روانسنجی ابزارهای سنجش سلامت (۲): بررسی روایی سازه و ملاکی، پایایی و قابلیت پاسخ‌گویی بـه تغییرات

عباس عبادی[۱]، زیبا نقی زاده[۲]، علی منتظری[۳]، زهرا شاهواری[۴]، محمود طاووسی[۳]، راضیه باقرزاده[۵*]

۱. مرکز تحقیقات علوم رفتاری، دانشکده پرستاری دانشگاه بقیه الله، تهران، ایران
۲. مرکز تحقیقات مراقبت‌های پرستاری و مامایی، دانشکده پرستاری و مامایی دانشگاه علوم پزشکی تهران، تهران، ایران
۳. مرکز تحقیقات سنجش سلامت، پژوهشکده علوم بهداشتی جهاد دانشگاهی، تهران، ایران
۴. دفتر تعهد حرفه‌ای، دانشگاه علوم پزشکی تهران، ایران
۵. دانشکده مامایی دانشکده پرستاری و مامایی، دانشگاه علوم پزشکی بوشهر، بوشهر، ایران

نشریه پایش
تاریخ پذیرش مقاله: ۱۳۹۵/۱۱/۱۳
سال شانزدهم، شماره چهارم، مرداد ـ شهریور ۱۳۹۶ صص ۴۵۵-۴۴۵
[نشر الکترونیک پیش از انتشار- ۲۴ اردیبهشت ۹۶]

چکیده

مقدمه: چگونگی روانسنجی یک ابزار نشان دهنده کیفیت آن ابزار است. اغلب ابزارهای طراحی شده، همه ملاک‌های روانسنجی را مد نظر قرار نداده‌اند.

مواد و روش کار: حاضر با بررسی متون موجود فارسی و انکلیسی به شرح بررسی روایی سازه، پایایی و قابلیت پاسخگویی به عنوان خصوصیات مهم روانسنجی ابزارهای سنجش سلامت پرداخته است. همچنین اثر سقف و کف و تفسیرپذیری نیز شرح داده شده است.

یافته ها: برای تأمین روایی سازه نیز روش‌های مختلفی وجود دارد که شامل روایی ساختاری، آزمون فرضیه و نیز، روایی همگرا، افتراقی و بین فرهنگی است. روایی ملاکی نیز بخشی از روایی است که در صورت قابلیت انجام به ارتقای کیفیت ابزار کمک می‌کند. طراحان ابزار نباید به انجام تحلیل عاملی به تنهایی بسنده کنند و باید دیگر روش‌های تأمین روایی سازه را نیز برای افزایش اعتبار ابزار خود استفاده نمایند. پایایی بخش مهمی از روانسنجی است که شامل همبستگی درونی، ثبات، هم‌ارزی و خطای استاندارد ابزار است. طراحان ابزار اغلب به همبستگی درونی و گاه به بررسی ثبات پرداخته‌اند. بررسی خطای استاندارد ابزار به عنوان جزء مهمی از پایایی در اغلب ابزارهای طراحی شده مورد غفلت واقع شده است. قابلیت پاسخگویی در صورتی تأیید می شود که کمترین تغییر قابل اهمیت از لحاظ کلینیکی از کمترین تغییر قابل تشخیص توسط ابزار بیشتر باشد. همچنین از سطح زیر محنی ROC نیز برای بررسی قابلیت پاسخگویی می‌توان استفاده کرد. اثر سقف و کف در صورتی وجود دارد که بیش از ۱۵٪ پاسخگویان به ترتیب حداکثر یا حداقل نمره قابل دستیابی را کسب کنند.

بحث و نتیجه گیری: آشنایی با اصول ترجمه و طراحی ابزار و بررسی معیارهای روانسنجی می‌تواند به پژوهشگران کمک کند تا به ابزاری معتبر برای پژوهش خود دست یابند.

کلیدواژه: روانسنجی، روایی سازه، پایایی، قابلیت پاسخگویی، ابزار سنجش سلامت

* نویسنده پاسخگو: بوشهر، دانشگاه علوم پزشکی بوشهر، دانشکده پرستاری و مامایی، گروه پرستاری و مامایی
تلفن: ۰۷۷۳۳۴۵۰۱۸۷
E-mail: r.bagherzadeh@bpums.ac.ir

مقدمه

طراحی یا انتخاب ابزار برای پژوهش، نیازمند توجه ویژه به معیارهای روانسنجی است. ابزارهای طراحی شده در زمینه سلامت، همه ملاک‌های روانسنجی را مد نظر قرار نداده‌اند. برخی از معیارهای روانسنجی بیش از بقیه مورد غفلت واقع شده‌اند [۱].

اغلب ابزارهای طراحی شده در کشور به اعتبار (validity) و پایایی (Reliability) پرداخته‌اند. آن‌ها در انجام پایایی بیشتر به همبستگی درونی (Internal consistency) و گاهاً ثبات (Stability) بسنده کرده‌اند و دیگر جنبه‌های پایایی مثل خطای استاندارد ابزار (Standard error of measurement - SEM) و یا تکرارپذیری (Reproducibility) را مد نظر قرار نداده‌اند [۴-۲]. در مقاله قبلی فرایند ترجمه و همچنین طراحی ابزار شرح داده شده و اشاره گردید که بخش های اصلی روانسنجی شامل روایی، پایایی و قابلیت پاسخگویی (Responsiveness) هستند؛ و نیز در خصوص روایی محتوا (Content validity) و صوری (Face validity) به طور کامل شرح داده شد [۵]. در این مقاله جنبه‌های دیگر روانسنجی مثل روایی سازه (Construct validity)، پایایی و قابلیت پاسخگویی به تغییرات، شرح داده شده و به برخی از ویژگی‌های روانسنجی که بر انجام آن‌ها اتفاق نظر وجود ندارد؛ اما برخی از محققان معتقدند که بررسی این ویژگی‌ها کیفیت ابزار را افزایش می‌دهند، پرداخته شده است. این ویژگی‌ها شامل بررسی توافق و بررسی اثر سقف و کف (Ceiling and floor effect) و تفسیر پذیری (Interpretability) هستند.

