند.

۶. انتخاب نرم‌افزار. محقق می‌تواند با انتخاب یک نرم‌افزار مناسب، نیاز به محاسبات دستی را مرتفع کرده و به روند تحلیل داده‌ها سرعت و دقت ببخشد. نکته‌ای که باید به آن توجه داشت، این است که نرم‌افزارهای موجود برای انجام تحلیل‌های تعمیم‌پذیری، قابلیت‌های متفاوتی دارند. برای مثال، می‌توان به توانایی آن‌ها در انجام مطالعات D، به‌کارگیری طرح‌های (متعادل- نامتعادل)، طرح‌های (تک متغیری- چند متغیری)، به‌کارگیری داده‌های گمشده[1]، نحوه‌ی ورود داده‌ها، سهولت کاربرد و ... اشاره کرد. در این قسمت، شرح مختصری از رایج‌ترین نرم‌افزارها بیان می‌شود. برنان (۲۰۰۱)، از سه برنامه نرم‌افزاری نام می‌برد که هرکدام ویژگی‌های متفاوتی دارند؛ GENOVA، urGENOVA و mGENOVA. دو برنامه‌ی اول، مختص به طرح‌های اندازه‌گیری تک متغیری و سومی

1. missing data

Archive of SID

www.SID.ir

مختص طرح‌های اندازه‌گیری چند متغیری است. GENOVA، برای برآورد مؤلفه‌های واریانس اثرات تصادفی طرح‌های متعادل تا حداکثر ۶ اثر (۵ رویه و ۱ هدف اندازه‌گیری) مناسب است. ورود داده‌ها به این نرم‌افزار به دو صورت (داده‌های خام یا میانگین مجذورات) امکان‌پذیر است. urGENOVA برای برآورد مؤلفه‌های واریانس اثرات تصادفی هر دو طرح متعادل و نامتعادل مناسب است. در نرم‌افزار مذکور، ورود داده‌ها تنها به‌صورت داده‌های خام امکان‌پذیر است. mGENOVA، نیز برای مجموعه‌ای از طرح‌های متعادل یا نامتعادل (با توجه به آشیانه‌ای بودن) بکار می‌رود. ورود داده‌ها در mGENOVA می‌تواند به دو شیوه (داده‌های خام یا ماتریس واریانس و کوواریانس) انجام گیرد. از دیگر نرم‌افزارهای موجود برای انجام تحلیل‌های تعمیم‌پذیری می‌توان از EDUG نام برد که مبتنی بر تحلیل واریانس است و برای طرح‌های متعادل تا حداکثر ۸ اثر (۷ رویه و ۱ هدف اندازه‌گیری) مناسب است. ورود داده‌ها به این نرم‌افزار به دو شیوه (داده‌های خام، مجموع مجذورات) انجام می‌گیرد. برخلاف دیگر نرم‌افزارها، قابلیت اجرای اصل تقارن با این نرم‌افزار وجود دارد و همچنین، هدف اندازه‌گیری را می‌توان به‌عنوان آشیانه شده در رویه‌های دیگر در نظر گرفت. در بین نرم‌افزارهای فوق‌الذکر، GENOVA و mGENOVA و EDUG قابلیت اجرای مطالعه‌ی D را دارند. باوجوداین، نرم‌افزارهای نامبرده در صورت کامل نبودن داده‌ها (وجود داده‌های گمشده) قادر به انجام تحلیل نیستند. لذا مشکل مقادیر گمشده، قبل از ورود داده‌ها به این نرم‌افزارها باید برطرف شود. سهولت دسترسی به هر چهار نرم‌افزار از ویژگی‌های مشترک آن‌هاست (برنان، ۲۰۰۱، a ۲۰۰۱؛ کاردینت و همکاران، ۲۰۱۰). همچنین، دو نرم‌افزار SPSS VARCOMP وSAS VARCOMP نیز از سوی محققان برای تحلیل‌های تعمیم‌پذیری به کار می‌رود (بریچ و همکاران، ۲۰۱۴).

۷. برآورد مؤلفه‌های واریانس مطالعه‌ی G. در طرح اندازه‌گیری مطالعه‌ی G یعنی؛ s×i×r، نمره کل مشاهده‌شده هر دانش‌آموز در یک سؤال که توسط یک ارزیاب داده شده است، برحسب مدل خطی زیر بیان می‌شود:

$x_{sir} = \mu$ میانگین اصلی

Archive of SID

www.SID.ir

$$
\begin{aligned}
&+ \mu_s - \mu \quad \text{اثر دانش آموز} \\
&+ \mu_i - \mu \quad \text{اثر سؤال} \\
&+ \mu_r - \mu \quad \text{اثر ارزیاب} \\
&+ \mu_{si} - \mu_s - \mu_i + \mu \quad \text{اثر دانش آموز} \times \text{سؤال} \\
&+ \mu_{sr} - \mu_s - \mu_r + \mu \quad \text{اثر دانش آموز} \times \text{ارزیاب} \\
&+ \mu_{ir} - \mu_i - \mu_r + \mu \quad \text{اثر سؤال} \times \text{ارزیاب} \\
&- \mu_{sr} - \mu_{ir} + \mu_s + \mu_i + \mu_r - \mu \quad \text{اثر باقیمانده} - \mu_{si} + x_{sir}
\end{aligned}
$$

نمره‌ی مشاهده‌شده، شامل یک اثر برای میانگین اصلی؛ سه اثر برای دانش‌آموز، سؤال و ارزیاب؛ سه اثر تعاملی دوطرفه (sr، si، ri) و اثر باقی‌مانده (تعامل سه‌طرفه sir به‌علاوه‌ی خطای تصادفی) است. طبق گفته‌ی برنان (۲۰۰۳)، همه‌ی اثرات (به‌غیراز μ) اثرات تصادفی نامیده می‌شوند؛ زیرا آن‌ها مرتبط با روش نمونه‌گیری تصادفی از جامعه و جهان هستند. همچنین، فرض می‌شود که همه‌ی اثرات در مدل، نا همبسته هستند. توزیع هر مؤلفه یا اثر به‌غیراز میانگین اصلی[1]، یک میانگین صفر و واریانس σ^2(بنام مؤلفه واریانس) دارد. فرض کنید که برای هر دانش‌آموز در جامعه، نمره‌ی میانگین (نمره مورد انتظار) در تمام سؤالات و همه‌ی ارزیابان در جهان مشاهدات قابل‌قبول به دست آمده است. واریانس این نمرات میانگین در جامعه‌ی دانش‌آموزان، واریانس نمره‌ی جهان (واریانس هدف اندازه‌گیری) نام دارد که آن را با نماد $\hat{\sigma}_s^2$ نشان می‌دهند. سایر مؤلفه‌های واریانس برآورد شده، واریانس خطا را تشکیل می‌دهند (شولسون و وب، ۱۹۹۱؛ برنان، ۲۰۰۱؛ وب و شولسون، ۲۰۰۵).

