



جزوه آموزشی درس سنجش و اندازه گیری

ویژه داوطلبان دکتری رشته روان شناسی تربیتی

(مرکز آموزشی نوگام)

تهیه و تدوین

دکتر مهسا امیربگلوی داریانی

پاییز ۱۴۰۳

۳	 سابقه تاریخی روان‌سنجی
۶	 تعاریف مفاهیم عمده روان‌سنجی
۸	 سطوح اندازه‌گیری
۱۱	 دسته‌بندی روش‌های اندازه‌گیری
۱۷	 هدف‌های کلی کاربرد آزمون‌ها
۲۰	 مراحل ساخت آزمون
۲۰	 طرح ریزی آزمون
۲۱	 اهداف آموزشی
۲۸	 انواع سؤال‌های آزمون
۳۳	 تحلیل سؤال‌های آزمون
۴۲	 نظریه کلاسیک اندازه‌گیری
۴۷	 اعتبار (Reliability)
۶۰	 روایی (Validity)
۷۱	 نرم یا هنجار
۷۷	 مقیاس‌های سنجش نگرش
۷۸	 نظریه صفت مکنون (نظریه سؤال - پاسخ)

روان‌سنجی شاخه‌ای از روان‌شناسی است که با اندازه‌گیری تفاوت‌های فردی سر و کار دارد. برای مطالعه علمی و مناسب نظریه‌های روان‌شناسی، استفاده از فنون روان‌سنجی و سنجش دقیق تفاوت‌های فردی با به کار بستن آزمون-های روانی مناسب، یا فنون دیگر امری حیاتی است.

سابقه تاریخی روان‌سنجی

این واقعیت که افراد از نظر توانایی، استعداد، شخصیت و رفتار تفاوت دارند و اینکه این تفاوت‌ها را می‌توان مورد سنجش قرار داد، احتمالاً از ابتدای تاریخ مکتوب زندگی انسان مورد توجه بوده است. افلاطون و ارسطو در آثار خود از تفاوت‌های فردی بین افراد و گروه‌ها سخن گفته‌اند. دوبوا استفاده از آزمون را به سه زمینه اساسی نسبت می‌دهد: خدمات اجتماعی، امتحانات آموزشگاهی و مطالعه تفاوت‌های فردی. اجرای آزمون‌های خدمات اجتماعی از ۳۰۰۰ سال پیش در چین آغاز شد، در این زمان یکی از امپراتوران تصمیم گرفت شایستگی مأموران خود را بسنجد.

اولین قدم برای اندازه‌گیری پدیده‌های روانی در آزمایشگاه‌ها و توسط فیزیکدان‌ها و فیزیولوژیست‌ها صورت گرفته است. کریستیان وُن وُلَف در ۱۷۳۲ رشته ریاضی روان‌شناسی را تحت عنوان «روان‌سنجی» دایر می‌کند. اولین آزمایشگاه روانشناسی آزمایشی توسط ویلهلم وونت در سال ۱۸۷۹ در شهر لایپزیک آلمان تأسیس شد. پس از وونت، فرانسیس گالتون انگلیسی در سال ۱۸۸۴ شروع به انجام تحقیقاتی کرد که او را به سمت اندازه‌گیری استعدادهای فردی و مطالعه تفاوت‌های فردی هدایت کرد. وی سعی کرد تا با محاسبه نسبت افرادی که از جامعه جلو می‌افتند نبوغ را اندازه‌گیری کند. سپس به تأسیس آزمایشگاه مردم‌شناسی (انسان‌سنجی) اقدام کرد. گالتون معتقد بود که بین توانایی ذهنی (هوش) و تمیز حسی رابطه وجود دارد. بدین منظور برای اندازه‌گیری قدرت تمیز حسی، آزمون‌های مختلفی نظیر: سوت گالتون، وزنه‌های گالتون، رنگ سنج و ... را ابداع کرد. خدمت مهم‌تر گالتون ابداع روش‌های آماری برای کمی ساختن نتایج حاصل از اجرای آزمون‌ها بود که بعد از وی کارل پیرسون انگلیسی کار او را تکمیل کرد و برخی از روش‌های آماری را که هسته مرکزی نظریه‌های اندازه‌گیری را تشکیل می‌دهد پیشنهاد کرد نظیر: همبستگی

گشتاوری پیرسون و آزمون نیکویی برازش خی دو. چارلز اسپیرمن کار گالتون و پیرسون را ادامه داد و فعالیت وی منجر به مفاهیم جدید اعتبار (پایایی) و تحلیل عاملی شد.

جیمز مک کین کتل آمریکایی اولین بار در ۱۸۹۰ اصطلاح «آزمون روانی» را وارد روان‌شناسی کرد. کتل نیز مانند گالتون عقیده داشت که با اندازه‌گیری فرایندهای ساده تمیز حسی و زمان واکنش می‌توان توانایی ذهنی را اندازه‌گیری کرد. در سال ۱۸۹۵ کارهای گالتون و کتل توسط آلفرد بینه و ویکتور هنری مورد انتقاد قرار گرفت. آنها معتقد بودند که هوش باید براساس توانایی‌های عالی‌تر ذهنی مانند حافظه، توجه، درک، تخیل و استدلال مورد اندازه‌گیری قرار گیرد.

شروع واقعی تهیه آزمون‌های روانی توسط آلفرد بینه و تئودور سیمون است. نخستین آزمون هوشی که در سال ۱۹۰۵ در فرانسه انتشار یافت دارای ۳۰ سؤال بود که از آسان به دشوار مرتب شده و سه کارکرد قضاوت، درک و فهم و استدلال را اندازه‌گیری می‌کرد. در سال ۱۹۰۸ ضمن تجدید نظر در این آزمون محتوای آن گسترش یافت و مفهوم سن عقلی در آن مطرح شد. در ۱۹۱۶ لوئیس ترمن به کمک دانشگاه استنفورد کالیفرنیا این آزمون را در کودکان آمریکایی هنجار کرد و آن را استنفورد-بینه نامید، همچنین اصلاح هوشبهر (IQ) را در این آزمون به کار برد. ویلیام اشترن آلمانی نخستین بار واژه هوشبهر را به کار برد که آن را نسبت سن عقلی (اندازه‌گیری شده) به سن تقویمی (واقعی) تعریف کرد. ترجمه‌ها و انطباق‌های متعددی از آزمون هوش بینه به عمل آمد است.

قدم بعدی در پیشرفت آزمون‌ها اجرای گروهی آنها بود. که برای اجرای سریعتر و بهتر آزمون بر روی سربازان آمریکایی در جنگ جهانی اول صورت گرفت، و بدین‌سان آزمون ارتشی آلفا و بتا ساخته شد. آزمون آلفا یک آزمون کلامی بود و برای افراد با سواد، که خواندن و نوشتن زبان انگلیسی را می‌دانستند به کار می‌رفت و آزمون بتا یک آزمون غیرکلامی بود و برای افراد بی‌سواد یا ناآشنا به زبان انگلیسی به کار برده می‌شد.

علاوه بر آزمون‌های هوش آزمون‌های شخصیت نیز در جنگ جهانی اول مورد توجه قرار گرفت. امیل کراپلین روان‌پزشک آلمانی «آزمون تداعی آزاد و اندیشه‌ها» را که قبلاً توسط گالتون تدوین شده بود در مورد بیماران روانی به کار بست و بعدها یونگ روش مشابهی را با عنوان «تداعی آزاد کلمات» برای تشخیص عقده‌های روانی بیماران به کار بست. وود ورث برای اولین بار برگه تهیه اطلاعات فردی را ساخت و ویل اسکات برای اولین بار آزمون صنعتی و تجاری را به کار برد.

نمونه سؤال

۱- چه کسی اولین آزمون انفرادی هوش را ساخت؟

(۱) بینه (۲) اشترن (۳) پیرسون (۴) براون

۲- تفاوت آزمون‌های ارتشی آلفا و بتا بیشتر از چه لحاظ است؟

(۱) فردی و گروهی (۲) کلامی و غیر کلامی

(۳) موضوع (۴) میزان دقت

۳- یکی از معایب آزمون‌های مک کین کتل این بود که:

(۱) فراایندهای سطوح پایین ذهن را اندازه می‌گرفتند.

(۲) فراخنای حافظه برای اعداد و کلمات را می‌سنجیدند.

(۳) طولانی بودند و خستگی آزمودنی را به همراه داشتند.

(۴) تهیه هنجار یا جدول درجه‌بندی آنها دشوار بود.

۴- مفهوم تحلیل عاملی اولین بار توسط چه کسی مطرح شد؟

(۱) کتل (۲) گالتون (۳) اسپیرمن (۴) پیرسون

تعاریف مفاهیم عمده روان‌سنجی

اندازه‌گیری (measurement):

قواعدی برای اختصاص دادن اعداد افراد، اشیاء، یا رویدادها به منظور کمی ساختن آنها.

ابزار اندازه‌گیری

ابزار اندازه‌گیری به عنوان فن یا روشی برای ربط دادن آنچه در دنیای واقعی مشاهده می‌شود به آنچه که مورد اندازه‌گیری قرار می‌گیرد (که در اصل یک مفهوم انتزاعی یا نظری) است، تعریف می‌شود. این مفاهیم انتزاعی همان سازه‌ها هستند که گاهی صفت با خصیصه مکنون یا نهفته نامیده می‌شوند.

سازه (Construct)

سازه یک مفهوم نظری و انتزاعی است که ابزار اندازه‌گیری برای توصیف آن ساخته می‌شود. بسیاری از ویژگی‌های روان‌شناختی آدمی مانند استعداد ریاضی، خودپنداره، اعتماد به نفس، خلاقیت و ... همگی سازه نامیده می‌شوند. برای اندازه‌گیری این سازه‌ها، از روش‌ها و ابزارهایی مانند پرسشنامه، آزمون، مصاحبه، مقیاس درجه‌بندی رفتار و برخی ابزارهای فیزیکی و مانند آن‌ها استفاده می‌شود تا تفاوت‌های فردی افراد در هر مورد تعیین شود.