۱- روایی سازه: قبل از انجام نمونه‌گیری برای روایی سازه باید همبستگی درونی پرسشنامه بررسی شود؛ که برای محاسبه آن معمولاً یک مطالعه پایلوت انجام می‌شود. همبستگی درونی نشان دهنده میزان همبستگی گویه‌های یک مقیاس/زیرمقیاس با هم است [۶]. همبستگی درونی برای ابزارهایی که گویه‌های هر مقیاس و یا زیرمقیاس با هم یک سازه را می‌سنجند قابل بررسی است. در ابزارهایی که هر گویه یک جنبه از یک پدیده پیچیده کلینیکی را می‌سنجد (مثل نمره آپگار)، همبستگی درونی لازم نیست. پس از تأیید همبستگی درونی گویه‌های ابزار، تحلیل عامل اکتشافی (Exploratory factor analysis) قابل انجام است تا مشخص شود آیا ابزار یک مقیاس واحد است یا متشکل از چند بعد (حیطه یا زیرمقیاس) است. سپس این بعدگذاری با تحلیل عاملی تأییدی (Confirmatory factor analysis) مورد تأیید قرار می‌گیرد [۷]. پس از مشخص شدن زیرمقیاس‌ها، محاسبه همبستگی درونی برای هر زیرمقیاس نیز ضروری است. همبستگی درونی بین ۰/۷ تا ۰/۹ مناسب است [۸]. البته حد بالایی قابل قبول برای همبستگی درونی توسط برخی محققان ۰/۹۵ نیز ذکر گردیده است [۷]. همبستگی درونی در بخش پایایی به طور مبسوط تر شرح داده شده است. برخی محققان معتقدند که قبل از تحلیل عاملی باید هموژنیتی (Homogeneity) کنترل شود. همبستگی درونی به معنای هموژنیتی نیست. همبستگی درونی شرط لازم برای هموژنیتی است اما شرط کافی نیست. یعنی تا تمام گویه‌ها با هم و با کل مقیاس همبستگی نداشته باشند ابزار هموژن نخواهد بود. برای بررسی هموژنیتی علاوه بر محاسبه همبستگی درونی گویه‌ها، باید همبستگی هر گویه به کل نیز سنجیده شود و گویه‌هایی که همبستگی کمتر از ۰/۳ با کل مقیاس دارند باید حذف شوند [۹]. همچنین باید یک ماتریس همبستگی بین گویه‌ها گرفته شود. چنانچه یک گویه ضریب همبستگی بیش از ۰/۳ با حداقل یک گویه دیگر از پرسشنامه نداشته باشد حذف می‌شود [۱۰] همچنین چنانچه ضریب همبستگی بین دو گویه بیش از ۰/۷ بود یکی از آن گویه‌ها حذف می‌شود [۱۱]. روایی سازه با تحلیل عاملی، آزمون فرضیه (Hypothesis test)، روایی همگرا (Convergent validity) و افتراقی (Discriminant validity) و روایی بین فرهنگی (Cross-cultural validity) قابل تأمین است. والتز و همکاران روایی گروه‌های شناخته شده (Known groups) و آزمون فرضیه را به عنوان دو روش متفاوت جهت تأمین روایی سازه بیان کرده اند [۱۲]. اما تروی و همکاران این دو را یکی دانسته‌اند [۷].

تحلیل عاملی به معنی تعیین ابعاد پرسشنامه در واقع بخشی از روایی سازه است که موکینک و همکاران آن را روایی ساختاری می‌گویند [۱۳]. اگر هیچ پیش فرضی در مورد ابعاد تشکیل دهنده ابزار نداشته باشیم تحلیل مؤلفه‌های اصلی (PrincipalComponent Analysis) یا تحلیل عامل اکتشافی انجام می‌شود ولی اگر پیش فرض پذیرفته شده قبلی در مورد ابعاد داشته باشیم باید تحلیل عامل تأییدی انجام داد [۱۴].

در واقع محقق تحلیل عامل اکتشافی در مورد روانسنجی هیچ چیز نمی داند. این تحلیل فقط گویه های مرتبط را به صورت خوشه جدا می‌کند. این خوشه ممکن است به دلائلی غیر از یکسان بودن از

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

لحاظ مفهوم، مثلاً به دلیل فرمت یا نوع گویه با هم مرتبط باشند. بنابراین فردی که تحلیل عامل اکتشافی را انجام می‌دهد باید در نظر داشته باشد که این خوشه‌های تشکیل شده با مفهوم منطبق هستند یا خیر. دیگر این‌که باید از چند گروه نمونه مستقل استفاده شود و برای هر گروه تحلیل عاملی انجام شود [۱۵]. در صورت حذف ، اضافه یا تغییر گویه، بهتر است یک نمونه‌گیری جدید انجام شود؛ یعنی یک نمونه از گروه هدف که مستقل از نمونه‌های قبل باشد گرفته شود و ویرایش جدید پرسشنامه بین آن‌ها توزیع شود و دوباره تحلیل عاملی انجام شود [۱۶]. تعیین تعداد نمونه برای انجام تحلیل عاملی ضروری است. برخی از محققین تعداد ۲۰۰ تا ۳۰۰ نمونه را کافی می‌دانند [۹]. کاستلو و همکاران بهترین روش برای تعیین حجم نمونه را نسبت نمونه به گویه می‌دانند. آن‌ها عقیده دارند که بهتر است به ازای هر گویه ابزار ۱۰ تا ۲۰ نمونه گرفته شود [۱۷]. روش نمونه‌گیری باید به صورت دقیق شرح داده شود.

در تحلیل عامل اکتشافی و روش مؤلفه‌های اصلی آماره کایزر مایر الکین (Kaiser-Meyer-Olkin- KMO) برای بررسی کفایت نمونه محاسبه می‌شود؛ که میزان ۰/۸ و بیشتر مناسب در نظر گرفته می‌شود [۱۰]. آزمون کرویت بارتلت (Bartlett's Test of Sphericity) برای مناسب بودن الگوی تحلیل عاملی انجام می‌شود. معنی‌دار بودن این تست به این معنی است که ماتریس همبستگی بین گویه‌ها تأیید شده و مدل تحلیل عاملی مناسب است. معیار دترمینان (Determinant) نیز برای بررسی هم‌خطی چند گانه بکار گرفته می‌شود که اگر بیشتر از ۰/۰۰۰۰۱ باشد، هم خطی چندگانه وجود ندارد. نکته دیگر هنگام انجام تحلیل عاملی این است که میزان اشتراکات قابل قبول چقدر در نظر گرفته شده است. همچنین میزان واریانس تبیین شده توسط هر عامل و مجموع واریانس تبیین شده باید ذکر شود [۱۴].

مشخص کردن روش تعیین تعداد عوامل استخراج شده لازم است. اغلب مطالعات از ارزش ویژه (Eigenvalues) بیشتر از یک و آزمون سنگریزه (Scree test) برای تعیین تعداد عواملی که استخراج می‌شود استفاده کرده‌اند. اما روش‌های دیگر و دقیق‌تری نیز برای تعیین تعداد عواملی که باید استخراج شود وجود دارد: تحلیل موازی (Parallel analysis) و یا MAP (Minimum average partial method). متأسفانه استفاده از این روش‌ها به علت عدم امکان استفاده مستقیم از نرم افزار SPSS رایج نیستند. شرح روش‌های مختلف استخراج عوامل از اهداف این مقاله نیست. این روش‌ها در مقاله هایتون و همکاران [۱۱] و همچنین مقاله لدسما و والرو مورا به تفصیل شرح داده شده‌اند [۱۸].

هنگام انجام تحلیل عامل اکتشافی یا تحلیل مؤلفه‌های اصلی نوع چرخش باید مشخص شود. چرخش برای ساده کردن و واضح کردن ساختار انجام می‌شود. چرخش‌ها دو دسته کلی متعامد (شامل سه نوع Varimax، quartimax و equamax) و غیر متعامد یا مورب (شامل سه نوع direct oblimin، promax و quartimin) هستند [۱۷]. چرخش‌های متعامد عواملی ایجاد می‌کنند که با هم همبستگی ندارند ولی چرخش مایل به عوامل اجازه می‌دهند که با هم همبستگی داشته باشند [۱۷, ۱۴].