واریانس نمره کل مشاهده‌شده، در سراسر جامعه‌ی افراد و سطوح در جهان مشاهدات قابل‌قبول به هفت مؤلفه واریانس مستقل مطابق زیر تجزیه می‌شود:

$$\sigma^2_{(x_{sir})} = \sigma^2_s + \sigma^2_i + \sigma^2_r + \sigma^2_{si} + \sigma^2_{sr} + \sigma^2_{ir} + \sigma^2_{sir}$$

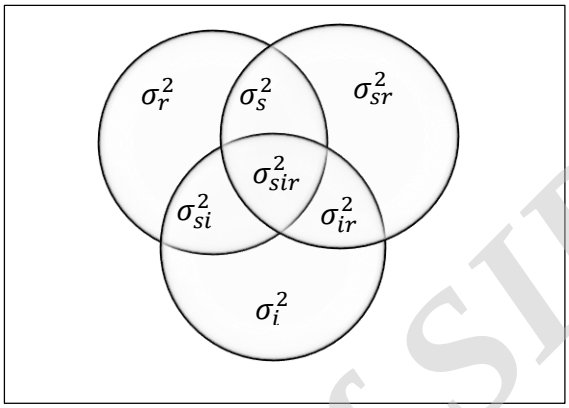
باید در نظر داشت که تعداد مؤلفه‌های واریانس، بستگی به طرح اندازه‌گیری دارد. تجزیه واریانس نمره‌ی کل مشاهده‌شده برحسب مؤلفه‌های واریانس تشکیل‌دهنده آن، از

1. grand mean

Archive of SID

www.SID.ir

طریق نمودار ون نشان داده شده است. لازم به ذکر است که اندازه‌ی دایره‌ها با نواحی درون آن‌ها، نشان‌دهنده‌ی اندازه‌ی مؤلفه‌های واریانس نیست (شولسون و وب، ۱۹۹۱).



نمودار ۱. مؤلفه‌های واریانس مربوط به طرح اندازه‌گیری s×i×r

مؤلفه‌های واریانس را می‌توان با استفاده از ترکیب نظری میانگین مجذورات که میانگین مجذورات مورد انتظار (EMS) نامیده می‌شود، برآورد کرد. بدین‌صورت که با استفاده از روش تحلیل واریانس (روش‌های مختلف دیگری نیز برای برآورد مؤلفه‌های واریانس وجود دارد که در این مقاله به آن‌ها اشاره شده است)، میانگین مجذورات (MS) برای هر مؤلفه محاسبه می‌شود که با قرار دادن آن در معادلات EMS مربوط به آن مؤلفه، مؤلفه واریانس برآورد شده[1] (EVC) به دست می‌آید.

در جدول شماره‌ی ۲، فرمول‌های مربوط به برآورد هر مؤلفه واریانس مطالعه G برای طرح اندازه‌گیری s×i×r ارائه شده است. هر طرح اندازه‌گیری، معادلات EMS مخصوص به خود را دارد.

2. Estimated variance components

Archive of SID

www.SID.ir

جدول ۲. فرمول‌های مربوط به برآورد مؤلفه‌های واریانس مطالعه **G** برای طرح اندازه‌گیری **s×i×r**

منبع تغییر	نوع تغییرپذیری	MS	EMS	EVC
s	نمره جهان	MS_s	$\sigma^2_{sir} + n_i\sigma^2_{sr} + n_r\sigma^2_{si} + n_i n_r\sigma^2_s$	$\hat{\sigma}^2_s = \frac{MS_s - MS_{si} - MS_{sr} + MS_{sir}}{n_i n_r}$
i	خطا	MS_i	$\sigma^2_{sir} + n_s\sigma^2_{ir} + n_r\sigma^2_{si} + n_s n_r\sigma^2_i$	$\hat{\sigma}^2_i = \frac{MS_i - MS_{si} - MS_{ir} + MS_{sir}}{n_s n_r}$
r	خطا	MS_r	$\sigma^2_{sir} + n_s\sigma^2_{ir} + n_i\sigma^2_{sr} + n_s n_i\sigma^2_r$	$\hat{\sigma}^2_r = \frac{MS_r - MS_{sr} - MS_{ir} + MS_{sir}}{n_s n_i}$
si	خطا	MS_{si}	$\sigma^2_{sir} + n_r\sigma^2_{si}$	$\hat{\sigma}^2_{si} = \frac{MS_{si} - MS_{sir}}{n_r}$
sr	خطا	MS_{sr}	$\sigma^2_{sir} + n_i\sigma^2_{sr}$	$\hat{\sigma}^2_{sr} = \frac{MS_{sr} - MS_{sir}}{n_i}$
ir	خطا	MS_{ir}	$\sigma^2_{sir} + n_s\sigma^2_{ir}$	$\hat{\sigma}^2_{ir} = \frac{MS_{ir} - MS_{sir}}{n_s}$
sir	خطا	MS_{sir}	σ^2_{sir}	$\hat{\sigma}^2_{sir} = MS_{sir}$

در جدول شماره‌ی **۳**، مؤلفه‌های واریانس برای هدف اندازه‌گیری، هر یک از رویه‌ها و همچنین، برای تعاملات موجود در طرح اندازه‌گیری s×i×r، برآورد شده است. آنچه در جدول برآوردها باید به آن توجه داشت و راهکار مناسبی برای آن در نظر گرفت، منفی بودن برآوردهاست. همان‌طور که در جدول پیداست، اندازه‌ی مؤلفه‌ی واریانس مربوط به اثر تعاملی دانش‌آموز با ارزیاب، منفی شده است. به سبب کوچک بودن این اندازه (نزدیک به صفر)، صفر را جایگزین آن می‌کنیم. این نکته حائز اهمیت است که این مؤلفه‌ها، برای نمرات هر فرد در یک سؤال و یک ارزیاب است.