آزمون (Test)

وسیله‌ای عینی و استاندارد شده که برای اندازه‌گیری نمونه‌ای از رفتار یا خصایص آدمی به کار می‌رود.

آزمودن (Testing)

وقتی برای اندازه‌گیری یکی از ویژگی‌های روانی یا تربیتی یک فرد یا گروهی از افراد از آزمون استفاده می‌شود، به این فعالیت آزمودن می‌گویند.

سنجش (Assessment)

سنجش از آزمودن و اندازه گیری مفهوم گسترده تری دارد. وقتی اندازه گیری صرفاً با استفاده از آزمون یا پرسشنامه است آزمودن صورت می پذیرد، اما در سنجش وسایل و فنون مختلفی برای جمع آوری اطلاعات به کار می روند از جمله آنها می توان به آزمون، پرسشنامه، مقیاس درجه بندی، فهرست واری، مصاحبه و ... اشاره کرد.

تفاوت دیگر سنجش و اندازه گیری و آزمودن این است که هم اندازه گیری و هم آزمودن با کمیّت سر و کار دارند، اما سنجش به کمیّت منحصر نمی شود و جنبه کیفی نیز دارد. سنجش نوعی قضاوت یا داوری در مورد اندازه یک سازه بر اساس یک یا چند ملاک مشخص و تعریف شده به منظور هدف های خاص مانند تشخیص اختلال روانی است. به عنوان مثال فرض کنید سازه هوش یک کودک، با آزمون و کسلر اندازه گیری شده و هوشبهر کلی وی ۶۵ برآورد شده، تا اینجا عمل اندازه گیری صورت گرفته است. اکنون می خواهیم تصمیم بگیریم آیا این کودک در مدارس عادی تحصیل کند یا مدارس استثنایی، در اینجا ما هوشبهر کودک را با ملاک مشخص کودکان کم توان ذهنی مقایسه و سپس درباره وی تصمیم گیری می کنیم. بنابراین سنجش مستلزم داوری و تصمیم گیری در مورد سازه اندازه گیری شده براساس یک ملاک مشخص است.

ارزشیابی (evaluation)

ارزشیابی یک فرایند نظامدار برای جمع آوری، تحلیل، و تفسیر یک برنامه، یک روش یا دوره آموزشی یا یک برنامه ریزی و اقدام اجتماعی در مقایسه با هدف یا هدف هایی که این برنامه روش یا اقدام برای آن طراحی و اجرا شده است. ملاک ارزشیابی هدف های (ارزش ها) از پیش تعیین شده است، و در نهایت منجر به داوری ارزشی و تصمیم گیری عملی و اجرایی درباره برنامه آموزشی می شود.

تفاوت اندازه گیری و ارزشیابی

ارزشیابی جامع تر از اندازه گیری است و علاوه بر جنبه های کمی جنبه های کیفی را نیز در بر می گیرد. در اندازه گیری داوری ارزشی وجود ندارد و فقط کمی کردن مطرح است. اندازه گیری به منظور مقایسه افراد از لحاظ ویژگی به خصوصی انجام می شود، اما در ارزشیابی مقایسه تک تک افراد مطرح نیست و هدف مطلوب بودن فرایند یا فرآورده یا موفقیت و شکست یک برنامه یا دوره است.

نمونه سؤال

۱- جمع آوری اطلاعات به منظور قضاوت در مورد پدیده‌ها و تصمیم‌گیری در مورد آنها چه نامیده می‌شود؟

(۱) سنجش (۲) ارزیابی (۳) اندازه‌گیری (۴) ارزشیابی

۲- کدام تعریف بیانگر تعریف سنجش است؟

- (۱) اختصاص کمیت به خصیصه‌ها یا صفات مختلف
(۲) تفسیر، ارزش‌گذاری و داوری در مورد عملکرد اختصاصی فرد
(۳) تحلیل چندجانبه عملکرد اختصاصی فرد با استفاده از فنون مختلف
(۴) اجرای آزمون پیشرفت تحصیلی به منظور تخمین عملکرد شغلی

۳- در کدام فعالیت تعیین کیفیت از ویژگی‌های عمده محسوب می‌شود؟

(۱) ارزشیابی (۲) روان‌سنجی (۳) ارزیابی (۴) اندازه‌گیری

سطوح اندازه‌گیری (مقیاس‌های اندازه‌گیری)

اندازه‌گیری در چهار سطح (مقیاس) مختلف انجام می‌شود، اسمی، ترتیبی، فاصله‌ای و نسبی. سطوح اندازه‌گیری با چهار ویژگی جداسازی، ترتیب مقداری، فاصله‌های مساوی و صفر مطلق تعیین می‌شود.

سطح اسمی (طبقه‌ای): (nominal)

اندازه‌گیری در سطح اسمی تنها ویژگی جداسازی را دارد. در این سطح اشیاء یا افراد را تنها برحسب صفات مشترکشان به گروه‌های مختلف طبقه‌بندی کرده و برای هر طبقه نام و یا نشانه‌ای را قائل می‌شویم. وقتی به طوری قراردادی زنان کد یک و مردان کد دو می‌گیرند، اندازه‌گیری در سطح اسمی است. در این سطح ترتیب وجود ندارد، فاصله‌ها مساوی نیست و صفر مطلق وجود ندارد. برای داده‌های اسمی نمی‌توان عملیات ریاضی به صورت معنادار انجام داد.

سطح ترتیبی (رتبه‌ای): (ordinal)

در این سطح بین اعداد یک رابطه منظم وجود دارد. در سطح ترتیبی عدد بزرگ‌تر به افرادی اختصاص داده می‌شود که صفت مورد اندازه‌گیری در آنها بیشتر است. این سطح اندازه‌گیری دارای ویژگی ترتیب و جداسازی است. مثل مرتب کردن افراد از کوتاه‌قدترین تا بلندقدترین. عدد بزرگ‌تر نشان‌دهنده قد بلندتر است، اما نمی‌توان فاصله اعداد را مساوی فرض کرد، و صفر مطلق نیز در این سطح وجود ندارد.

سطح فاصله‌ای: (interval)

در سطح فاصله‌ای علاوه بر اینکه همانند سطح رتبه‌ای دارای نظم و ترتیب منطقی است میزان تقریبی تفاوت میان مقادیر مختلف صفت مورد اندازه‌گیری را نیز نشان می‌دهد. در این مقیاس یک نقطه صفر وجود دارد که به صفر قراردادی معروف است، بدین معنا که صفر به معنای فقدان کامل یک ویژگی نیست. مثلاً صفر دماسنج به معنای عدم وجود دما (یا حرارت) نیست. نمره‌هایی مثل هوشبهر یا نمره‌های امتحانی در مقیاس فاصله‌ای قرار دارند. اندازه‌گیری در سطح فاصله‌ای ویژگی‌های جداسازی، ترتیب مقداری و فاصله‌های مساوی را دارد. اما صفر مطلق به معنای عدم وجود، در آن بی‌معناست. در این مقیاس عملیات ریاضی جمع و تفریق امکان‌پذیر است، اما عملیات ضرب و تقسیم را نمی‌توان برای آنها به کار برد.

سطح نسبی: (ratio)

این سطح دارای همه خصوصیات مقیاس فاصله‌ای است و علاوه بر آن دارای صفر مطلق نیز هست، اندازه‌گیری در سطح نسبی دارای هر چهار ویژگی (جداسازی، ترتیب، فاصله مساوی و صفر مطلق) است. بسیاری از ویژگی‌های فیزیکی نظیر طول، وزن و... در این سطح قرار دارند. تمامی اعمال ریاضی در این سطح قابل انجام است.

طبقه‌بندی مقیاس‌های پیوسته و گسسته

مقیاس گسسته: در مقیاس‌های گسسته یا مجزا متغیر مورد نظر تنها می‌تواند یکی از ارزش‌های معین را به خود اختصاص دهد. به عنوان مثال تعداد فرزندان یک خانواده، تعداد فرزندان همیشه یک عدد درست است و اعشار نمی‌گیرد. متغیرهای دیگر نظر جنسیت، تعداد سؤال‌های یک آزمون و نظیر آن‌ها از جمله مقیاس‌های گسسته هستند.

مقیاس پیوسته: در مقیاس پیوسته برخلاف مقیاس گسسته متغیر مورد نظر می تواند در فاصله بین نقاط مقیاس هر تعداد ارزش را به خود اختصاص دهد. مانند سرعت، وزن، هوش، پیشرفت تحصیلی و ... ویژگی مهم این مقیاس ها تساوی واحدهای اندازه گیری است.

نمونه سؤال

۱- مقیاس اندازه گیری مورد استفاده در آزمون های چند گزینه کدام نوع است؟

(۱) اسمی (۲) ترتیبی (۳) فاصله ای (۴) نسبی

۲- کدام یک از داده های زیر دارای مقیاس فاصله ای است؟

(۱) درجه حرارت (۲) رنگ چشم (۳) اندازه قد (۴) میزان موافقت با گویه های نگرش سنج

نمونه سؤال کنگور دکتري روان شناسی تربیتی

۱- در سرشماری سال ۱۳۸۵ از مردم پرسیده اند که « در خانه به چه زبانی (فارسی، ترکی، کردی، لری، سایر زبان ها) صحبت می کنند؟ » نتایج به دست آمده در کدام مقیاس قرار می گیرد؟ (سال ۱۳۹۸)

(۱) نسبی (۲) اسمی (۳) ترتیبی (۴) فاصله ای

۲- با گسسته در نظر گرفتن یک متغیر پیوسته کدام حالت برای واریانس محتمل تر است؟ (سال ۱۴۰۰)

(۱) کم می شود. (۲) زیاد می شود.