میزان مقادیر گمشده (Missing values) و مدیریت آن در هنگام انجام تحلیل عاملی مهم بوده و باید گزارش شود [۱۴]. ممکن است برای مدیریت مقادیر گمشده هنگام جمع‌آوری نمونه‌ها اقدام شود؛ مثلاً با کنترل هر پرسشنامه و درخواست از شرکت کننده برای پر کردن گویه‌هایی که پاسخ داده نشده‌اند؛ همچنین ممکن است هنگام تحلیل آماری مقادیر گمشده مدیریت شوند؛ مثلاً جایگزین کردن مقادیر گمشده با میانگین یا حذف مشاهده‌هایی که در یکی از گویه‌های خود مقدار گمشده دارند.

در آزمون فرضیه، روایی سازه به وسیله آزمون کردن فرضیه‌های مطرح شده (مثلاً بررسی همبستگی یا تفاوت بین گروه‌های شناخته شده، یا ارتباط نمرات این ابزار با دیگر ابزارهای مرتبط) بررسی می‌شود. فرضیه‌های مطرح شده باید ویژه باشند؛ در غیر این صورت نتایج با تورش همراه خواهند بود [۷]. برای این که فرضیه درستی مطرح و سپس آزمون شود می‌توان از نتایج مطالعات پیشین و فراتحلیل استفاده کرد. آزمون فرضیه در صورتی نشان دهنده روایی سازه است که ۷۵ درصد از نتایج با فرضیه‌های مطرح شده در نمونه‌ای از حداقل ۵۰ نفر برای هر زیرگروه، هم‌خوانی داشته باشند [۷]. به عنوان مثال فراتحلیل ها نشان داده‌اند که رضایت جنسی یک پیش‌بینی کننده رضایت از زندگی است. چنانچه یک ابزار برای سنجش رضایت جنسی طراحی کرده‌ایم یک فرضیه قابل آزمون که به تأمین روایی سازه کمک می کند، بررسی ارتباط بین نمرات پرسشنامه طراحی شده با نمرات پرسشنامه رضایت از زندگی است. شاهواری و همکاران به عنوان یکی از روش‌های تأمین روایی سازه پرسشنامه رضایت جنسی فرضیه "رضایت جنسی در زنان با میزان تحصیلات مرتبط است" را آزمون نمودند. آنها تفاوت میانگین نمره رضایت جنسی را در گروه‌های مختلف تحصیلی مقایسه

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

نمودند. نتایج نشان داد که رضایت جنسی در افراد با تحصیلات پایین نسبت به افراد با تحصیلات بالاتر، کمتر است. این نویسندگان آزمون این فرضیه را تحت عنوان روایی گروه‌های شناخته شده گزارش نمودند [۱۹].

روایی همگرا نیز جزئی از روایی سازه است. روایی همگرا به این معنی است که اندازه‌گیری‌های متفاوت با یک سازه باید نتایج مشابه داشته باشند [۱۲]؛ یعنی این که وقتی گروه هدف، ابزار طراحی شده و ابزاری که سازه مشابه را می‌سنجد تکمیل کنند، بین نمرات این دو باید همبستگی وجود داشته باشد. وقتی ابزار جدید با ابزار موجود، که برای اندازه گیری همان سازه ساخته شده است، دارای همبستگی باشد، آزمون جدید نیز برای اندازه‌گیری آن سازه رواست. روش چند ویژگی- چند روش (Multi trait - multi method approach) و استفاده از متوسط واریانس تبیین شده از جمله روش‌های موجود برای تأمین روایی همگرا است [۲۰].

روایی افتراقی نیز از جمله معیارهای روانسنجی ابزار و روشی برای تأمین روایی سازه است. روایی افتراقی نیز با چند روش ؛ ازجمله تحلیل عامل تأییدی و استفاده از تفاوت کای دو، روش چند ویژگی- چند روش یا استفاده از متوسط واریانس تبیین شده قابل بررسی است [۲۱]. روایی همگرا و افتراقی در این مقاله شرح داده نشده است.

۲- روایی ملاکی: این روایی نشان می‌دهد که ابزار ساخته شده تا چه حد با یک استاندارد بیرونی مرتبط است [۲۲]. اگر یک استاندارد بیرونی وجود داشته باشد، باید روایی ملاکی بررسی شود و همبستگی حداقل ۰/۷ بین ابزار و استاندارد بیرونی قابل قبول است [۷] ریو و همکاران برای بررسی روایی ملاکی فرم کوتاه پرسشنامه خلق و احساس (برای بررسی سندرم افسردگی کودکان) از برنامه مصاحبه تشخیصی در کودکان و تشخیص افسردگی آنان به عنوان یک معیار استاندارد استفاده کردند [۲۳].

۳- پایایی: پایایی شامل همبستگی یا همسانی درونی، هم‌ارزی، ثبات و خطای استاندارد ابزار است. البته برخی محققان معتقدند که همبستگی درونی جزء پایایی نیست و ثبات به اضافه خطای استاندارد ابزار اجزاء پایایی را تشکیل می‌دهند [۷]. آن‌ها معتقدند که همبستگی درونی روشی برای اثبات تک عاملی بودن ابزار است و نشان می دهد همه گویه‌های ابزار با هم یک مفهوم را می‌سنجند [۲۰]. همچنین موکینک و همکاران بررسی توافق (Agreement) را پیشنهاد می‌کنند و معتقدند ثبات و توافق دو مفهوم متفاوتند که در کنار هم تکرارپذیری را تشکیل می‌دهند [۱۳].

وقتی آزمون را یک بار اجرا می کنیم یا وقتی یک آزمون در یک موقعیت مشابه توسط چند ارزیاب نمره‌دهی می‌شود (Inter-rater reliability) در واقع همبستگی درونی بررسی می‌شود. روش دو نیمه کردن (Split-half) نیز نوعی بررسی همبستگی درونی است [۲۴]. شایعترین روش بررسی همبستگی درونی محاسبه آلفای کرونباخ (Cronbach's alpha) است. در ابزارهای لیکرتی چند گزینه‌ای برای مشخص شدن همبستگی درونی، آزمون آلفای کرونباخ و برای ابزارهایی که پاسخ آنها دو گزینه‌ای است، معمولاً آزمون کودر-ریچاردسون (Kuder–Richardson) ۲۰ یا ۲۱ انجام می‌شود [۲۵, ۱۲]. اگر سازه از وجوه مختلف تشکیل شده باشد یعنی مقیاس دارای چند زیرمقیاس باشد، آلفای کرونباخ علاوه بر کل مقیاس باید برای تمام زیرمقیاس‌ها محاسبه شود. میزان پایایی به دست آمده باید با توجه به فرمت گویه‌ها و خصوصیات متریک که روی نتایج نهایی آلفای کرونباخ تأثیر می‌گذارد قضاوت شوند. مثلاً میزان دشواری گویه می‌تواند روی پایایی تأثیر بگذارد. گاهی گویه‌ها یا پرسش‌ها آنقدر عمومی هستند، که پاسخ‌های مشابهی دارند و این خود یک پایایی را ایجاد می کند که به دلیل فرمت پرسشنامه است و منتج از مفهوم اصلی نیست [۱۵]. همبستگی درونی تحت تأثیر تعداد گویه است. برخی محققان معتقدند با افزایش گویه‌ها، نرسیدن به میزان بالای پایایی به ندرت اتفاق می‌افتد و با ابزار بالای ۴۰ گویه، استفاده از آلفای کرونباخ برای تأیید پایایی عملاً بی فایده است. همبستگی درونی بین ۰/۷ تا ۰/۹ مناسب است [۸]. حد بالایی قابل قبول برای همبستگی درونی توسط برخی محققان ۰/۹۵ نیز ذکر گردیده است [۷]. آلفای حدود ۰/۹۵ یا بیشتر به جای اینکه نشانه پایایی ضعیف باشد در واقع نشان دهنده روایی ناکافی است [۹].