جدول ۳. برآوردهای آنوا از مؤلفه‌های واریانس مطالعه‌ی G برای طرح اندازه‌گیری **s×i×r**

منبع تغییر	ss	df	MS	EVC	%	SE
s	۰/۴۶۲	۹	۰/۰۵۱	$\hat{\sigma}^2_s = \frac{0.051 - 0.012 - 0.005 + 0.0114}{5 \times 3} = 0.0030$	۱۶/۸۵	۰/۰۰۱۵
i	۰/۳۵۵	۴	۰/۰۸۹	$\hat{\sigma}^2_i = \frac{0.089 - 0.012 - .0.012 + 0.0114}{10 \times 3} = 0.0025$	۱۴/۰۴	۰/۰۰۱۷
r	۰/۰۷۰	۲	۰/۰۳۵	$\hat{\sigma}^2_r = \frac{0.035 - 0.005 - 0.012 + 0.0114}{10 \times 5} = 0.0006$	۳/۳۷	۰/۰۰۰۵
si	۰/۴۲۷	۳۶	۰/۰۱۲	$\hat{\sigma}^2_{si} = \frac{0.012 - 0.0114}{3} = 0.0002$	۱/۱۲	۰/۰۰۱۱
sr	۰/۰۹۶	۱۸	۰/۰۰۵	$\hat{\sigma}^2_{sr} = \frac{0.005 - 0.0114}{5} = -0.0013 = 0$	۰	۰/۰۰۰۵
ir	۰/۰۹۶	۸	۰/۰۱۲	$\hat{\sigma}^2_{ir} = \frac{0.012 - 0.0114}{10} = 0.0001$	۰/۵۶	۰/۰۰۰۶
sir	۰/۸۲۰	۷۲	۰/۰۱۱۴	$\hat{\sigma}^2_{sir} = 0.0114$	۶۴/۰۴	۰/۰۰۱۹

Archive of SID

www.SID.ir

۸ تفسیر مؤلفه‌های واریانس. مؤلفه‌های واریانس برآورد شده به مقیاس اندازه‌گیری بکار رفته، وابسته‌اند. ازاین‌رو، بر مبنای اندازه‌ی نسبی‌شان تفسیر می‌شوند. برای تفسیر کردن، می‌توان هر یک از مؤلفه‌ها را جداگانه تفسیر کرد و به سهم درصدی هر مؤلفه از واریانس کل (مجموع مؤلفه‌های واریانس) اشاره نمود. همچنین، با یک مقایسه‌ی نسبی مشخص کرد، کدام رویه‌ها و یا تعاملاتشان بیشترین سهم تغییرپذیری را به خود اختصاص داده‌اند. از دیگر شیوه‌های تفسیر، می‌توان به مقایسه‌های دوبه‌دوی مؤلفه‌ها و یا استفاده از انحراف استاندارد اشاره کرد (شولسون و وب، ۱۹۹۱؛ برنان، ۲۰۰۱). در مثال مطرح‌شده، مؤلفه واریانس باقی‌مانده $\hat{\sigma}^2_{sir}$، بزرگ‌ترین منبع تغییرپذیری است به‌طوری‌که ۶۴/۰۴ درصد از واریانس کل را به خود اختصاص داده است. این مؤلفه، نشان می‌دهد که سهم بزرگی از واریانس ناشی از؛ تعامل سه راهه بین دانش‌آموزان، سؤالات و ارزیابان و دیگر منابع تغییرپذیری منظم و تصادفی است که در مطالعه، اندازه‌گیری نشده‌اند. در مرتبه‌ی دوم، بیشترین سهم تغییرپذیری به مؤلفه واریانس برای دانش‌آموز یا واریانس نمره جهان $\hat{\sigma}^2_s$ اختصاص دارد که سهم آن از واریانس کل، ۱۶/۸۵ درصد است. این مؤلفه، تغییرپذیری منظم میان دانش‌آموزان را در توانایی اندازه‌گیری شده نشان می‌دهد که اندازه‌ی بزرگ آن مطلوب یک محقق است. در مرتبه‌ی سوم، سؤالات (با اختصاص دادن ۱۴/۰۴ درصد از واریانس کل) قرار دارند. مؤلفه واریانس سؤالات $\hat{\sigma}^2_i$، نسبتاً بزرگ است و نشان می‌دهد که سؤالات از بعد دشواری متفاوت هستند. مؤلفه واریانس کوچک برای تعامل دانش‌آموز با سؤال $\hat{\sigma}^2_{si}$، نشان می‌دهد جایگاه نسبی افراد از سؤالی به سؤال دیگر تفاوت چندانی نمی‌کند. مؤلفه‌ی واریانس مربوط به ارزیاب و همه‌ی مؤلفه‌های مرتبط با ارزیاب نسبتاً کوچک هستند. سهم نسبتاً کوچک مؤلفه واریانس ارزیاب $\hat{\sigma}^2_r$ از واریانس کل (۳/۳۷٪) نشان می‌دهد که ارزیابان در یک مقیاس سخت‌گیری- سهل گیری نسبتاً مشابه عمل کرده‌اند. به‌عبارتی‌دیگر، میانگین نمره‌گذاری‌هایشان نسبتاً نزدیک به هم بوده است. مطلوب است که سهم رویه ارزیاب از واریانس کوچک باشد. تعامل دانش‌آموز با ارزیاب $\hat{\sigma}^2_{sr}$، سهمی در تغییرپذیری نمرات دانش‌آموزان ندارد. بدین معناست که سخت‌گیری- سهل گیری یک ارزیاب در بین

Archive of SID

www.SID.ir

دانش‌آموزان ثبات داشته است. تعامل ارزیابان با سؤالات $\hat{\sigma}^2_{ir}$ (۰/۵۶٪ از واریانس کل) نشان می‌دهد که ارزیابان در نمره‌گذاری سؤالات مختلف آزمون تقریباً مشابه عمل کرده‌اند.

روش دیگر برای تفسیر، مقایسه‌های دوبه‌دوی مؤلفه‌هاست. برای مثال می‌توان گفت؛ واریانس برآورد شده ناشی از سؤالات تقریباً ۴ برابر واریانس ناشی از ارزیابان است که نشان می‌دهد سؤالات نسبت به ارزیابان به‌طور قابل‌ملاحظه‌ای منبع بزرگ‌تری از تغییرپذیری در نمرات دانش‌آموزان هستند. به‌بیان‌دیگر، تفاوت سؤالات در میانگین دشواری خیلی بیشتر از تفاوت ارزیابان در میانگین سخت‌گیری- سهل گیری است.

ب: مطالعه‌ی D. همان‌طور که برنان (۲۰۰۱) عنوان کرده است، این مطالعات بر برآورد، استفاده و تفسیر مؤلفه‌های واریانس برای تصمیم‌گیری در مورد اهداف اندازه‌گیری با روش‌های اندازه‌گیری مشخص‌شده‌ای تأکید می‌کند. بنا به گفته‌ی مارکولیدس (۱۹۹۶) نیز، از مؤلفه‌های واریانس برآورد شده در مطالعه G می‌توان جهت برآورد واریانس نمره جهان، واریانس‌های خطا و ضرایب اعتبار (تعمیم‌پذیری) برای جهان تعمیم و طرح‌های مطالعه D استفاده کرد.