(۳) تغییری نمی کند. (۴) بستگی به رابطه متغیر با سایر عوامل دارد.

دسته‌بندی روش‌های اندازه‌گیری

روش‌های اندازه‌گیری و آزمون‌ها را بر حسب ملاک‌های متفاوت می‌توان به انواع مختلف طبقه‌بندی کرد:

دسته‌بندی براساس ویژگی مورد اندازه‌گیری

آزمون‌ها براساس ویژگی‌های مورد اندازه‌گیری را می‌توان به دو دسته آزمون‌های توانایی و پرسشنامه‌های سنجش عاطفی تقسیم کرد.

آزمون‌های توانایی

آزمون‌های توانایی، آموخته‌ها، مهارت‌ها و استعداد‌های افراد را می‌سنجد، دسته‌ای توانایی‌های شناختی را اندازه می‌گیرد، و تعدادی توانایی‌های روانی حرکتی. توانایی‌های شناختی در زمینه‌های تحصیلی و فعالیت‌های فکری و ذهنی مورد استفاده قرار می‌گیرد. اما آزمون‌های روانی حرکتی در موقعیت‌های صنعتی، نظامی و تربیت بدنی قابل استفاده است. آزمون‌های توانایی را می‌توان به دو دسته کلی آزمون‌های استعداد و آزمون‌های پیشرفت تحصیلی تقسیم کرد.

آزمون‌های استعداد توانایی یا آمادگی فرد را برای انجام دادن کارهایی که در پیش دارد و همچنین ظرفیت او را برای یادگیری‌ها مختلف می‌سنجد. به عبارت دیگر اندازه‌گیری توانایی‌ها و قابلیت‌های بالقوه فرد مدنظر است و ناظر به آینده و پیش‌بینی موفقیت‌های فرد است. آزمون‌های استعداد را می‌توان به دو دسته کلی آزمون‌های استعداد کلی (هوش) و آزمون‌های استعداد خاص (مثلاً استعداد ریاضی، استعداد فنی، استعداد موسیقی و ...) طبقه‌بندی کرد.

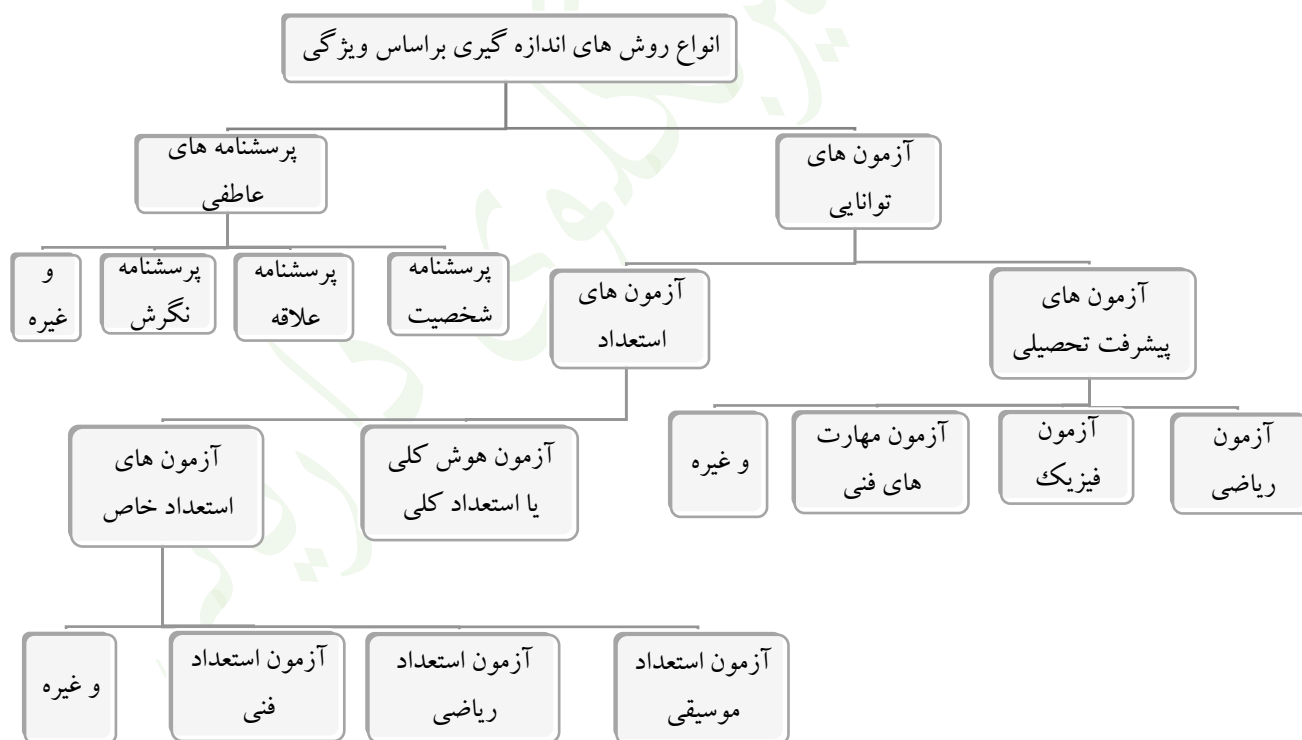
آزمون‌های پیشرفت تحصیلی دانش و مهارت‌هایی که فرد تا لحظه اجرای آزمون کسب کرده را اندازه می‌گیرد. این آزمون‌ها ناظر به گذشته و تعیین آموخته‌های قبلی فرد است. آزمون‌های پیشرفت تحصیلی و استعداد، ماده‌ها و محتوای نسبتاً مشترک دارند، اما به لحاظ هدف متمایزند.

پرسشنامه‌های اندازه‌گیری عاطفی

این پرسشنامه‌ها معطوف به ویژگی‌های عاطفی و شخصیتی افراد است. این پرسشنامه‌ها وضع موجود آزمون‌شونده را در موقعیت طبیعی زندگی می‌سنجد، و بیشتر به امور مشاوره در حل مسائل روانی، تربیتی و حرفه‌ای کمک می‌کند. این پرسشنامه‌ها دارای انواعی زیادی است، از جمله پرسشنامه‌های شخصیت، نگرش، علاقه، و ...

پرسشنامه‌های شخصیت را می‌توان به سه دسته پرسش‌های گزارش شخصی، پرسشنامه‌های جامعه‌سنجی (افراد دیگر نظرشان را نسبت به فرد می‌گویند)، و فنون مشاهده‌ای تقسیم کرد. در طبقه‌بندی دیگر آزمون‌های شخصیت را می‌توان به دو دسته آزمون‌های عینی (سازمان یافته، که در آن قضاوت آزمونگر بی‌تأثیر است مثل آزمون نئو و MMPI) و آزمون‌های فرافکن (که تحت تأثیر قضاوت آزمونگر قرار دارد، مثل آزمون‌های رورشاخ، جملات ناتمام ... طبقه‌بندی کرد.

ابزارهای اندازه‌گیری علاقه و نگرش، به دوست داشتن و نداشتن‌های فرد اشاره می‌کند و به رجحان‌های فرد درباره فعالیت‌ها، نهادهای اجتماعی یا گروه‌ها ارتباط دارد. علائق احساس فرد نسبت به یک فعالیت است مثل رغبت سنج‌های شغلی - تحصیلی، و نگرش احساس فرد نسبت به یک شی یک نهاد اجتماعی یا گروه است.



آزمون‌های استاندارد شده در مقابل آزمون‌های معلم ساخته

آزمون‌های استاندارد شده (میزان شده) براساس دستورالعمل خاص و به شیوه یکسان اجرا، نمره گذاری و تفسیر می‌شوند. در آزمون‌های استاندارد شده، برای تفسیر نمره آزمودنی از نرم‌ها و هنجارها بر اساس عملکرد گروه مرجع استفاده می‌شود و نمره آزمودنی با آن مقایسه و نسبت به گروه نرم وضعیت وی تعیین می‌گردد. در مقابل آزمون‌های معلم ساخته برای ارزشیابی‌های غیررسمی است و فاقد نرم و هنجار برای مقایسه است.

آزمون‌های فردی در مقابل آزمون‌های گروهی

آزمون فردی توسط اجرا کننده تعلیم دیده و هر بار در مورد یک آزمودنی اجرا می‌شود، مثل آزمون وکسلر یا لکه‌های جوهر رورشاخ. آزمون گروهی در هر جلسه روی تعداد زیادی آزمودنی اجرا می‌شود مانند آزمون ریون.

آزمون‌های سرعت در برابر آزمون‌های قدرت

آزمون سرعت آزمونی است که آزمودنی باید در زمانی محدود و معین، هر اندازه که می‌تواند به سؤال‌های آن پاسخ دهد. آزمون قدرت آزمونی است که در آن زمان غیرمحدود یا زمان نسبتاً زیاد است و آزمودنی وسعت دانش و میزان درک و فهم و مهارت خود را در پاسخ دادن به سؤال‌ها نشان می‌دهد. در آزمون‌های سرعت معمولاً سطح دشواری سؤال‌ها یکسان است، در آزمون‌های قدرت معمولاً دشواری از آسان به مشکل تنظیم می‌شود.

آزمون‌های عینی در مقابل آزمون‌های ذهنی

آزمون‌های عینی قضاوت آزمایش کننده در نمره گذاری بی‌تأثیر است یا اثر ناچیز دارد، مثل آزمون‌های چندگزینه‌ای و بسته پاسخ، در آزمون‌های ذهنی نمره گذاری پاسخ آزمودنی تا حد زیادی از قضاوت آزمایش کننده تأثیر می‌پذیرد، مثل آزمون‌های باز پاسخ (تشریحی) یا آزمون‌های فراقکن شخصیت.