وقتی دو یا بیش از دو ارزیاب یک گروه را ارزیابی کرده و به آنها نمره می دهند (Inter-rater)، همبستگی بین نمرات ارزیابان محاسبه می‌شود. اگر ارزیاب‌ها دو نفر باشند و داده‌ها فاصله‌ای بوده یا بتوان آن را فاصله‌ای فرض کرد، از ضریب همبستگی درون‌خوشه ای یا درون رده‌ای (Intra class correlation coefficient- ICC) استفاده می‌شود. ولی اگر تعداد ارزیاب‌ها بیشتر از دو نفر باشد ضریب آلفا حساب می شود. وقتی می خواهیم ضریب آلفا محاسبه کنیم جدولی تشکیل می دهیم که ستون‌ها قضاوت ارزیابان

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

و سطرها موضوعات را نشان دهد. در مطالعه هادیان و همکاران برای ارزیابی پایایی مقیاس برگ (مقیاس تعادلی) دو نفر تراپیست بر اساس معیارهای مقیاس برگ، کودکان را بررسی و به آنها نمره دادند. تراپیست دوم از نمره تراپیست اول بی اطلاع بود. تراپیست اول بر اساس معیار برگ کودکان را ارزیابی و به آنها نمره می‌داد، فرد سومی نمره‌ها را جمع‌آوری می‌کرد بعد از ۱۵ دقیقه استراحت تراپیست دوم نمره می‌داد. در انتها بین دو نمره ICC گرفته شد [۲۶].

در مقیاس‌های اسمی برای بررسی پایایی بین ارزیابان از آزمون کاپای کوهن (Cohen's kappa) استفاده می‌شود. این آزمون در مقیاس‌های اسمی دو حالته و چند حالته و همچنین مقیاس رتبه‌ای کاربرد دارد [۲۷]. البته می توان از محاسبه درصد توافق هم استفاده کرد، اما ضریب کاپا روش مناسب‌تری است چون اثر توافق شانسی را هم لحاظ می کند. استفاده از آزمون کاپا معمولاً در تشخیص‌های کلینیکی کاربرد دارد. مینیک و همکاران برای بررسی پایایی آزمون غربالگری حرکات عملکردی از پایایی بین ارزیابان و آزمون کاپا استفاده نمودند [۲۸]. کاپای ۰ تا ۰/۲ عدم توافق، ۰/۲۱ تا ۰/۳۹ حداقل توافق، ۰/۴ تا ۰/۵۹ توافق ضعیف، ۰/۶ تا ۰/۷۹توافق متوسط، ۰/۸ تا ۰/۹ توافق قوی و بالاتر از ۰/۹ توافق عالی محسوب می‌شود [۲۹].

استفاده از فرم‌های موازی روشی برای بررسی هم‌ارزی به عنوان پایایی است. استفاده از این روش مشکل است؛ چون یافتن فرم موازی عملاً مشکل است. دو ابزار در صورتی موازی فرض می‌شوند که: ۱- سازه مشابهی را روی افراد مشابه با فرایند مشابه بسنجند. ۲- تقریباً میانگین یکسان داشته باشند. ۳- انحراف معیار مساوی داشته باشند. ۴- همبستگی آنها با یک متغیر سوم یکسان باشد [۱۲].

برای بررسی پایایی با استفاده از فرم‌های موازی، نمونه‌ها به طور تصادفی به دو دسته تقسیم می شوند. یک گروه ابتدا فرم یک و با فاصله کوتاهی فرم ۲ را پر می کنند و گروه دیگر اول فرم ۲ و بعد با فاصله کمی فرم یک را پر می کنند. سپس همبستگی بین این دو سنجیده می‌شود؛ که به آن ضریب هم‌ارزی گویند. ضریب هم‌ارزی بالاتر یعنی این‌که دو فرم می توانند به جای هم استفاده شوند. ضریب همبستگی ۰/۸ تا ۰/۹ مناسب در نظر گرفته می شود. گاهی دو فرم به همین ترتیب ولی با فاصله‌ای مشابه آزمون –بازآزمون (Test-retest) تکمیل می‌شود که در واقع در اینجا علاوه بر هم‌ارزی ثبات هم بررسی شده؛ که به همبستگی به دست آمده ضریب هم‌ارزی و ثبات گویند [۳۰].

ثبات با توجه به نوع ابزار به دو روش شامل دوبار انجام دادن آزمون برای یک گروه با یک فاصله زمانی (آزمون - بازآزمون) و به کارگیری ابزار توسط ارزیابان یکسان در دو موقعیت متفاوت (Intra-rater) قابل انجام است. برخی اوقات یک ابزار توسط تعدادی پاسخگو پاسخ داده می‌شود و سپس یک ارزیاب به پاسخ‌ها بر اساس معیارهای از پیش تعریف شده نمره می‌دهد. این نمرات در برگه دیگری غیر از برگه سؤالات ثبت می‌شود. بعد از یک فاصله زمانی (معمولاً دو هفته) همان ارزیاب مجدداً به همان پاسخ‌های پاسخگویان که قبلاً ثبت شده بود نمره می‌دهد و سپس همبستگی بین نمرات بار اول و دوم محاسبه می‌شود [۱۲]. اگر داده‌ها (نمرات) فاصله‌ای باشند یا بتوان آن‌ها را فاصله‌ای فرض کرد همبستگی درون‌رده‌ای محاسبه می‌شود. اگر داده‌ها اسمی یا رتبه‌ای باشد (مثلاً به صورت مردود- قبول یا به صورت ضعیف- متوسط و عالی) از آزمون کاپا استفاده می‌شود.