۹. تعیین جهان تعمیم. مهم‌ترین مسئله‌ی مطالعه‌ی D، تعیین جهان تعمیم است. در این گام، مشخص می‌کنیم که جهان تعمیم (جهانی که علاقه‌مند به تعمیم دادن نتایج به آن هستیم) با جهان مشاهدات قابل‌قبول یکسان است و یا زیرمجموعه‌ای از رویه‌ها و سطوحشان در جهان مشاهدات قابل‌قبول است. در مورد اخیر، می‌بایست مشخصات جهان تعمیم به‌وضوح بیان شود. در مثال، جهان تعمیم شامل همه ارزیابان و سؤالات در جهان مشاهدات قابل‌قبول است. ازآنجاکه هر دورویه نامحدود (تصادفی) فرض شده‌اند، جهان تعمیم نیز نامحدود در نظر گرفته می‌شود.

۱۰. طرح اندازه‌گیری. مطالعه‌ی D می‌تواند طرح اندازه‌گیری مطالعه G و یا طرح‌های دیگری (در صورت امکان) را در نظر بگیرد. در مورد اخیر، می‌بایست طرح اندازه‌گیری مجدداً تعیین شود. مؤلفه‌های واریانس برآورده شده مطالعه‌ی G، بر روی نمرات واحد دانش‌آموز - سؤال - ارزیاب تمرکز می‌کند، درحالی‌که مطالعه‌ی D بر نمرات میانگین تمرکز می‌کند. به همین خاطر در ادبیات برنان، برای رویه‌های بکار رفته در طرح مطالعه‌ی D از حروف بزرگ

Archive of SID

www.SID.ir

استفاده می‌شود. برای مثالمان، طرح اندازه‌گیری مطالعه‌ی D همان طرح مطالعه G است که به‌صورت s × I × R نشان داده می‌شود. در این طرح نیز، افراد هدف اندازه‌گیری‌اند.

۱۱.برآورد مؤلفه‌های واریانس مطالعه‌ی D. همان‌طور که در جدول شماره ۴ آمده است، جهت محاسبه‌ی مؤلفه‌های واریانس مطالعه‌ی D، از مؤلفه‌های واریانس برآورد شده‌ی مرحله‌ی ۷ (مؤلفه‌های واریانس برآورد شده‌ی مطالعه‌ی G) بدین‌صورت استفاده می‌کنیم که: مؤلفه مطالعه G بر $\acute{n}_i$ تقسیم می‌شود، اگر مؤلفه شامل i باشد بی‌آنکه r را شامل شود. همچنین، بر $\acute{n}_r$ تقسیم می‌شود، اگر مؤلفه شامل r باشد بی‌آنکه i را شامل شود. بر $\acute{n}_i\acute{n}_r$ تقسیم می‌شود، اگر شامل هر دو r و i باشد. به‌عنوان‌مثال، برای برآورد مؤلفه واریانس ارزیاب ($\hat{\sigma}^2_R$)، مقدار برآورد شده‌ی این مؤلفه در مطالعه‌ی G ($\hat{\sigma}^2_r$= ۰/۰۰۰۶) تقسیم‌بر حجم نمونه ارزیاب در مطالعه D ($\acute{n}_r$= ۳) می‌شود. قابل‌ذکر است که این قاعده‌ی تقسیم‌بر حجم نمونه، برای مؤلفه هدف اندازه‌گیری (واریانس نمره جهان) به کار نمی‌رود. مقدار این مؤلفه در هر دو مطالعه برابر است. همان‌طور که گفته شد، حجم نمونه برای مطالعه‌ی D ($\acute{n}_r$ و $\acute{n}_i$) می‌تواند همان حجم نمونه در مطالعه‌ی G (n_r و n_i) باشد و یا متفاوت از آن باشد. در اینجا، همان حجم نمونه در مطالعه‌ی G در نظر گرفته شده است.

جدول ٤. معادلات و محاسبات مربوط به برآورد مؤلفه‌های واریانس مطالعه‌ی **D**

EVC در مطالعه‌ی D	EVC در مطالعه‌ی G
$\hat{\sigma}^2_s$= ۰/۰۰۳۰	$\hat{\sigma}^2_s$= ۰/۰۰۳۰
$\hat{\sigma}^2_I = \frac{\hat{\sigma}^2_i}{\acute{n}_i}$= ۰/۰۰۰۵	$\hat{\sigma}^2_i$= ۰/۰۰۲۵
$\frac{\hat{\sigma}^2_r}{\acute{n}_r}\hat{\sigma}^2_R$ =۰/۰۰۰۲	$\hat{\sigma}^2_r$= ۰/۰۰۰۶
$\hat{\sigma}^2_{sI} = \frac{\hat{\sigma}^2_{si}}{\acute{n}_i}$= ۰/۰۰۰۰۴	$\hat{\sigma}^2_{si}$= ۰/۰۰۰۲
$\hat{\sigma}^2_{sR} = \frac{\hat{\sigma}^2_{sr}}{\acute{n}_r}$ =۰	$\hat{\sigma}^2_{sr}$= −۰/۰۰۱۳ = ۰
$\hat{\sigma}^2_{IR} = \frac{\hat{\sigma}^2_{ir}}{\acute{n}_i\acute{n}_r}$ = ۰/۰۰۰۰۱	$\hat{\sigma}^2_{ir}$= ۰/۰۰۰۱
$\hat{\sigma}^2_{sIR} = \frac{\hat{\sigma}^2_{sir}}{\acute{n}_i\acute{n}_r}$ = ۰/۰۰۰۷۶	$\hat{\sigma}^2_{sir}$= ۰/۰۱۱۴
n_i= $\acute{n}_i$= ۵ n_r= $\acute{n}_r$= ۳	

۱۲.نوع تفسیر یا نوع تصمیم. در یک وضعیت اندازه‌گیری، نوع تفسیر نمره (هنجار مرجع در مقابل ملاک مرجع) تعیین می‌کند که کدام تصمیم (نسبی یا مطلق) مناسب است (فن و سان، ۲۰۱۳). در این گام، نوع تصمیم و یا نوع تفسیر باید مشخص شود. در تصمیم نسبی،

Archive of SID

www.SID.ir

تصمیم‌گیری درباره‌ی دانش‌آموزان، با توجه به جایگاهشان در ارتباط با عملکرد دیگر دانش‌آموزان انجام می‌گیرد. در تصمیم مطلق، تصمیم‌گیری بر اساس عملکرد هر دانش‌آموز بر اساس ملاک تعیین‌شده صورت می‌گیرد. در مثال، بر اساس گام ۱، هر دو نوع تصمیم در نظر گرفته می‌شوند.