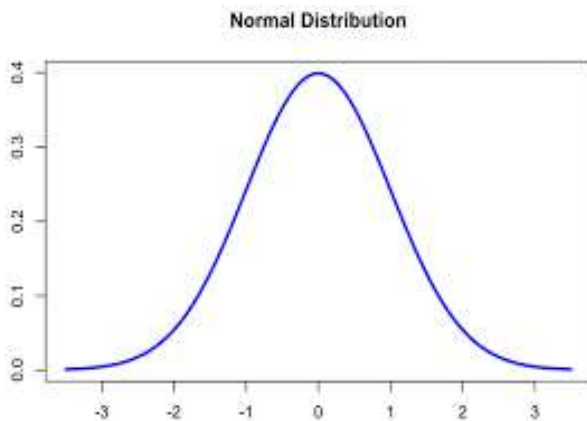
آزمون‌های وابسته به فرهنگ در برابر آزمون‌های ناوابسته به فرهنگ

آزمون‌های وابسته به فرهنگ براساس مشخصات یک فرهنگ خاص ساخته شده و پاسخ دادن به آن مستلزم آشنایی آزمودنی با آن فرهنگ است. اما پاسخ دادن به آزمون‌های ناوابسته به فرهنگ مستلزم آشنایی با فرهنگ خاصی نیست، مثل آزمون هوش کتل، یا فراخنای حافظه.

آزمون‌های ملاک مرجع در مقابل آزمون‌های نرم مرجع

آزمون‌های نرم مرجع (هنجار مرجع، گروه مرجع):

این دسته آزمون‌ها برای مشخص کردن کارکرد فرد در رابطه با کارکرد دیگران در یک ابزار سنجش به خصوص به کار می‌رود. معنا و اهمیت نمره فرد از طریق مقایسه پدیدار می‌شود و وضع نسبی فرد با گروه نرمی مشخص می‌شود.

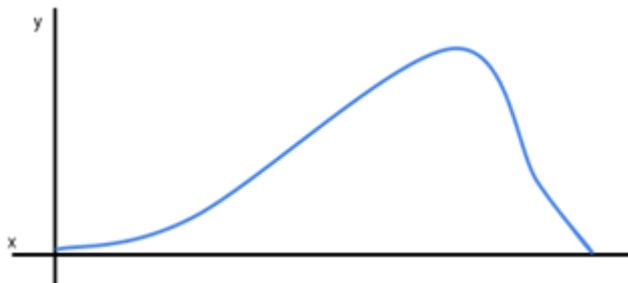


آزمون‌های هنجاری (ارزشیابی وابسته به هنجار) مبتنی بر ملاک

نسبی است. در سنجش هوش، استعداد، شخصیت و سایر جنبه‌های عاطفی به کار می‌رود. معروف‌ترین این آزمون‌ها مانند کنکور سراسری است. توزیع فراوانی در آزمون‌های هنجاری معمولاً به صورت توزیع نرمال و طبیعی است. در این توزیع اغلب نمره‌ها در اطراف میانگین و تعداد کمی بسیار بالاتر و یا بسیار پایین‌تر از میانگین است.

آزمون‌های ملاک مرجع (محتوا مرجع، هدف مرجع، چیرگی در حیطه، تسلط):

در این آزمون‌ها عملکرد آزمودنی با یک ملاک مشخص، بدون توجه به عملکرد سایر افراد گروه، مورد قضاوت قرار می‌گیرد. آزمون‌های ملاکی (ارزشیابی وابسته به ملاک) مبتنی بر ملاک مطلق است و به طور عمده در سنجش پیشرفت تحصیلی به منظور تعیین میزان موفقیت یادگیرندگان در رسیدن به هدف‌های آموزشی اجرا می‌شوند. مانند امتحانات نهایی مقاطع مختلف تحصیلی، که قبولی و عدم قبولی داوطلبان با توجه به ملاک نسبتاً مستقل و مطلق تعیین می‌شود. برخلاف آزمون‌های هنجار مرجع توزیع نمره-



های آزمون‌های ملاکی به توزیع نرمال ندارد و توزیع نمره‌های آن دارای کجی منفی است. بدین معنا که اکثریت افراد نمره‌های بالا کسب کرده‌اند و اکثریت بزرگی به هدف‌های آموزشی مورد نظر (ملاک مطلق) رسیده‌اند.

دسته‌بندی ارزشیابی‌های آموزشی بر اساس زمان وهدف

ارزشیابی آغازین: این ارزشیابی در ابتدای دوره پیش از انجام فعالیت‌های آموزشی انجام می‌شود، بدین منظور که آیا یادگیرندگان بر دانش‌ها و مهارت‌های پیش نیاز درس جدید تسلط دارند (اندازه‌گیری رفتارهای ورودی) و اینکه یادگیرندگان چه مقدار از هدف‌ها و محتوای درس تازه را قبلاً یاد گرفته‌اند (آزمون جایابی یا آزمون پایه‌گزینی).

ارزشیابی تکوینی (formative evaluation): ارزشیابی تکوینی یا فرایندی عمدتاً به منظور کمک به اصلاح موضوع ارزشیابی یعنی برنامه یا روش آموزشی مورد استفاده قرار می‌گیرد. این ارزشیابی در طول دوره آموزشی انجام می‌شود. در ارزشیابی تکوینی در پایان هر واحد درسی یک آزمون دقیق و مختصر که حاوی هدف‌های آموزشی در آن واحد است اجرا می‌شود، و اغلب آزمون‌های مورد استفاده آزمون‌های ملاکی با ملاک مطلق است.

ارزشیابی تشخیصی: این نوع ارزشیابی با هدف تشخیص مشکلات یادگیری دانش‌آموزان در جریان آموزش به کار می‌رود، و زمانی مورد استفاده قرار می‌گیرد که معلم با مشکلات مبرم و مکرری در مورد یک یا چند دانش‌آموز روبرو می‌شود. ابزارهای مورد استفاده سنجش تشخیصی اغلب آزمون‌های تشخیصی وابسته به هنجار است.

ارزشیابی تراکمی (summative evaluation): در ارزشیابی تراکمی تمام آموخته‌های دانش‌آموزان در طول یک دوره آموزشی تعیین می‌شوند و هدف آن نمره دادن به دانش‌آموزان و قضاوت درباره اثربخشی کار معلم و برنامه درسی یا مقایسه برنامه مختلف درسی با یکدیگر است. از آنجایی که معمولاً در پایان دوره آموزشی انجام می‌شود به آن ارزشیابی نهایی، پایانی یا جامع نیز گفته می‌شود. فنون مورد استفاده در ارزشیابی تراکمی به هدف‌های آموزشی دوره بستگی دارد و می‌تواند انواع آزمون‌ها، مشاهده، سنجش، عملکردها و ... را دربرگیرد.

آزمون‌های عملکردی در برابر آزمون‌های مدادکاغذی

آزمون‌های عملی مستلزم آن است که آزمودنی اشیاء را دستکاری و قطعاتی را به هم وصل کند و مهارت خود را در کار عملی نشان دهد. در آزمون مداد کاغذی آزمودنی پاسخ خود را به سؤال با علامت زدن در پاسخنامه یا به صورت کتبی مشخص می‌کند.

سنجش عملکردی

یک دسته از روش های سنجش واقعی سنجش عملکردی است. در سنجش عملکردی فرایندها یا فرآورده های یادگیری به طور مستقیم سنجش می شود. مواردی نظیر مهارت ها (مثل سخن گفتن، مهارت های ورزشی یا شغلی و ...)، عادت - های کاری (توانایی برنامه ریزی، پشتکار، ...)، نگرش های اجتماعی (علاقه مندی به رفاه دیگران، احترام به قانون، ...)، علاقه ها (علاقه مکانیکی، هنری، ...)، ارج گذاری و قدردانی (ابراز رضایت از طبیعت، خدمات اجتماعی و ...)، و سازگاری (رابطه با همسالان، سازگاری اجتماعی و ...) را به سختی می توان با آزمون های مداد کاغذی سنجید. این نوع سنجش، فعالیت های شناختی سطح بالا، حوزه های عاطفی و روانی - حرکتی را شامل می شود.

گرانلاند چهار دسته سنجش عملکردی را با نام های ۱) آزمون های کتبی عملکردی (مداد کاغذی، مثل ساختن یک جدول مشخصات آزمون)، ۲) آزمون های شناسایی (مثل شناسایی یک ابزار و توصیف کارکرد آن)، ۳) انجام عملکرد در موقعیت شبیه سازی شده (مثل انجام حرکات شنا بیرون از آب) و ۴) نمونه کار (مثل آزمون رانندگی در موقعیت هایی مشابه موقعیت واقعی) معرفی کرده است. نیکتو نیز روش ها و فنون سنجش عملکردی را در چهار دسته توضیح می دهد ۱) سنجش ساختارمند یا کنترل شده، ۲) سنجش در موقعیت های طبیعی (مثل مشاهده دانش آموز در زمین بازی و رابطه وی با همسالان)، ۳) پروژه های طولانی مدت (مثل انجام یک تحقیق کتابخانه ای)، و ۴) کارپوشه. در سنجش ساختارمند یا کنترل شده معلم یا سنجش گر هم بر تکالیف سنجش و هم بر شرایطی که دانش آموز باید پاسخ دهد کنترل اعمال می کند. در این نوع سنجش عملکردی از آزمون شونده خواسته می شود نه تنها پاسخ خود را به طور کتبی بنویسد، بلکه درباره آن توضیح دهد و دلیل بیاورد، با این روش هم فرایند اندیشه سنجش می شود و هم فرآورده آن.