در آزمون- بازآزمون، که شایعترین روش انجام ثبات است، آزمون دوبار با یک فاصله زمانی در یک گروه اجرا می‌شود. فاصله زمانی بین آزمون و بازآزمون باید تا اندازه‌ای طولانی باشد که فرد پاسخ‌های قبلی خود را به یاد نداشته باشد و آنقدر کوتاه باشد که مطمئن باشیم تغییری که پاسخ‌ها را مخدوش کند رخ نداده است. فاصله بین آزمون و بازآزمون بین یک تا دو هفته مناسب در نظر گرفته می‌شود و ممکن است دلیلی برای فاصله زمانی خارج از این محدوده وجود داشته باشد. مهم این است که علت فاصله زمانی بین آزمون و بازآزمون شرح داده شود [۱۲, ۷]. پس از جمع‌آوری هر دو نوبت داده‌ها، اگر داده‌ها فاصله‌ای باشد و یا فاصله‌ای محسوب شود و چولگی نداشته باشد ضریب همبستگی درون‌خوشه‌ای برای زیرمقیاس وکل پرسشنامه محاسبه می‌شود. شایان ذکر است انجام آزمون همبستگی پیرسون روش صحیحی نیست. خود ICC دو نوع است؛ $ICC_{agreement}$ و $ICC_{consistency}$. نوع Absolute agreement با مدل Two-way random روش ارجح است. ICC بین دو آزمون ۰/۸ یا بیشتر نشاندهنده ثبات رضایت بخش است [۳۱]. هنگام انجام آزمون- بازآزمون میزان و مدیریت مقادیر گمشده باید مد نظر قرار گیرد و گزارش شود. نکته دیگری که باید مورد توجه قرار گیرد با ثبات بودن نمونه‌ها در فاصله آزمون-بازآزمون از لحاظ ویژگی مورد سنجش است [۱۲]. خطای

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

استاندارد ابزار SEM باید محاسبه شود. کوچک بودن خطای استاندارد ابزار مهم است؛ چون تغییری از لحاظ کلینیکی قابل اهمیت است که بالاتر از خطای استاندارد ابزار باشد. خطای استاندارد ابزار دقت نمره هر فرد را کمی‌سازی می‌کند. خطای استاندارد ابزار از فرمول زیر به دست می‌آید، که در این فرمول SD به معنی انحراف معیار (Standard deviation) مجموع دو نمونه آزمون و بازآزمون می‌باشد [۳۲]. SEM بر دو نوع است؛ در فرمول زیر بر حسب این‌که از $ICC_{agreement}$ یا $ICC_{consistency}$ استفاده کنیم، $SEM_{agreement}$ یا $SEM_{consistency}$ حاصل می شود.

$$SEM = SD\times\sqrt{(1-ICC)}$$

علاوه بر خطای استاندارد ابزار، پارامتر توافق ابزار نیز باید بررسی شود. پارامتر توافق ابزار در صورتی مثبت است که کمترین تغییر قابل تشخیص یا قابل کشف (Smallest detectable change- SDC) بیشتر از کمترین تغییر قابل اهمیت (Minimal important change- MIC) باشد. در واقع، کمترین تغییر قابل تشخیص، تغییری است که واقعی بوده و ناشی از خطای ابزار نیست. SDC می‌تواند با ضریب‌های اطمینان مختلف محاسبه شود. مثلاً ضریب اطمینان ۹۵ درصد، به این معنی است که مقدار حاصل شده با P value کمتر از ۰/۰۵، کمترین مقدار تغییر قابل اندازه‌گیری بالای خطای استاندارد اندازه‌گیری را نشان می‌دهد [۷, ۳۳]. با ضریب اطمینان ۹۵ درصد فرمول SDC بدین صورت است:

$$SDC=1.96\times\sqrt{2}\times SEM$$

برای اندازه‌گیری کمترین تغییر قابل اهمیت (MIC) دو روش اصلی شامل استفاده از یک معیار بیرونی (Anchor-based approaches)، و استفاده از روش توزیع (Distribution-based approaches) وجود دارد. در روش توزیع MIC با توجه به ویژگی‌های آماری نمونه‌ها محاسبه انجام می‌شود. روش استفاده از یک معیار بیرونی روش ارجح می‌باشد [۷, ۳۳]. برای هر روش چند فرمول وجود دارد. به عنوان مثال ماکسیموویچ و همکاران از روش استفاده از معیار بیرونی و با استفاده از سطح زیر منحنی (ROC- Receiver Operating Characteristic) یعنی (AUC- Area Under Curve) کمترین تغییر قابل اهمیت را محاسبه نمودند [۳۴]. می‌توان به عنوان معیار بیرونی یک سؤال با طیف پاسخ ۷ تا ۱۵ گزینه‌ای در پایان ابزار مطرح نمود. مثلاً یک سؤال با طیف پاسخ خیلی بدتر تا خیلی بهتر و سپس تفاوت نمره پرسشنامه در کسانی که خود را بدون تغییر گزارش نموده‌اند با کسانی که خود را کمی بهتر گزارش کرده‌اند به عنوان کمترین تغییر قابل اهمیت در نظر گرفت [۳۵]. وان کمپن و همکاران از این روش برای محاسبه کمترین تغییر قابل اهمیت استفاده نمودند [۳۳]. مواهرن و میدوز علاوه بر استفاده از یک معیار بیرونی از روش توزیع نیز برای محاسبه کمترین تغییر قابل اهمیت استفاده نمودند؛ این محققان حاصل‌ضرب اندازه اثر (Effect size) در انحراف معیار پایه را به عنوان کمترین تغییر قابل اهمیت محاسبه نمودند [۳۶]. برای اطلاع از روش‌های مختلف محاسبه کمترین تغییر قابل اهمیت به مقاله رایت و همکاران رجوع شود [۳۷]. طبق معیارهای کاسمین، برای تعیین توافق حداقل ۵۰ نمونه لازم است [۷]. همبستگی درون‌رده‌ای (پایایی) عددی بین ۰ و ۱ است که واحد ندارد؛ اما خطای استاندارد ابزار و کمترین تغییر قابل تشخیص واحد دارند. مثلاً اگر وزن را بر حسب کیلوگرم اندازه‌گیری می‌کنیم واحد خطای استاندارد ابزار و کمترین تغییر قابل تشخیص هم کیلوگرم خواهد بود پس کاربرد کلینیکی دارد. مثلاً اگر ابزاری برای اندازه گیری وزن داریم که خطایی معادل ۳۰۰ گرم دارد، کمترین تغییر قابل تشخیص با اطمینان ۹۵٪، ۸۳۲ گرم است. این ابزار می تواند برای بزرگسالان به کار رود چون معمولاً تغییر وزن زیر یک کیلوگرم در بزرگسالان قابل اهمیت نیست. این ابزار برای نوزادان و یا مثلاً وزن کردن آرد در آشپزخانه قابل استفاده نیست؛ زیرا در این موارد تغییرات زیر ۸۰۰ گرم قابل اهمیت است. بنابراین وقتی مفهوم مشخصی از تفاوت قابل اهمیت وجود دارد، خطای استاندارد ابزار اطلاعات مفیدی را مثل مورد بالا ارائه می دهد. در مواردی که نمره دهی پرسشنامه دقیقاً مشخص نیست؛ وضعیت به گونه‌ای دیگر است. مثلاً در نظر بگیرید یک ابزار جدید با تعدادی گویه برای بررسی وضعیت عملکردی طراحی شده است. این ابزار نمره‌ای بین صفر تا پنجاه دارد. یک ارتوپد می‌خواهد بداند یک نمره ۱۴ با خطای استاندارد ۲ چه معنی می‌دهد. چون او هیچ ایده‌ای ندارد که چه تغییر نمره‌ای تغییر کلینیکی را نشان می‌دهد. با آزمون ICC او متوجه می‌شود که آیا ابزار می‌تواند بین بیماران افتراق قائل شود یا خیر. اما او نمی‌تواند مطمئن شود که ابزار برای مانیتورینگ وضعیت عملکرد بیمارانش در طول زمان مناسب باشد. آگاهی از این توانایی ابزار به اطلاعات بیشتر در مورد تفسیر نمرات نیاز دارد. با بررسی نمرات افرادی که ناتوانی عملکردی