۱۳. محاسبه واریانس‌های خطا و ضرایب مربوطه به همراه تفسیر. هر یک از دو تصمیم نسبی و مطلق، واریانس‌های خطا و ضرایب مخصوص به خود را دارند. در نمودار شماره‌ی ۲، تفاوت میان واریانس خطای نسبی و مطلق در طرح اندازه‌گیری s × I × R نشان داده شده است که قسمت‌های هاشور خورده‌ی آن، سهم واریانس خطا را تحت سطوح مختلف نشان می‌دهد. شکل الف، واریانس خطای نسبی σ_{δ}^{2} را نشان می‌دهد که شامل همه‌ی مؤلفه‌های واریانس تعاملی است که هدف اندازه‌گیری را در برمی‌گیرد. شکل ب، مربوط به واریانس خطای مطلق σ_{Δ}^{2} است که شامل همه مؤلفه‌های واریانس مدل به‌جز هدف اندازه‌گیری است.

الف: خطای نسبی ب: خطای مطلق

نمودار ۲. سهم واریانس خطای نسبی و مطلق در طرح دورویه‌ای متقاطع

در جدول شماره‌ی ۵، معادلات و محاسبات واریانس‌های خطا و ضرایب مربوطه برای طرح s × I × R آورده شده است.

Archive of SID

www.SID.ir

جدول ۵. معادلات و محاسبات مربوط به واریانس‌های خطا و ضرایب برای طرح s × I × R

واریانس خطای نسبی $\sigma^2_{(\delta)}$	$\hat{\sigma}^2_{\delta} = \hat{\sigma}^2_{sI} + \hat{\sigma}^2_{sR} + \hat{\sigma}^2_{sIR}$ =۰/۰۰۰۴ +۰+۰/۰۰۰۷۶ = ۰/۰۰۰۸
و ضریب تعمیم‌پذیری $(E\hat{\rho}^2)$	$E\hat{\rho}^2 = \frac{\hat{\sigma}^2_S}{\hat{\sigma}^2_S + \hat{\sigma}^2_{\delta}} = \frac{0.0030}{0.0030 + 0.0008} = 0.79$
واریانس خطای مطلق $\sigma^2_{(\Delta)}$	$\hat{\sigma}^2_{\Delta} = \hat{\sigma}^2_I + \hat{\sigma}^2_R + \hat{\sigma}^2_{sI} + \hat{\sigma}^2_{sR} + \hat{\sigma}^2_{IR} + \hat{\sigma}^2_{sIR}$ = ۰/۰۰۰۵+۰/۰۰۰۲ +۰/۰۰۰۴ +۰+۰/۰۰۰۱ + ۰/۰۰۷۶ = ۰/۰۰۱۵
و شاخص اتکاپذیری (φ)	$\hat{\varphi} = \frac{\hat{\sigma}^2_S}{\hat{\sigma}^2_S + \hat{\sigma}^2_{\Delta}} = \frac{0.0030}{0.0030 + 0.0015} = 0.67$

تفسیر: همان‌طور که وب و همکاران (۲۰۰۷) مطرح کرده‌اند، برای تصمیم‌گیری در مورد افراد مبتنی بر نمرات مشاهده‌شده‌شان، ضریب اعتبار ۰/۸۰ و بالاتر غالباً به‌قدر کافی معتبر تلقی می‌شود و درصورتی که تصمیمات، پیامدهای چشمگیری داشته باشند، مقادیر ۰/۹۰ به بالاتر ترجیح داده می‌شود. با توجه به اینکه اندازه‌ی هر دو نوع ضریب، کمتر از حداقل سطح مطلوب (۰/۸۰) است، می‌توان نتیجه گرفت؛ اندازه‌گیری نمرات دانش‌آموزان از اعتبار کافی برخوردار نیست. به‌عبارتی‌دیگر، ضریب تعمیم‌پذیری (۰/۷۹) نشان می‌دهد که امکان تفکیک معتبر دانش‌آموزان ازنظر توانایی اندازه‌گیری شده وجود ندارد. همچنین، شاخص اتکاپذیری (۰/۶۷) نشان می‌دهد که مشخص کردن جایگاه یک دانش‌آموز در مقیاس توانایی اندازه‌گیری شده به‌صورت معتبر، امکان‌پذیر نیست.

۱۴. خطای استاندارد اندازه‌گیری و فواصل اطمینان. در استانداردهای سنجش تربیتی و روان-شناختی آمده است که همه‌ی برآوردهای اعتبار باید همراه خطاهای استاندارد اندازه‌گیری (SEM) و فواصل اطمینان باشد (بریچ و همکاران، ۲۰۱۴). SEM، کاربر را در مورد اندازه خطایی که نتایج را در بافت اندازه‌گیری نسبی و مطلق متأثر می‌سازد، مطلع می‌کند (کاردینت و همکاران، ۲۰۱۰). اگر فرض کنیم خطاها به‌طور طبیعی توزیع شده‌اند، با گرفتن ریشه دوم واریانس‌های خطای نسبی و مطلق به ترتیب SEM نسبی و مطلق به دست می‌آید. در مثال مفروض، برای داده‌هایی که در مقیاس ۰/۵-۰ بودند، ۰/۰۲۸۲ = SEM نسبی و ۰/۰۳۸۷ = SEM مطلق به دست آمد. هر چه مقدار SEM (نسبی یا مطلق) بزرگ‌تر باشد، دقت اندازه‌گیری کاهش می‌یابد و فاصله اطمینان نیز بزرگ‌تر خواهد بود. فواصل اطمینان رایج،

Archive of SID

www.SID.ir

۹۵٪ و ۹۹٪ هستند که به ترتیب با نمرات Z، ۱/۹۶ و ۲/۵۸ مرتبط هستند. برای محاسبه فاصله اطمینان دلخواه، مقدار SEM به‌دست‌آمده را در فرمول زیر جایگزین می‌کنیم.

$$\bar{X} \pm (Z \times SEM)$$

۱۵.بهینه‌سازی طرح اندازه‌گیری. محقق در این گام می‌تواند، با دست‌کاری رویه‌ها (مانند؛ تغییر در تعداد سطوح، ماهیت و روابط رویه‌ها) خطا را در طرح اندازه‌گیری کاهش دهد و اعتبار را به‌اندازه‌ی دلخواه برساند. در جدول شماره ۶، ضرایب تعمیم‌پذیری و اتکاپذیری برای چند طرح (با تغییر در تعداد سطوح رویه‌ها در طرح اندازه‌گیری $s \times I \times R$) محاسبه شده است. همچنین، در جدول مذکور چگونگی استفاده از اطلاعات به‌دست‌آمده در مطالعه‌ی G برای به‌کارگیری در طرح‌های مختلف مطالعه‌ی D نشان داده شده است.