کارپوشه به مجموعه ای از کارهای دانش آموز یا دانشجو گفته می شود که کوشش، پیشرفت و موفقیت تحصیلی او را در یک زمینه خاص نشان می دهد. یکی از امتیازهای این روش این است که علاوه بر فرآورده یادگیری می توان فرایند یادگیری را نیز مورد سنجش قرار داد.

سنجش مشاهده‌ای

روش‌های سنجش مشاهده‌ای شامل روش‌هایی است که معلم یا سنجشگر به کمک آن مشاهدات خود را نظم می‌دهند. معروف‌ترین این روش‌ها فهرست واری، مقیاس درجه‌بندی و واقعه نگاری هستند.

فهرست واری: شامل فهرستی از رفتارها و ویژگی‌ها، فعالیت‌ها یا تولیدات دانش‌آموزان، همراه با جایی در مقابل هر ماده فهرست، برای تعیین یا چک کردن بود یا نبود آن ماده است. در فهرست واری هم می‌توان ویژگی‌های مثبت فرایند یا فرآورده را مورد نظر قرار داد هم ویژگی‌های منفی و هم ترکیب آنها.

مقیاس درجه‌بندی: در مقیاس درجه‌بندی درجه یا مقدار ویژگی مورد نظر سنجش می‌شود. این مقیاس‌ها انواع متنوعی دارد که مهم‌ترین آنها مقیاس درجه‌بندی عددی، مقیاس درجه‌بندی نگاره‌ای، و مقیاس درجه‌بندی نگاره‌ای توصیفی است.

واقعه نگاری: به توصیف‌های واقعی از رویدادها و اتفاقات معنی‌داری که معلم یا سنجشگر در نتیجه مشاهده رفتار دانش‌آموزان به دست می‌آورند گفته می‌شود.

هدف‌های کلی کاربرد آزمون‌ها

آزمون‌های روانی در موقعیت‌های آموزشی، راهنمایی تحصیلی و شغلی، و در موقعیت‌های بالینی توسط معلمان، مشاوران، روان‌شناسان بالینی و سایر متخصصان برای هدف‌های مختلف مورد استفاده قرار می‌گیرند. کرانباخ موارد استفاده از آزمون‌ها را به شرح زیر مطرح کرده است:

۱- پیش‌بینی: آزمون‌ها برای اندازه‌گیری توانایی پیشرفت تحصیلی و سایر ویژگی‌های افراد که پایه تصمیم‌گیری است به کار می‌رود. پیش‌بینی مبتنی بر داده‌های کمی معتبرتر و دقیق‌تر از تفکر و نظرهای مبهم است.

۲- گزینش: در سازمان‌های استخدام‌کننده آزمون‌ها برای گزینش یا رد بعضی از افراد به کار می‌روند. تصمیم‌گیری در مورد پذیرش یک فرد، تصمیم در مورد گزینش او به شمار می‌رود.

۳- طبقه‌بندی: طبقه‌بندی عبارت از قرار دادن افراد در گروه‌های مختلف براساس ملاک‌های منطقی است. به عنوان مثال برای طبقه‌بندی می‌توان به مواردی از قبیل شناخت انواع بیماری‌های روانی (آزمون‌های تشخیصی) و انتخاب

برنامه درسی مدارس یا دانشکده‌ها و یا گزینش افراد برای انواع حرفه‌های تخصصی در خدمات نظامی و غیره اشاره کرد.

۴- ارزشیابی: آزمون به منظور قضاوت و ارزشیابی در مورد برنامه‌ها، روش‌ها و مانند اینها به کار می‌رود.

نمونه سؤال

۱- اگر هدف آزمون تعیین این باشد که فرد در گذشته چه چیزهایی یاد گرفته است، به کارگیری کدام تست مناسب‌تر است؟

(۱) استعداد (۲) پیشرفت تحصیلی (۳) رغبت (۴) شخصیت

۲- کدام یک از آزمون‌های زیر جز آزمون‌های غیرشناختی به شمار می‌رود؟

(۱) هوش (۲) فرافکن (۳) معلومات (۴) استعداد

۳- کدام یک از انواع ارزشیابی تقریباً در حاشیه قرار دارد و مکمل عمل آموزشی نیست؟

(۱) تشخیصی (۲) تکوینی (۳) مستمر (۴) مجموعی

۴- امتحانات کوتاهی که معلمان در طول یک دوره تحصیلی به عمل می‌آورند، چه نوع ارزیابی است؟

(۱) تشخیصی (۲) تکوینی (۳) تراکمی (۴) پایانی

۵- هنگامی که از بین گروهی داوطلب استخدام در یک سازمان تعدادی را انتخاب کنیم کدام جنبه یا کارکرد آزمون مورد استفاده قرار می‌گیرد؟

(۱) تشخیص (۲) گزینش (۳) طبقه‌بندی (۴) ارزشیابی

نمونه سؤال کنگور، دکتری روان‌شناسی تربیتی

۳- کدام ویژگی، آزمون‌های ملاک مرجع را از آزمون‌های نرم مرجع متمایز می‌کند؟ (۱۴۰۳)

(۱) نوع تکالیف (۲) نوع سؤالات

(۳) شیوه نمره گذاری (۴) ملاک قضاوت

۴- ملاک قبولی در چهار امتحان در زیر ارائه شده است، با توجه به این ملاک‌ها، کدام موارد از نوع هنجار مرجع هستند؟ (سال ۱۳۹۷)

(۱) ۱۵ نفری که دارای بالاترین نمره هستند.

(۲) افرادی که به ۷۰ درصد از سؤال‌ها پاسخ درست داده‌اند.

(۳) افرادی که نمره ۲۴ و بالاتر کسب کرده‌اند.

(۴) افرادی که نمره آنها یک انحراف بالاتر از میانگین.

الف) یک و سه (۲) دو و چهار (۳) یک و چهار (۴) دو و سه

۵- ملاک مورد استفاده در اندازه‌گیری صفت مورد بررسی در آزمون‌های هنجاری، ملاکی، استاندارد شده و معلم ساخته از چه نوعی است؟ (سال ۱۳۹۶)

(۱) نسبی - نسبی - مطلق - مطلق (۲) نسبی - مطلق - نسبی - مطلق

(۳) مطلق - نسبی - نسبی - مطلق (۴) مطلق - نسبی - مطلق - نسبی

۶- در کدام مورد، فرایند انجام کار مورد ارزشیابی قرار گرفته است؟ (سال ۱۳۹۸)

(۱) از دانش آموز می‌خواهیم درباره امید انشاء بنویسد و آن را ارزشیابی می‌کنیم.

(۲) از هنرجو می‌خواهیم یک مدار الکتریکی دو پل بسازد و به آن امتیاز می‌دهیم.

(۳) از هنرجو می‌خواهیم تابلوی یک منظره را نقاشی کند و به آن امتیاز می‌دهیم.

(۴) از دانشجوی پرستاری می‌خواهیم فشار خون بیمار را اندازه بگیرد و به آن امتیاز می‌دهیم.

۷- کدام مورد، ویژگی‌های سنجش عملکردی را براساس پژوهش‌های صاحب‌نظران تأیید می‌کند؟ (سال ۱۴۰۰)

- (۱) تأکید بر سنجش مستقیم، استفاده از مسائل ذهنی، تأکید بر کاربرست.
- (۲) تأکید بر کاربرست، استفاده از مسائل واقعی و سنجش غیرمستقیم.
- (۳) **تأکید بر سنجش مستقیم، استفاده از مسائل واقعی، ترغیب و تشویق تفکر باز.**
- (۴) تأکید و تشویق تفکر باز، استفاده از مسائل ذهنی، تأکید بر کاربرست.

مراحل ساخت آزمون

فهرست زیر مراحل ساخت آزمون را نشان می‌دهد:

- (۱) سازه (خصیصه مکنون) که آزمون برای سنجش آن طراحی می‌شود را تعریف کنید و دامنه و حدود آن را مشخص کنید.
- (۲) آزمون را طرح‌ریزی کنید، حوزه‌هایی را که سؤال‌های آزمون مورد اندازه‌گیری قرار می‌دهد به طور منظم طرح‌ریزی کنید (تهیه جدول هدف - محتوا).
- (۳) برای هر حیطه سؤال بنویسد، یک ونیم تا سه برابر سؤال‌هایی که در آزمون نهایی به کار خواهید برد.
- (۴) تمام سؤال‌ها را روی یک گروه نمونه معرف جامعه که کمتر از ۵۰ نفر نباشد، اجرا کنید.
- (۵) سؤال‌ها را تجزیه و تحلیل و انتخاب کنید (شاخص‌های درجه دشواری و ضریب تشخیص، اعتبار و روایی، تحلیل عاملی، منحنی‌های ویژگی سؤال).
- (۶) آزمون اصلاح شده را روی نمونه دیگری (معرف نماینده واقعی جامعه) اجرا کنید. این مرحله هنجاریابی (test norming) یا میزان کردن (test standardization) نامیده می‌شود.

طرح‌ریزی آزمون

در این مرحله حوزه و حیطه‌ای که سؤال‌های آزمون قرار است اندازه‌گیری کند، به طور منظم طرح‌ریزی می‌شود. این طرح‌ریزی در تعیین روایی محتوایی (منطقی) یا نمونه‌گیری از حیطه دارای اهمیت است. به ویژه در رابطه با آزمون-های پیشرفت تحصیلی. به منظور طرح‌ریزی آزمون اغلب از جدولی با عنوان جدول مشخصات آزمون (امتحان)

استفاده می‌شود. این جدول شامل هدف‌های رفتاری است که آزمون قصد اندازه‌گیری آن را دارد و محتوای درسی که قرار است اندازه‌گیری شود. این جدول هدف‌های رفتاری را به محتوای درس ربط می‌دهد. سپس وزنی که باید به هر یک از زمینه‌های یادگیری داده شود مشخص شده و براساس آن درصد و تعداد سؤال‌های هر بخش تعیین می‌شود. وزن‌دهی براساس اهمیت نسبی هر هدف و محتوا، حجم مطلب، و مقدار زمان آموزش صرف شده برای آن، صورت می‌گیرد.