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

خفیف، متوسط و شدید دارند احساس معنی دار بودن نمرات افزایش می‌یابد. مقایسه با ابزارهای دیگر، بینش بیشتری را نسبت به ارزش نمرات ایجاد می‌کند. بررسی کمترین تغییر قابل اهمیت از نظر کلینیکی برای ابزارهای مختلف به درک این موضوع که چه تغییری در نمرات، مرتبط با تغییرات کلینیکی هستند، کمک می کند. تنها این اطلاعات هستند که امکان ارزیابی این که آیا پارامتر توافق یک ابزار برای تشخیص تغییرات کلینیکی کفایت می کند را فراهم می کنند [۳۸].

۴- قابلیت پاسخگویی به تغییرات یا حساسیت: قابلیت پرسشنامه برای تشخیص تغییرات در طول زمان را نشان می‌دهد [۶]. این قابلیت را به نام روایی آینده‌نگر نیز ذکر می‌کنند. در واقع در قابلیت پاسخگویی با طرح فرضیه، تغییرات میانگین بین گروه‌های شناخته شده یا همبستگی بین تغییرات در اندازه‌گیری‌ها آزمون می‌شود [۷]. ابزار باید قادر باشد تا تغییرات کلینیکی را از خطای ابزار تشخیص دهد. قابلیت پاسخگویی با توجه به مقایسه SDC با MIC (که در تکرارپذیری توضیح داده شد) بررسی می‌شود. اگر MIC بزرگتر از SDC باشد، یا نسبت پاسخگویی (که عبارت است از نسبت میانگین تغییرات پس از مداخله به انحراف معیار از میانگین پایه) حداقل ۱/۹۶ باشد، قابلیت پاسخگویی پرسشنامه تأیید می شود [۷, ۳۹]. از سطح زیر منحنی ROC (AUC) که توانایی ابزار را برای افتراق بیماران با و بدون تغییر بر طبق یک معیار بیرونی نشان می‌دهد نیز می‌توان استفاده کرد. سطح زیر منحنی برابر یا بیش از ۰/۷ نشان دهنده قابلیت پاسخگویی مناسب است [۷]. برخی محققان معتقدند در تفسیر میزان کمترین تغییر قابل اهمیت، باید میزان خطای نوع دوم در نظر گرفته شود. چنانچه معیار قابل اهمیت بودن تغییر، بزرگتر از کمترین خطای قابل اندازه‌گیری باشد ممکن است این معنی‌دار بودن در اثر خطای نوع دوم باشد؛ چون در محاسبه SDC خطای نوع دوم لحاظ نشده است. این محققان معتقدند برای این‌که پاسخگویی به یک تغییر یا مداخله مهم تلقی شود کمترین تغییر قابل اهمیت بایستی از چهار برابر خطای استاندارد ابزار بیشتر باشد [۳۵]. از آن‌جا که حداقل نمونه برای اندازه‌گیری MIC، ۵۰ نفر است، برای بررسی قابلیت پاسخگویی حداقل ۵۰ نمونه لازم است [۷]. افشار و همکاران برای بررسی قابلیت پاسخگویی به تغییرات پرسشنامه سنجش پیامد برای سندرم تونل کارپ، پاسخگویی به عمل جراحی را به عنوان یک تغییر بررسی نمودند. پرسشنامه سنجش پیامد برای سندرم تونل کارپ توسط بیماران قبل از عمل جراحی و دو بار بعد از عمل جراحی (به فاصله شش ماه) تکمیل شد و برای مشخص شدن قابلیت پاسخگویی به تغییر، اندازه اثر محاسبه گردید [۲].

۵- دیگر ویژگی‌های روانسنجی:

۵-۱- اثر سقف و اثر کف: این اثر هنگامی وجود دارد که بیش از ۱۵ درصد پاسخ دهندگان به ترتیب بیشترین یا کمترین نمره قابل دستیابی را کسب کنند. وجود اثر سقف و کف نشان دهنده این است که احتمالاً گویه‌های نشان دهنده حداکثر و حداقل شدت پدیده در پرسشنامه گنجانده نشده‌اند؛ که نشان دهنده روایی محتوای ناکافی است؛ در نتیجه بیماران با حداقل و حداکثر نمره قابل کسب از هم افتراق داده نمی‌شوند بنابراین پایایی کاهش می‌یابد [۷] برای بررسی اثر سقف یا کف حداقل ۵۰ نمونه لازم است. چولگی (Skewness) متغیرها نه تنها به عنوان یک شاخص تشخیص نرمال نبودن در نظر گرفته می‌شود بلکه به عنوان یک شاخص وجود اثر سقف و کف نیز کاربرد دارد؛ فرض بر این است که وقتی چولگی مثبت است اثر کف و وقتی چولگی منفی است اثر سقف وجود دارد؛ هو و کارول عقیده دارند چولگی برای وجود اثر سقف و کف نه کافی است و نه ضروری؛ آن‌ها در مقاله خود این مقوله را با ذکر مثال و تصویر نشان دادند [۴۰].

۵-۲- تفسیرپذیری: تفسیرپذیری به معنی میزان توانایی برای ارجاع معانی کیفی به نمرات کمی است. محققان باید در مورد این‌که چه تغییراتی در نمرات از لحاظ کلینیکی معنی‌دار است، گزارش دهند. انواع مختلف اطلاعات می‌تواند به تفسیرپذیری ابزار کمک کند: ۱- میانگین و انحراف معیار (زیرگروه‌ها) جمعیت مرجع. ۲- میانگین و انحراف معیار زیرگروهی از بیماران که انتظار می‌رود از لحاظ نمره متفاوت باشند؛ مثلاً گروه‌های با تشخیص متفاوت، گروه‌های سنی، جنسیتی، مراکز مراقبت اولیه در مقابل مراکز ثانویه. ۳- میانگین و انحراف معیار بیماران قبل و پس از درمانی که کارایی آن مشخص شده است. ۴- میانگین و انحراف معیار زیرگروه‌های بیماران بر اساس رتبه‌بندی جهانی بیماری. برای مثال اگر میانگین و انحراف معیار برای حداقل ۴ زیرگروه گزارش شود تفسیرپذیری بررسی شده است. مثلاً میانگین و انحراف معیار برای جمعیت عمومی بر اساس سن و جنسیت طبقه‌بندی شود. به علاوه کمترین تغییر قابل اهمیت از نظر کلینیکی باید تعریف شود تا قادر باشد تغییرات را در طول زمان تفسیر نموده و برای محاسبه حجم نمونه استفاده شود. توصیه می‌شود از روش استفاده از معیار بیرونی برای محاسبه

کمترین تغییر قابل اهمیت استفاده شود. حداقل نمونه برای محاسبه کمترین تغییر قابل اهمیت ۵۰ نفر است [۷].