۱-۱۵. تغییر تعداد سطوح رویه‌ها. یکی از سؤالات مطرح‌شده در گام اول، بررسی تعداد سطوح لازم از ارزیابان و سؤالات برای رسیدن به ضرایب اعتبار دلخواه بود. همان‌طور که در قسمت تفسیر مؤلفه‌های واریانس مطالعه G گفته شد، سؤالات نسبت به ارزیابان، منبع بزرگ‌تری از تغییرپذیری در نمرات دانش‌آموزان هستند؛ بنابراین، افزودن تعداد سطوح سؤالات، واریانس خطا را بسیار بیشتر از افزودن ارزیابان کاهش می‌دهد که نتیجه‌ی آن، افزایش ضرایب اعتبار خواهد بود. با توجه به جدول شماره‌ی ۶، اگر تعداد ارزیابان را به یک نفر کاهش داده و تعداد سطوح سؤالات را ۲، ۳ و ۴ برابر کنیم، ضرایب تعمیم‌پذیری به‌دست‌آمده به ترتیب برابر با ۰/۷۳، ۰/۸۰ و ۰/۸۴ است. بااین‌حال، ضرایب اتکاپذیری در هر سه مورد همچنان پایین‌تر از حداقل سطح مطلوب (۰/۸۰) است. اگر تعداد ارزیابان موجود در طرح ۲ نفر باشد و تعداد سطوح سؤالات را ۲، ۳ و ۴ برابر کنیم، ضرایب تعمیم‌پذیری به‌دست‌آمده به ترتیب برابر ۰/۸۴، ۰/۸۸ و ۰/۹۱ خواهد شد و ضرایب اتکاپذیری نیز، به ترتیب عبارت‌اند از؛ ۰/۷۲، ۰/۷۸ و ۰/۸۱. اگر تعداد سطوح ارزیابان ثابت نگه داشته شود و تعداد سطوح سؤالات به ۲، ۳ و ۴ برابر افزایش یابد، ضرایب تعمیم‌پذیری از ۰/۷۹ به ترتیب به ۰/۸۸، ۰/۹۲ و ۰/۹۴ و ضرایب اتکاپذیری نیز از ۰/۶۷ به ۰/۷۸، ۰/۸۲ و ۰/۸۵ خواهد رسید. تنها در حالت‌های (۲ ارزیاب و ۲۰ سؤال، ۳ ارزیاب و ۱۵ سؤال، ۳ ارزیاب و ۲۰ سؤال) هر دو ضرایب اعتبار با هم از حداقل سطح مطلوب و بالاتر از آن برخوردارند. ازآنجاکه طرح

Archive of SID

www.SID.ir

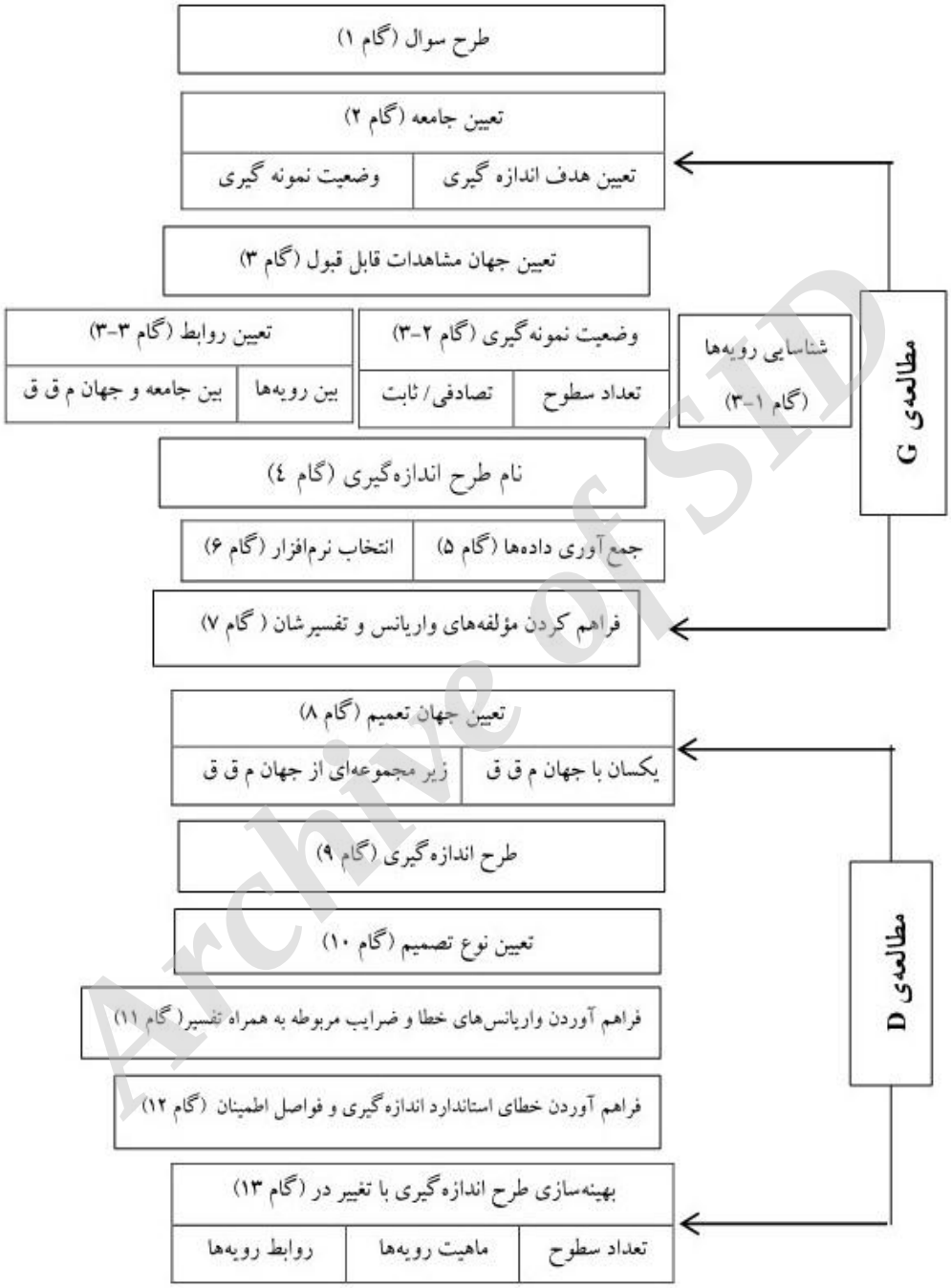
سؤال و روند ارزیابی مستلزم صرف هزینه و زمان است، با انجام این گام می‌توان ترکیب مناسبی از سؤالات و ارزیابان را با توجه به‌اندازه اعتبار موردنظر و دیگر محدودیت‌های عملی، به دست آورد. بنا به گفته‌ی ون لیون (۱۹۹۷)، به لحاظ عملی، اجرا کردن سؤالات بیشتر که توسط یک ارزیاب نمره‌گذاری می‌شوند به‌صرفه‌تر از داشتن تعداد کمی سؤال با ارزیابان چندگانه است. ازاین‌رو، مناسب‌ترین طرح را می‌توان داشتن ۲ ارزیاب و ۲۰ سؤال دانست.