اهداف آموزشی

هدف آموزش و پرورش به طور کلی ایجاد تغییرات مطلوب در رفتار دانش‌آموز است و هدف امتحان اندازه‌گیری و سنجش این تغییرات است. هدف‌های آموزشی را می‌توان در سه سطح تعریف کرد.

۱) در سطح اول هدف‌های آموزشی به صورت کلی و بیشتر انتزاعی (غایت‌ها و بیانات آرمانی) تعریف می‌شوند. این هدف‌های کلی و انتزاعی به صورت هدف‌های دراز مدت بیان می‌شوند و به طور معمول محصول نهایی یک دوره کامل آموزشی را توصیف می‌کنند. این اهداف معمولاً به وسیله برنامه‌ریزان در سطح وزارت تهیه و تدوین می‌شود. مثلاً هدف کلی درس زبان فارسی کسب مهارت‌های اساسی در خواندن و نوشتن توسط دانش‌آموز است.

۲) در سطح دوم هدف‌های کلی به هدف‌های جزئی‌تر و در رابطه با رفتارهای قابل مشاهده و اندازه‌گیری دانش‌آموز تعریف می‌شود. به عنوان مثال دانش‌آموز بتواند کلمات فارسی را با املاء درست بنویسد و آنچه را می‌خواند بفهمد.

۳) در سطح سوم هدف‌های آموزشی به صورت رفتارهای کاملاً مشخص و در رابطه با محتوای دقیق هر یک از بخش‌های یک درس خاص به گونه‌ای روشن، دقیق و دور از ابهام تعریف می‌شود. مثلاً از ده کلمه جدید درس که به او دیکته شده، دست کم ۸ کلمه را با املاء درست بنویسد.

نکته: در تعریف هدف‌های رفتاری باید از فعل‌هایی استفاده شود که اولاً رفتار مورد نظر را به روشنی توصیف کند، ثانیاً از نظر همه افراد مرتبط مفهوم یکسانی داشته باشد. به عنوان مثال فعل‌های نام می‌برد، شرح می‌دهد، مثال

می‌زند، تعریف می‌کند، مقایسه می‌کند و... قابل استفاده است. اما مواردی نظری بفهمد، یادبگیرد، کسب کند روشن و واضح نیست.

منابع اهداف آموزشی شامل:

- (۱) نیازهای جامعه
- (۲) نیازهای یادگیرنده
- (۳) نیازهای ماده درسی

سطوح مختلف هدف‌های آموزشی (تربیتی)

هدف‌های آموزشی توصیفی از رفتار فرد در رابطه با کاری که می‌تواند با اطلاعات موجود در مطالب درسی (محتوا) انجام دهد است. مقصود رفتار و اعمال مطلوبی است که باید در پایان هر برنامه درسی در دانش‌آموزان مشاهده شود. از هدف‌های آموزشی (تربیتی) طبقه‌بندی‌های مختلفی صورت گرفته است که مهم‌ترین آنها به کارهای بلوم بر می‌گردد. طبقه‌بندی بلوم در سه حیطه مشخص شده است:

- (۱) حیطه شناختی
- (۲) حیطه عاطفی
- (۳) حیطه روانی - حرکتی

حیطه شناختی:

هدف‌های حوزه شناختی به جریان‌هایی که با فعالیت‌های ذهنی و فکری آدمی سر و کار دارند مربوط می‌شوند. آزمون‌های پیشرفت تحصیلی متداول که به آزمون‌های مداد-کاغذی معروفند به طور عمده در حیطه شناختی قرار می‌گیرند. در طبقه‌بندی اولیه بلوم اهداف حیطه شناختی در شش سطح از ساده به دشوار طبقه‌بندی می‌شدند. این اهداف به ترتیب عبارت بودند از: دانش، فهمیدن، به کار بستن، تحلیل، ترکیب و ارزشیابی. در طبقه‌بندی تجدیدنظر شده از حوزه شناختی دو بُعد دانش و فرایندهای شناختی وجود دارد. بعد دانش شامل چهار نوع دانش (دانش واقعی،

دانش مفهومی، دانش روندی، و دانش فراشناختی است) و بعد فرایندهای شناختی شامل ۶ سطح (به یادآوردن، فهمیدن، به کار بستن، تحلیل کردن، ارزشیابی کردن، و آفریدن/خلق کردن) است.

(۱) دانش: منظور از دانش این است که یادگیرنده بتواند از طریق یادآوری یا بازشناسی شواهدی ارائه دهد حاکی از اینکه او اندیشه‌ها یا پدیده‌ای را که در جریان آموزش تجربه کرده، به خاطر سپرده است. یادگیری در این سطح بیشتر جنبه حفظی دارد. مثال: دانش واژه‌ها، آشنایی با تعاریف، تاریخ رویدادها مهم، دانش اصطلاحات. اصطلاحات درس‌ها را تعریف کند، یا محصولات عمده یک استان را نام ببرد، یک ضرب المثل را به یاد بیاورد. افعال: تعریف می‌کند، فهرست می‌کند، نام ببرد، از حفظ بخواند و...

(۲) فراگیری (درک، فهمیدن): فراگیری مستلزم درک معنا و مفهوم آشکار و پنهان یک مطلب و بیان آن با جمله‌ها و عبارت‌های توسط خود فراگیر است. تبدیل مطالب آموخته شده از شکلی به شکل دیگر، تفسیر، تبیین، تشریح، ترجمه کردن، خلاصه کردن، استنباط کردن، تعمیم دادن، توضیح دادن. مثال: یک قطعه شعر را به صورت نثر بنویسد، مفهوم نمونه را به زبان خود شرح دهد، مطالب درس را خلاصه کند. معنای ضرب المثل را بگوید.

(۳) به کار بستن: فراگیر بتواند اطلاعات و آموخته‌های خود را در موارد عملی و یا موقعیت‌های جدید به کار ببرد، کاربرد آموخته‌ها در موقعیت‌های زندگی. مثال حل مسائل ریاضی در نمونه جدید، با استفاده از قواعد هندسی مساحت یک زمین را محاسبه کند.

(۴) تحلیل: خرد کردن مباحث به اجزای سازنده آنها و کشف روابط اجزا و سازمان‌بندی آنها. مثال چند موجود زنده را از روی شباهت و تفاوت طبقه‌بندی کند. عناصر مهم یک متن را تشخیص دهد، جمله ادبی را از نظر دستوری تجزیه کند.

(۵) ترکیب: به هم پیوستن عناصر و اجزا به منظور تشکیل یک واحد کل، خلاقیت و آفرینندگی. گردآوری اطلاعات متعدد و پیوند دادن آنها. به عنوان مثال نوشتن مقاله، نوشتن شعر، به طور کلی ابداع و خلق یک اثر علمی، هنری، اجتماعی و ...

(۶) ارزشیابی (قضاوت): داوری آگاهانه درباره اندیشه‌ها، کارها، روش‌ها و راه‌حل‌ها برای مقصود معین. مثل مقایسه نظریه‌ها، داوری پایان نامه، مقاله انتقادی، عقاید مطرح شده در دو متن را مقایسه کنید.

طبقه تجدیدنظر شده در بُعد دانش
دانش امور واقعی: عناصر اساسی مورد نیاز یادگیرنده برای آشنا شدن با یک رشته علمی یا حل مسائل مربوط به آن. (۱) دانش اصطلاحات: دانش علامت‌ها و نمادهای کلامی و غیرکلامی در رشته‌های مختلف (مثل نمادهای آماری) (۲) دانش اجزا و عناصر خاص: دانش رویدادها، مکان‌ها، اشخاص، تاریخ‌ها و منابع اطلاعاتی (مثل دانستن مراکز استان‌ها)
دانش مفهومی: دانش مقوله‌ها، طبقه‌ها، و روابط میان آنها (۱) دانش طبقه‌بندی و طبقه‌ها: دانش طبقه‌ها، دسته‌ها، تقسیمات و ترتیبات خاصی در رشته‌ها و موضوعات گوناگون (مثل طبقه‌بندی انواع جانوران) (۲) دانش اصل‌ها و تعمیم‌ها: دانش انتزاع‌های ویژه‌ای است که به کمک آنها مشاهدات مربوط به پدیده‌ها خلاصه می‌شوند (مثل قضیه حد مرکزی در آمار) (۳) دانش نظریه‌ها، الگوها و ساخت‌ها: دانش اصل‌ها و تعمیم‌های همراه با روابط میان آنها که معرف یک دیدگاه نظام‌مند، منسجم و روشن از یک پدیده، مسأله یا موضوع پیچیده است (مثل نظریه کلاسیک اندازه‌گیری)
دانش روندی: چگونگی انجام دادن کارها، روش‌های تحقیق و ملاک‌ها برای به کار بردن مهارت‌ها، الگوریتم‌ها، تکنیک‌ها و روش‌ها است. (۱) دانش مهارت‌ها و الگوریتم‌های خاص یک موضوع: دانش روند یا روال انجام دادن کاری بدون تأکید بر توانایی استفاده از آن (مثل دانش مراحل حل یک تقسیم، مراحل برنامه نویسی) (۲) دانش فن‌ها و روش‌های خاص یک موضوع: دانشی که عمده‌تاً حاصل توافق، هم‌رأیی یا هنجارهای درونی یک نظام علمی است تا دانشی که به طور مستقیم از راه مشاهده، آزمایش یا اکتشاف به دست می‌آید (مثال دانش روش‌های پژوهش در علوم اجتماعی) (۳) دانش ملاک‌های تعیین زمان استفاده از روش مناسب: دانش در این مورد که چه وقت روش‌ها و شیوه‌های خاص را به کار گیرند (مثال چه وقت از کدام روش آماری استفاده شود)
دانش فراشناختی: دانش شناخت به طور کلی و همین‌طور آگاهی و دانش فرد درباره شناخت خود. (۱) دانش راهبردی: دانش به راهبردهای کلی برای یادگیری، تفکر و حل مسأله در حیطه‌های مختلف (به عنوان مثال دانش به اینکه مرور و تکرار سبب افزایش یادگیری) (۲) دانش مربوط به تکالیف شناختی شامل دانش موقعیتی و مشروط: تعیین اینکه چه وقت و چه موقع این راهبردها را در شرایط مختلف به کار بگیرند (مثلاً سبک مطالعه برای آزمون تشریحی و تستی متفاوت است) (۳) دانش خود (خودآگاهی): آگاهی یک فرد از سطح دانش خود. دانش به اینکه شخص در بعضی زمینه‌ها آگاه و در بعضی زمینه‌ها بی اطلاع است.