بحث و نتیجه‌گیری

با توجه به مطالب ذکر شده، مشخص می‌گردد که علاوه بر پایایی و روایی، ویژگی سومی به نام قابلیت پاسخگویی به تغییرات وجود دارد که فرایند روانسنجی را تکمیل می‌نماید. در تحلیل عامل اکتشافی نباید به طور کامل به تحلیل آماری تکیه شود. طراح ابزار که بر مفهوم مورد سنجش توسط ابزار و سازه‌های احتمالی آن اشراف کامل دارد، نقش مهمی در انجام و تفسیر تحلیل عامل اکتشافی دارد. بهینه این است که طراحان تنها به انجام تحلیل عاملی به عنوان روایی سازه بسنده نکنند و از دیگر روش‌های روایی سازه نیز استفاده کنند. محاسبه خطای استاندارد ابزار باید به عنوان جزئی از پایایی محاسبه شود. برخی ویژگی‌ها هستند که گرچه صاحب‌نظران حوزه ابزارسازی بر انجام آن‌ها اتفاق نظر ندارد، ولی بررسی آن‌ها می تواند کیفیت ابزار را افزایش دهد.

در پایان پس از تشریح فرآیند ترجمه، طراحی و روانسنجی ابزارهای سنجش سلامت (در این مقاله و مقاله قبلی) برای تکمیل موضوع توجه محققان محترم را به چند نکته کاربردی جلب می نماید:

- پرهیز از گسترش بی رویه مطالعات روانسنجی:

روان سنجی ابزارهای سنجش سلامت اگرچه امری ضروری به نظر می رسد، اما گسترش بی رویه آن به ویژه توسط پژوهشگرانی که باید علی القاعده به درمان و آلام بیماران بپردازند چندان شایسته نیست. این محققان بهتر است از ابزارهای موجود استفاده کرده و پیامدهای یک درمان و یا مقایسه پیامد درمانهای گوناگون را ارزیابی کنند.

- توجه به مالکیت معنوی:

قبل از هر چیز باید اشاره شود که ابزارهای سنجش سلامت نیز مانند هر محصول علمی دیگر مشمول مقررات مالکیت معنوی است و استفاده از هر ابزاری منوط به کسب اجازه از طراح آن است؛ مگر در مواردی که به صراحت مشخص شده باشد که استفاده از پرسشنامه نیازی به کسب اجازه ندارد.

- رعایت اولویت‌ها در انتخاب ابزارها:

برای انجام دقیق و بهینه تحقیقات اولویت کار باید بر استفاده از پرسشنامه های کوتاه و متناسب با اهداف اختصاصی متمرکز شده و حتی الامکان باید از رویکرد استفاده از چند پرسشنامه در یک مطالعه پرهیز کرد. همچنین توصیه می شود تا آنجا که ممکن است از ابزارهای معتبر موجود برای سنجش شاخص های سلامت در مطالعات استفاده شده و از طراحی غیر ضرور ابزارهای جدید جلوگیری شود؛ این توصیه برای طراحی و روانسنجی ابزارهای جدید که یکبار مصرف هستند مورد تاکید است.

- تعیین نقطه برش

هنگام ارائه نتایج باید از دسته بندی و گروه بندی افراد تحت مطالعه بدون داشتن دلایل علمی و بی مبنا پرهیز شود، مگر آنکه نقطه یا نقاط برش توسط پدید آورندگان از قبل ارائه شده باشد. در صورت نیاز جدی به تعیین نقطه یا نقاط برش ارائه دلیل یا دلایل علمی در این خصوص ضروری است.

پرسشنامه های تغییر یافته:

چنانچه در ابزاری سوال یا سوالاتی اضافه یا کم شود باید توجه داشت که این کار محتاج کسب اجازه است. این نوع پرسشنامه ها در واقع گونه تغییر شکل یافته تلقی شده و نیازمند روانسنجی است.

- گزارش نتایج روایی سازه:

هنگام ارائه نتایج تحلیل عاملی اکتشافی، باید اعداد مربوط به بار های عاملی تمام عامل ها در جدول ذی ربط درج گردد.

- تفاوت روایی و پایایی با روایی درونی و بیرونی:

محققان نباید روایی درونی (Internal validity) و بیرونی (External validity) را با روایی و پایایی اشتباه بگیرند. روایی داخلی نمایانگر آن است که تا چه اندازه یافته های تحقیق از صحت و دقت لازم برخوردار است. این روایی با توانا ساختن پژوهشگر در جمع‌آوری اطلاعات و تجزیه و تحلیل آنها، حذف کلیه عوامل مداخله گر و تعبیر و تفسیر درست آنها سروکار دارد. روایی بیرونی به قابلیت تعمیم‌پذیری یافته‌های تحقیق ارتباط دارد. به این معنی که آیا نتایج آزمایش قابل اطمینان بوده و می‌توان آن را به جامعه-ای که نمونه از آن انتخاب شده است، تعمیم داد یا نه.

سهم نویسندگان

زیبا تقی‌زاده: نظارت بر تدوین مقاله و تصحیح آن
عباس عبادی: نظارت بر تدوین مقاله و تصحیح آن
علی منتظری: ویرایش آخر و اصلاح نهایی مقاله
زهرا شاهواری: همکاری در نوشتن مقاله
محمود طاووسی: همکاری در نوشتن مقاله
راضیه باقرزاده: طرح اولیه و نوشتن مقاله

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

منابع

1. Uijen AA, Heinst CW, Schellevis FG, van den Bosch WJ, van de Laar FA, Terwee CB, et al. Measurement properties of questionnaires measuring continuity of care: a systematic review. PloS one 2012;7:e42256 Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0042256 [Accessed 15 may 2016]

2. Afshar A, Yekta Z, Etemadi A, Mirzatoloee F. Outcome measurement questionnaires for carpal tunnel syndrome. Iranian Journal of Orthopedic Surgery 2005;4:346-50 [Persin]

3. Eshaghi S, Farajzadegan Z, Babak A. Healthy lifestyle assessment questionnaire in elderly: translation, reliability and validity. Payesh 2010;9:91-99 [Persin]

4. Pajaohande A, Farzad V, Kadivar P. Devising and Validating the Adolescent Attachment Styles Questionnaire (AASQ). International Journal of Education and Applied Sciences 2014;1:26-36

5. Taghizadeh Z, Ebadi A, Montazeri A, Shahvari Z, Tavoosi M, Bagherzadeh R. Psychometric properties of health assessment instruments: Part 1: Translation, designing, face and content validity. Payesh 2017;3: 343-357

6. Reneman MF, Dijkstra A, Geertzen JH, Dijkstra PU. Psychometric properties of chronic pain acceptance questionnaires: a systematic review. European Journal of Pain 2010;14:457-465

7. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 2007;60:34-42

8. Nunnally J, Bernstein I. Psychometric Theory. 3th Edition, McGraw-Hill: New York, 1994

9. Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychological Assessment 1995;7:309-319