۲-۱۵. تغییر در ماهیت و روابط رویه‌ها. محقق در این گام فرعی می‌تواند با توجه به هدف و سؤالات در نظر گرفته‌شده و همچنین ملاحظات عملی و منطقی، ماهیت رویه‌ها (رویه ارزیاب یا رویه‌ی سؤال را ثابت در نظر بگیرد) یا روابط آن‌ها (متقاطع/ آشیانه‌ای) را تغییر دهد و طرح‌های متنوع و متفاوتی را بکار بگیرد.

لازم به ذکر است که استفاده از نرم‌افزار نیاز به محاسبات دستی را مرتفع می‌کند که در این صورت، بعضی از مراحل ذکرشده در قسمت پیشین حذف می‌شوند. در این قسمت، فرآیند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری که به‌طور مفصل تشریح شد، در قالب نمودار شماره‌ی ۳ نشان داده شده است (البته با حذف مراحل مربوط به محاسبات دستی).

Archive of SID

www.SID.ir



نمودار ۳: فرآیند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری

Archive of SID

www.SID.ir

جدول ٦. مؤلفه‌های واریانس و ضرایب تعمیم‌پذیری برآورد شده برای طرح‌های مختلف مطالعه‌ی **D**

مطالعه‌ی G		مطالعه‌ی D								
n_r= ۱	n'_r	۱			۲			۳		
n_i= ۱	n'_i	۱۰	۱۵	۲۰	۱۰	۱۵	۲۰	۱۰	۱۵	۲۰
$\hat{\sigma}^2_s$= ۰/۰۰۳۰	$\hat{\sigma}^2_s$	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰
$\hat{\sigma}^2_i$= ۰/۰۰۲۵	$\hat{\sigma}^2_I$	۰/۰۰۰۲۵	۰/۰۰۰۱۷	۰/۰۰۰۱۲۵	۰/۰۰۰۲۵	۰/۰۰۰۱۷	۰/۰۰۰۱۲۵	۰/۰۰۰۲۵	۰/۰۰۰۱۷	۰/۰۰۰۱۲۵
$\hat{\sigma}^2_r$= ۰/۰۰۰۶	$\hat{\sigma}^2_R$	۰/۰۰۰۶	۰/۰۰۰۶	۰/۰۰۰۶	۰/۰۰۰۳	۰/۰۰۰۳	۰/۰۰۰۳	۰/۰۰۰۲	۰/۰۰۰۲	۰/۰۰۰۲
$\hat{\sigma}^2_{si}$= ۰/۰۰۰۲	$\hat{\sigma}^2_{sI}$	۰/۰۰۰۰۲	۰/۰۰۰۰۱	۰/۰۰۰۰۱	۰/۰۰۰۰۲	۰/۰۰۰۰۱	۰/۰۰۰۰۱	۰/۰۰۰۰۲	۰/۰۰۰۰۱۳	۰/۰۰۰۰۱
$\hat{\sigma}^2_{sr}$= ۰	$\hat{\sigma}^2_{sR}$	۰	۰	۰	۰	۰	۰	۰	۰	۰
$\hat{\sigma}^2_{ir}$= ۰/۰۰۰۱	$\hat{\sigma}^2_{IR}$	۰/۰۰۰۰۱	۰/۰۰۰۰۱	۰/۰۰۰۰۰۵	۰/۰۰۰۰۰۵	۰/۰۰۰۰۰۳	۰/۰۰۰۰۰۲۵	۰/۰۰۰۰۰۳	۰/۰۰۰۰۰۲	۰/۰۰۰۰۰۲
$\hat{\sigma}^2_{sir}$= ۰/۰۱۱۴	$\hat{\sigma}^2_{sIR}$	۰/۰۰۱۱	۰/۰۰۰۷۶	۰/۰۰۰۵۷	۰/۰۰۰۵۷	۰/۰۰۰۳۸	۰/۰۰۰۲۸۵	۰/۰۰۰۳۸	۰/۰۰۰۲۵۳	۰/۰۰۰۱۹
	$\hat{\sigma}^2_s$	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰	۰/۰۰۳۰
	$\sigma^2_{(\delta)}$	۰/۰۰۱۱۲	۰/۰۰۰۷۷	۰/۰۰۰۵۸	۰/۰۰۰۵۹	۰/۰۰۰۳۹	۰/۰۰۰۲۹۵	۰/۰۰۰۴	۰/۰۰۰۲۶۶	۰/۰۰۰۲
	$\sigma^2_{(\Delta)}$	۰/۰۰۱۹۸	۰/۰۰۱۵۵	۰/۰۰۱۳۱	۰/۰۰۱۱۴۵	۰/۰۰۰۸۶۳	۰/۰۰۰۷۲۲	۰/۰۰۰۸۵۳	۰/۰۰۰۶۳۸	۰/۰۰۰۵۲۷
	$E\rho^2$	۰/۷۳	۰/۸۰	۰/۸۴	۰/۸۴	۰/۸۸	۰/۹۱	۰/۸۸	۰/۹۲	۰/۹۴
	φ	۰/۶۰	۰/۶۶	۰/۷۰	۰/۷۲	۰/۷۸	۰/۸۱	۰/۷۸	۰/۸۲	۰/۸۵

بحث و نتیجه‌گیری

درحالی‌که از مدت‌ها پیش اکثر متخصصان اندازه‌گیری برای سنجش اعتبار، استفاده از GT را بجای CTT توصیه کرده‌اند (مشکواش و اکانر[1]، ۲۰۰۶)، محققان کشورمان صرفاً به CTT متکی هستند. مهم‌ترین دلیل آن را می‌توان، عدم آشنایی محققان و طراحان آزمون با مبانی نظری و قابلیت‌های GT دانست. ازاین‌رو، در مقاله حاضر، مبانی و مفاهیم GT تک متغیری (ازجمله؛ جهان مشاهدات قابل‌قبول، جهان تعمیم، نقش مطالعات G و D، برآورد مؤلفه‌های واریانس، انواع طرح‌های تعمیم‌پذیری و انواع ضرایب) به‌سادگی و به‌دوراز ابهام معرفی و شرح داده شد. علاوه بر این، به برخی از شباهت‌ها و تفاوت‌های این نظریه با CTT اشاره گردید تا نقاط قوت و سودمندی GT آشکار گردد.