حیطه عاطفی:

این حیطه هدف‌هایی مانند انگیزه‌ها، احساسات، ارزش‌ها و نگرش‌ها، گرایش به خواندن و تمایل و رغبت نسبت به پدیده‌ها و مقررات را در برمی‌گیرد. بازده‌های یادگیری متعلق به حوزه‌های غیرشناختی مانند علاقه‌ها و ارزش‌گذاری‌ها در حیطه عاطفی است.

(۱) دریافت و توجه کردن: حساسیت به وجود بعضی پدیده‌ها و محرک‌های محیطی و میل به دریافت کردن آنها یا توجه کردن به آنها. این دسته هدف‌ها شامل رفتارهایی هستند که از وجود پدیده خبر می‌دهند و میل به تحمل در برابر دیدن و یا شنیدن را نمایان می‌سازند. به عنوان مثال یادگیرنده مایل است به اتفاقاتی که در کلاس درس می‌افتد توجه کند، یادگیرنده می‌پذیرد که دانش‌آموزان نژادهای مختلف وجود دارند.

(۲) پاسخ دادن: توجه فعالانه به پدیده‌ها و انجام کارهایی درباره آنها، یا فعالیت حاکی از پذیرش یک عقیده یا خط مشی. پس از دریافت و توجه نسبت به یک موضوع از شاگرد انتظار می‌رود که گامی فراتر نهد و احساس خود را فعالانه نمایان سازد. در این مرحله درک مسأله کافی نیست و شخص باید واکنشی در مورد مسأله مورد نظر نشان دهد. به عنوان مثال یادگیرنده به طور داوطلبانه کتاب‌ها و مجله‌های خاص کودکان را مطالعه می‌کند، یا در زمینه خاصی به فعالیت و پژوهش بپردازد.

(۳) ارج نهادن (ارزش‌گذاری): احساس یا باور پایدار حاکی از ارزشمند بودن یک چیز، یک اندیشه، یک شخص، یا یک گروه، یا ابراز عقیده یا نگرش درباره ارزش یک چیز. ارزش‌گذاری مبتنی بر درونی کردن دسته معینی از ارزش‌هاست. به عنوان مثال یادگیرنده در یک بحث گروهی از عقاید یکی از دوستان خود که به آن باور دارد دفاع می‌کند.

(۴) سازماندهی ارزش‌ها: دسته‌بندی و نظام‌دار کردن ارزش‌ها به گونه‌ای که بعضی از آنها مهمتر از بعضی دیگر جلوه کنند. به عنوان مثال یادگیرنده نیاز به یافتن نوعی تعامل بین رشد اقتصادی و حفظ منابع طبیعی تشخیص می‌دهد، و یا تنظیم کردن زمان استراحت و فعالیت‌های آموزشی نمونه دیگری از این سطح حیطه عاطفی است.

(۵) درونی شدن ارزش‌ها (شخصیت پذیرفتن): تبدیل نظام ارزشی به صورت یک سبک زندگی یا یک دیدگاه فلسفی. به عنوان مثال یادگیرنده در موقعیت‌های مختلف متناسب با نظام ارزشی خود عمل می‌کند.

حیطه روانی حرکتی:

حیطه روانی حرکتی به زمینه مهارت‌های حرکتی یا بدنی مربوط است. به سخن دیگر، هر فعالیتی که علاوه بر جنبه روانی دارای جنبه جسمانی نیز باشد، در این حوزه جای دارد. مهارت‌های مختلف مثل انجام آزمایش در آزمایشگاه، کار در کارگاه، فعالیت‌های ورزشی و ایجاد ارتباط کلامی و غیر کلامی در حیطه اهداف روانی حرکتی است. هدف - های این حیطه به کمک ابزار سنجش از جمله پرسشنامه، مشاهده، مقیاس درجه بندی، فهرست واریسی، واقعه نگاری، ثبت حضور و غیاب، گزارش‌های کار عملی و مانند این‌ها سنجش می‌شوند. طبقه بندی هارو (۱۹۷۷) از هدف‌های روانی حرکتی شامل: (۱) حرکات بازتابی، (۲) حرکات اساسی - بنیادی، (۳) توانایی ادراکی، (۴) توانایی بدنی، (۵) حرکات مهارتی، و ارتباطات غیر کلامی است. طبقه بندی سیمپسون از حوزه روانی حرکتی شامل موارد زیر است:

- ۱) ادراک (حسی): استفاده از حواس برای هدایت کنش‌های حرکتی، به عنوان مثال یادگیرنده با با چشیدن غذا مزه آن را تشخیص می‌دهد.
- ۲) آمادگی یا آمایه: آماده بودن برای انجام یک عمل یا یک رشته اعمال، به عنوان مثال یادگیرنده به تویی که به سمت او پرت می‌شود ضربه می‌زند.
- ۳) پاسخ هدایت شونده: عمل کردن به کمک هدایت یک الگو، به عنوان مثال راکت تنیس را مطابق آنچه مربی می‌گوید حرکت می‌دهد.
- ۴) مکانیسم یا عادت: انجام اعمالی که نسبتاً مشخص و عادت‌ی و خودکار هستند.
- ۵) پاسخ پیچیده آشکار: انجام رشته اعمالی که تا حدودی به صورت خودکار یا عادت‌ی درآمده یا ترکیب تعدادی از عادت‌ها، مثلاً یادگیرنده بتواند طول یک استخر را شنا کند.
- ۶) انطباق: استفاده از مهارت‌های قبلاً آموخته شده برای انجام تکالیف تازه اما مربوط به تکالیف گذشته.
- ۷) ابتکار: خلق الگوهای تازه حرکتی برای حل مسائل غیر معمول و برخورد با موقعیت‌های جدید. مثل ساخت یک قطعه موسیقی و اجرای آن.

۸- کدام مورد در خصوص واقعه نگاری به عنوان یکی از روش‌های سنجش مشاهده‌ای، درست است؟ (سال ۱۴۰۲)

۱) برای سنجش هدف‌های آموزشی حوزه عاطفی و روانی - حرکتی مناسب است.

- ۲) یک روش مشاهده غیرمستقیم برای سنجش رویدادها و رفتارهای واقعی است.
- ۳) عمدتاً برای ارزشیابی رویدادها و اتفاقات واقعی در محیط‌های آموزشی به کار می‌رود.
- ۴) برای کمک به ایجاد مهارت خودارزشیابی و نظارت بر پیشرفت آموزشی استفاده می‌شود.

۹- اگر یک دانشجوی روان‌شناسی پس از گذراندن چهار نیمسال تحصیلی، داوطلبانه کتاب‌های روان‌شناسی شخصیت را تهیه و مطالعه و تحولات این گرایش تخصصی را مستمراً پیگیری نماید. این موضوع از دیدگاه بلوم در کدام حوزه اهداف آموزش و پرورش طبقه‌بندی می‌شود و در چه سطحی است؟ (سال ۱۴۰۰)

- ۱) شناختی - فهمیدن
- ۲) شناختی - به کار بستن

۳) عاطفی - پاسخ دادن

- ۴) عاطفی - ارزش گذاردن

۱۰- عبارت « دانش آموز باید بتواند از ۱۰ علامت اختصاری شیمی، حداقل ۹ علامت را به درستی تشخیص دهد». چه نوع هدفی است و در کدام سطح شناختی طبقه بندی تجدید نظر شده بلوم دسته بندی می‌شود؟ (سال ۱۴۰۰)

- ۱) کامل - به یاد آوردن
- ۲) کامل - فهمیدن

- ۳) ناقص - به یاد آوردن
- ۴) ناقص - فهمیدن

۱۱- در طبقه بندی بلوم، هدف زیر در کدام سطح فرایند شناختی قرار دارد؟ (سال ۱۳۹۸)

(تفسیر خروجی نرم افزار مربوط به پایایی یک آزمون)

- ۱) فهمیدن
- ۲) کار بستن
- ۳) ارزشیابی کردن
- ۴) تجزیه و تحلیل کردن

۱۲- وقتی دانشجویی با شرکت در یک کارگاه مقاله نویسی، به این باور می رسد که مهارت در مقاله نویسی بسیار ارزشمند است، این باور دانشجو در کدام حوزه اهداف آموزش و پرورش از نظر بلوم طبقه بندی می شود؟ (سال ۱۳۹۷)