10. Plitcha SB, Kelvin EA. Munro's Statistical Methods for Health Care Research. 6th Edition, Lippincot Williams and Wilkins: Philadelphia, 2013

11. Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational Research Methods 2004;7:191-205

12. Waltz CF, Strickland OL, Lenz ER. Measurement in nursing and health research. 4 th Edition, Springer Publishing Company: New York, 2010

13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 2010;63:737-45

14. Henson RK, Roberts JK. Use of exploratory factor analysis in published research common errors and some comment on improved practice. Educational and Psychological Measurement 2006;66:393-416

15. Carretero-Dios H, Pérez C. Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. International Journal of Clinical and Health Psychology 2007;7:863-882

16. Hinkin TR. Scale development principles and practices, in Swanson RA, Holton EF. Research in organizations: Foundations and methods of inquiry. 1th Edition. Berrett-Koehler Publishers: San Francisco, 2005:161-179

17. Costello AB, Osborne JW. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. Practical Assessment Research & Evaluation 2005; 10 Available from: http://pareonline.net/getvn.asp?v=10&n=7 [Accessed 10 may 2016]

18. Ledesma RD, Valero-Mora P. Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. Practical Assessment, Research & Evaluation 2007;12:1-11

19. Shahvari Z, Raisi F, Yekta ZP, Ebadi A, Kazemnejad A. Married Women's Sexual Satisfaction Questionnaire; A Developmental and Psychometric Evaluation. Iranian Red Crescent Medical Journal 2015;17: e26488 Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443391/ [Accessed 20 April 2016]

20. Ping RA. Testing latent variable models with survey data 2004 Online Paper Available from: http://home. att. net/~ rpingjr/lv1/toc1. Htm [Accessed 20 April 2016]

21. Zaiţ A, Bertea PSPE. Methods for testing discriminant validity. Management & Marketing Journal 2011;9:217-224

22. Colton D, Covert RW. Designing and Constructing Instruments for Social Research and Evaluation. 1st Edition, Jossey-Bass: San Francisco, 2007

23. Rhew IC, Simpson K, Tracy M, Lymp J, McCauley E, Tsuang D, et al. Criterion validity of the Short Mood and Feelings Questionnaire and one-and two-item depression screens in young adolescents. Child and Adolescent Psychiatry and Mental Health, 2010;4:1-11

24. Drost EA. Validity and reliability in social science research. Education Research and Perspectives 2011;38:105-23

25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010;19:539-549

26. Hadian M, Nakhostin Ansari N, Asgari T, Abdolvahab M, Jalili M. Inter& intra rater reliability of Berg Balance Scale for evaluation of the balance in children with spastic hemieplegia. Journal of Modern Rehabilitation 2007;1:31-37

27. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 2005;85:257-268

28. Minick KI, Kiesel KB, Burton L, Taylor A, Plisky P, Butler RJ. Interrater reliability of the functional movement screen. The Journal of Strength & Conditioning Research 2010;24:479-486

29. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica 2012;22:276-282

30. Algina J, Crocker L. Introduction to classical and modern test theory. Rinehart and Winston Inc: New York, 1986

31. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. Optometry & Vision Science 2007;84:663-674

32. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. The Journal of Strength & Conditioning Research 2005;19:231-240

33. Van Kampen DA, Willems WJ, van Beers LW, Castelein RM, Scholtes VA, Terwee CB. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). Journal of orthopaedic surgery and research 2013;8 online paper
Available from http://www.josr-online.com/content/8/1/40 [Accessed 16 December 2015]

34. Maksymowych WP, Lambert RG, Brown LS, Pangan AL. Defining the minimally important change for the SpondyloArthritis Research Consortium of Canada spine and sacroiliac joint magnetic resonance imaging indices for ankylosing spondylitis. The Journal of Rheumatology 2012;39:1666-1674

35. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. Journal of Clinical Eepidemiology 2009;62:1062-1067

36. Mulhern B, Meadows K. Investigating the minimally important difference of the Diabetes Health Profile (DHP-18) and the EQ-5D and SF-6D in a UK diabetes mellitus population. Health 2013;9:1045-54

37. Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). Journal of Manual & Manipulative Therapy 2012;20:160-6

38. De Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology 2006;59:1033-9

39. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Diseases 1987;40:171-8

40. Ho AD, Carol CY. Descriptive Statistics for Modern Test Score Distributions Skewness, Kurtosis, Discreteness, and Ceiling Effects. Educational and Psychological Measurement 2015;75:365-88

Payesh

Journal of the Iranian Institute for Health Sciences Research

ABSTRACT

Translation, development and psychometric properties of health related measures-Part 2: construct validity, reliability and responsiveness

Abbas Ebadi[1], Ziba Taghizadeh[2], Ali Montazeri[3], Zahra Shahvari[4], Mahmoud Tavousi[3], Razieh Bagherzadeh[5*]

1. Behavioral Sciences Research Center (BSRC), Nursing Faculty, Baqiyatallah University of Medical Sciences, Tehran, Iran
2. School of Nursing and Midwifery, Tehran University of Medical Sciences. Tehran, Iran
3. Health Metrics Research Center, Iranian Institute for Health Sciences Research, ACECR, Tehran, Iran
4. Medical professionalism office, Tehran University of Medical Sciences. Tehran, Iran.
5. School of Nursing and Midwifery, Bushehr University of Medical Sciences. Bushehr, Iran

Payesh 2017; 4: 445- 455
Accepted for publication: 1 February 2017
[EPub a head of print-14 May 2017]

Objective (s): Data collection is one of the important stages of research process and requires valid and reliable instruments. The aim of this article was to introduce important steps of translation and construction process of a questionnaire. Adopting an appropriate instrument for research and developing an optimal instrument needs knowledge on psychometric criteria. The present paper explains the construct validity reliability and responsiveness as the important components of psychometric properties of health related instruments.
Methods: The present paper was written based on exisiting documentaries.
Results: There are different methods for evaluating construct validity, which include structural validity (factor analysis), hypothesis testing, convergent and discriminant validity, and cross-cultural validity. Criterion validity is also a type of validity, which can improve the quality of an instrument if being applicable. Reliability is an essential part of psychometric evaluation, which includes internal consistency, stability, equivalency and standard error of measurement. Instrument developers have often been concerned with internal consistency and stability. Examining standard error of measurement as an important component of reliability has been ignored in the development of most instruments. Responsiveness is approved if the smallest detectable change is less than the clinical minimal important change. Furthermore, the area under the ROC curve can also be used for examining responsiveness. Floor and ceiling effects exist if more than 15% of respondents get the highest or lowest possible scores.
Conclusion: Psychometric knowledge should be considered as basic requirement of translation and development of health-relate measures. This could help investigators to collect valid and reliable information when collecting the data.

Key Words: Psychometric properties, Construct validity, Reliability, Responsiveness

[DOR: 20.1001.1.16807626.1396.16.4.5.8]

* Corresponding author: School of Nursing and Midwifery, Bushehr University of Medical Sciences, Bushehr, Iran
Tel: 07733450178
E-mail: r.bagherzadeh@bpums.ac.ir