همان‌طور که می‌دانیم، یکی از چالش‌های عمده‌ای که هر وضعیت اندازه‌گیری با آن روبه‌روست، شناسایی و کاهش خطاهای اندازه‌گیری است. طبق بیان همبلتون و جونز[2] (۱۹۹۳)، نظریه یا مدل مناسب به اینکه چگونه در یک وضعیت اندازه‌گیری سهم خطا را می‌توان به حداقل رساند، کمک می‌کند. برخلاف CTT، GT این قابلیت را دارد که منابع

1. Mushquash & O'connor
2. Hambleton & Jones

Archive of SID

www.SID.ir

چندگانه خطای اندازه‌گیری را همزمان شناسایی، تفکیک و هر یک را برآورد کند. از این طریق می‌توان منبع (ها) خطایی که مقدار مؤلفه واریانس آن و یا مقدار مؤلفه‌های واریانس اثرهای تعاملی آن زیاد است را شناسایی کرد و در ادامه با راهکارهای بهینه‌سازی، مقدار واریانس ناشی از آن را کاهش داد و فرآیند اندازه‌گیری را بهبود بخشید. همان‌طور که فن و سان (۲۰۱۳) بیان نموده‌اند؛ قابلیت این نظریه در انجام طرح‌های مختلف اندازه‌گیری مطالعات D، به محققان این امکان را می‌دهد که اندازه اعتبار را در مطالعه واقعی‌شان پیش‌بینی کنند. چارچوب مفهومی محکم GT، تمایز گذاشتن میان تصمیم‌های نسبی و مطلق و به‌کارگیری اصل تقارن از دیگر نقاط قوت این نظریه است.

همچنین، در این مقاله تلاش شد که چگونگی انجام یک مطالعه اندازه‌گیری جهت سنجش اعتبار داده‌ها در قالب یک مثال نشان داده شود. در GT، طراحی دقیق یک مطالعه اندازه‌گیری، اجرا، تحلیل و تفسیر نتایج آن، انواع مختلفی از مسائل و ملاحظات را در برمی‌گیرد که محققان هنگام استفاده از این نظریه می‌بایست آن‌ها را در نظر داشته باشند تا بتوانند از سودمندی این نظریه در راستای سنجش اعتبار داده‌هایشان بهره‌ی کافی را ببرند. بدین منظور، فرآیند طراحی، تحلیل و تفسیر یک مطالعه اندازه‌گیری به‌صورت گام‌به‌گام در قالب مطالعات G و D توصیف شد. علاوه بر این، معادلات و محاسبات مربوطه نیز برای علاقه‌مندان تشریح گردید که به تسهیل درک آن‌ها از این نظریه کمک می‌کند.

لازم به ذکر است که برای برآورد اعتبار، از طرح‌های مختلفی می‌توان استفاده کرد که در این مقاله تنها به یکی از رایج‌ترین آن‌ها (طرح دورویه‌ای متقاطع با رویه‌های تصادفی) پرداخته شد. همچنین، بااینکه مثالمان در حوزه‌ی آموزش‌وپرورش بود، برای محققان سایر حوزه‌ها ازجمله روان‌شناختی، پزشکی، بازاریابی، مدیریت و ... نیز مفید خواهد بود. در صورت استفاده از این طرح، محققان می‌توانند رویه‌های موجود در وضعیت‌های اندازه‌گیری‌شان را جایگزین رویه‌های به‌کاررفته در این مثال کنند و اعتبار داده‌هایشان را به دست بیاورند. کمبود منابع از مشکلات جدی محققان و علاقه‌مندان به کار در حیطه‌ی GT در کشورمان است. با انتشار مقالات، ترجمه و تألیف کتاب در زمینه‌ی GT ازیک‌طرف و قرار گرفتن GT جزو برنامه آموزشی رشته‌ی سنجش و اندازه‌گیری و دیگر رشته‌های مرتبط

Archive of SID

www.SID.ir

از طرف دیگر، می‌توان انتظار داشت که در آینده قابلیت‌ها و کاربردهای این نظریه از سوی محققان کشورمان به کار گرفته شود.

منابع

کیامنش، علیرضا (۱۳۷۴). نظریه‌ی تعمیم‌پذیری در اندازه‌گیری آموزشی، نشریه‌ی علوم تربیتی، ۱۶(۲-۱): ۴۸-۲۵.

Alkharusi, H. (2012). *Generalizability Theory: An Analysis of Variance Approach to Measurement Problems in Educational Assessment*, Journal of Studies in Education, 2(1):184-196.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning*, *32*(2): 245-258.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4): 27-34.

Brennan, R.L. (1997). A perspective on the history of generalizability theory. *Educational Measurement*: *Issues and Practice, 16*: 14-20.

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer-Verlag.

Brennan, R. L. (2001a). mGENOVA [Computer software and manual]. Iowa City, IA: University of Iowa.

Brennan, R. L. (2003). Coefficients and indices in generalizability theory. *Center for Advanced Studies in Measurement and Assessment, CASMA Research Report, 1*: 1-44.

Brennan, R. L. (2010). Generalizability Theory. New York: *Springer-Verlag*.

Brennan, R.L. (2011a). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1): 1-21.

Brennan, R. L. (2011b). Using Generalizability Theory to Address Reliability Issues for PARCC Assessments: A White Paper. *Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa*.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of school psychology*, *52*(1): 13-35.

Brown, J. D. (2005). Statistics corner, questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *9*(1): 12-16.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language

Archive of SID

www.SID.ir

samples. *Journal of Early Intervention, 28*(2): 139-153. Cardinet, J., Johnson, S., & Pini, G. (2010). Applying Generalizability theory using Edug. Taylor & Francis Group.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13: 119–135.

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

Fan, X., & Sun, S. (2013). Generalizability theory as a unifying framework measurement of reliability in adolescent research. *The Journal of Early Adolescence*, 0272431613482044.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on. *Educational measurement: issues and practice*, *12*(3): 38-47.

Kumazawa, T. (2009). Revision of a criterion-referenced vocabulary test using generalizability theory .*JALT journal*, *31*(1): 81–100.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: the covariance structure analysis approach. *Structural equation modeling*, 3(3): 290-299.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior research methods*, *38*(3): 542-547.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44: 922–932.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973 – 1980. *British Journal of Mathematical and Statistical Psychology*, *34*:133-166.

Suen, H. K., & Lei, P. W. (2007). Classical versus Generalizability theory of measurement. *Educational Measurement*, *4*: 1-13.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: An application to inter-rater reliability. *Journal of Agricultural Education*, *38*:36-42.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, 21: 81–90.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. *Handbook of statistics*, *26*(4): 81-124.

Webb, N.M., & Shavelson, R.J. (2005).Generalizability theory: Overview. *Encyclopedia of statistics in behavioral science*, 2, 717-719.

Archive of SID

www.SID.ir