(۱) شناختی (۲) عاطفی (۳) شناختی و عاطفی (۴) روانی - حرکتی

انواع سؤال‌های آزمون

مرحله بعدی در ساخت آزمون نوشتن سؤال‌های آزمون است، که یک و نیم تا سه برابر سؤال‌های نهایی باید طراحی شود. سؤال‌های آزمون انواع مختلفی دارد. روش‌های مختلف اندازه‌گیری و سنجش از جمله آزمون‌ها را می‌توان با توجه به نوع سؤال یا ماده دسته‌بندی کرد. به طور کلی، وسایل مختلف اندازه‌گیری با توجه به نوع سؤال یا ماده به دو دسته بسته پاسخ و باز پاسخ تقسیم می‌شوند. در نوع بسته پاسخ هم سؤال و هم پاسخ در اختیار پاسخ دهنده قرار می‌گیرد و باید درباره آن تصمیم بگیرد نظیر: آزمون‌های صحیح-غلط، چندگزینه‌ای، و جورکردنی. در آزمون‌های باز پاسخ، پاسخ دهنده خود پاسخ سؤال را تولید می‌کند، مانند سؤال‌های تشریحی، کوتاه پاسخ و کامل کردنی. در طبقه‌بندی دیگری از پرسش‌های آزمون، آزمون‌ها به دو دسته عینی و انشائی تقسیم می‌شوند:

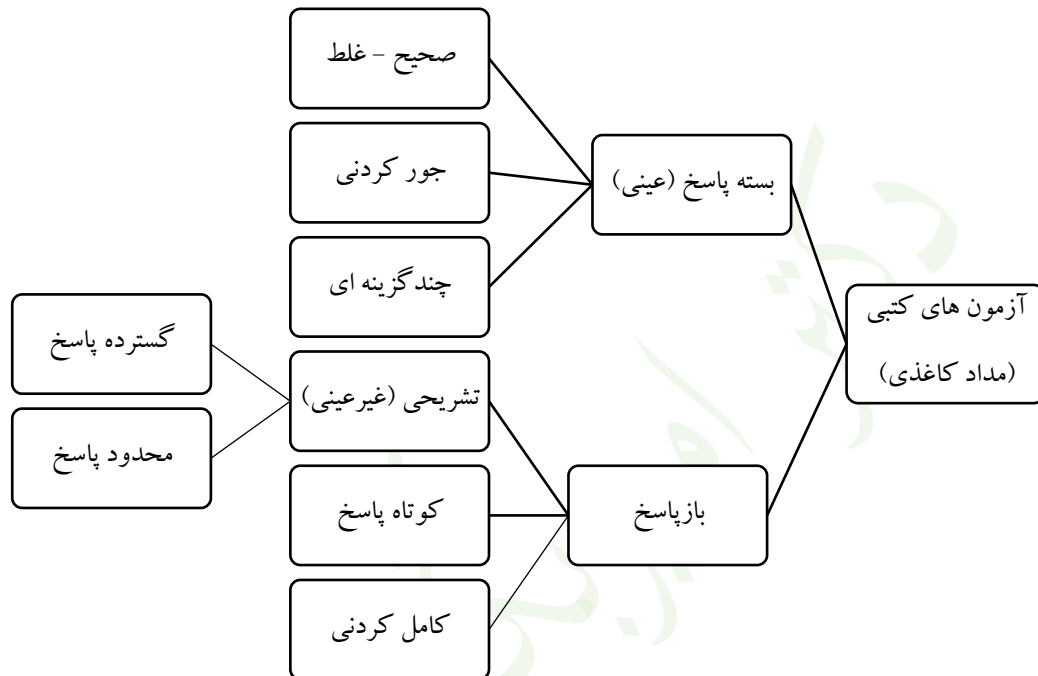
سؤال‌های عینی

سؤال‌های عینی به سؤال‌هایی گفته می‌شود که نمره‌گذاری آن مستقل از نظر شخصی و قضاوت ذهنی ارزشیاب است. یعنی تصحیح‌کنندگان مختلف به آن نمره یکسان می‌دهند.

پرسش‌های عینی به طور کلی یا از نوع یادآوری هستند و یا از نوع تشخیص یا شناسایی:

یادآوری: در این نوع سؤال، پاسخ دهنده جواب را از میان آموخته‌های قبلی خود پیدا می‌کند و مشخص می‌سازد، مانند سؤال‌های کوتاه پاسخ و سؤال‌های کامل کردنی.

تشخیص و شناسایی: در این نوع، برای هر سؤال تعدادی پاسخ پیشنهاد شده است و آزمایش شونده باید پاسخ درست را از بین آن‌ها پیدا کرده و مشخص سازد. مانند سؤال‌های چندگزینه‌ای، صحیح - غلط و جور کردنی.



انواع سؤال‌های عینی:

• سؤال‌های چندگزینه‌ای

آزمون چندگزینه‌ای معروف‌ترین و پرکاربردترین آزمون نوع بسته پاسخ است. هر پرسش چندگزینه‌ای شامل دو بخش است، متن اصلی (تنه یا ساقه) و گزینه‌ها (گزینه کلید، و گزینه‌های انحرافی). قسمت اصلی به دو صورت نوشته می‌شود یکی جمله پرسشی (استفهامی) و دیگری جمله ناقص. سؤال‌های چندگزینه‌ای انواع مختلفی دارد: نوع پرسشی، نوع ناتمام، نوع تنها گزینه درست، نوع بهترین گزینه، نوع منفی، نوع جایگزینی و نوع پاسخ ترکیبی. از آزمون‌های چندگزینه‌ای برای هدف‌های آموزشی در سطح یادآوری، فهمیدن، به کار بستن و تحلیل می‌توان استفاده کرد.

• پرسش‌های صحیح - غلط

هر پرسش صحیح - غلط به صورت یک جمله مثبت کامل است که پاسخ دهنده باید صحیح و غلط بودن آن را تشخیص دهد. به عبارتی یک پرسش چندگزینه‌ای با دو گزینه (درست - نادرست، بلی - خیر، موافق - مخالف و ...)

است. این نوع پرسش‌ها در رابطه با هدف‌های آموزشی حوزه شناختی: یادآوری، فهمیدن، کاربرستن، و در حوزه عاطفی: دریافت کردن، پاسخ دادن، ارزش‌گذاری باور و نگرش، و سازمان دادن قابل استفاده است.

- پرسش‌های جورکردنی

آزمون‌های جورکردنی از تعدادی تمرین تشکیل می‌شود که هر تمرین شامل سه بخش است: راهنمای انجام تمرین، فهرست پرسش‌ها، و فهرست پاسخ‌ها. پاسخ‌دهنده باید پاسخ مربوط به هر پرسش را مشخص کرده و علامت بزند. پرسش‌های جورکردنی انواع مختلفی دارد که دو نوع معروف آن معمولی و نوع رده‌بندی یا فهرست اصلی است.

در سؤال‌های جورکردنی معمولی باید تعداد پاسخ‌ها از پرسش‌ها بیشتر باشد تا پاسخ آخرین پرسش به خودی خود معلوم نشود، و ترتیب قرار گرفتن هر پاسخ نسبت به جواب باید به طور تصادفی و نامنظم باشد.

در روش رده‌بندی، فهرستی از پاسخ وجود دارد (مثلاً انواع جانوران)، که از هر پاسخ بیش از یک بار می‌توان استفاده کرد، از دانش‌آموز خواسته می‌شود تا هر سؤال را طبق این طرح رده‌بندی کند، به همین منظور تعداد پرسش‌ها از تعداد پاسخ‌ها می‌تواند بیشتر باشد.

سؤال‌های جورکردنی بیشتر برای سنجش توانایی انتقال یادگیری، توضیح و درک مطلب مناسب‌اند. این نوع پرسش‌ها برای اهداف آموزشی یادآوری، کاربرستن، و تحلیل مناسب است.

- پرسش‌های کوتاه پاسخ

سؤال‌های کوتاه پاسخ نوع دیگری از پرسش‌های عینی هستند که دانش‌آموز خود باید پاسخ آن را تهیه کند. برخی این نوع آزمون‌ها را حد وسط آزمون تشریحی و آزمون عینی می‌دانند. این آزمون‌ها به سه دسته پرسشی، کامل کردنی و تشخیصی یا تداعی تقسیم می‌شوند. در پرسش‌های کوتاه پاسخ از نوع کامل کردنی، سؤال به صورت یک جمله ناقص نوشته می‌شود و آزمودنی باید جای خالی را پر کند. پرسش‌های کامل کردنی را تکمیلی یا از نوع یادآوری نیز می‌نامند. در پرسش‌های نوع تداعی سؤال به صورت یک موضوع و مطلب ارائه می‌شود و آزمون‌شونده باید جواب را تداعی کند، مثلاً نام استان داده می‌شود، آزمودنی باید مرکز استان را بنویسد. این آزمون‌ها برای

اندازه‌گیری هدف‌های آموزشی سطح پایین به ویژه دانش مفیدند. برای مسائل مختصر ریاضی، تبدیل و تفسیر می‌توان آن را به سطح فهمیدن و به کار بستن افزایش داد.

سؤال‌های انشائی (تشریحی)

در سؤال‌های انشایی برخلاف پرسش‌های عینی آزمودنی به جای بازشناسی از میان پاسخ‌های احتمالی، باید اطلاعات را به یاد بیاورد، آنها را منظم کند، سازمان بدهد و بیان کند. سؤال‌های انشایی برای اندازه‌گیری اهداف سطح بالاتر نظیر ترکیب و آفرینندگی، و ارزشیابی به کار می‌رود. این نوع سؤال‌ها به دو دسته محدود پاسخ و گسترده پاسخ تقسیم می‌شوند. اگر پاسخ دهنده باید پاسخ سؤال را بر مبنای الگوی تعیین شده از طرف تهیه کننده سؤال، به موارد مشخص و معینی محدود سازد، به عبارت دیگر آزادی عمل پاسخ دهنده محدود است، به این سؤال انشائی کنترل شده یا محدود پاسخ می‌گویند. مزیت این روش کنترل شده مشخص بودن تکلیف و سهولت تصحیح و نمره گذاری آن است. نوع دوم، سؤال‌های گسترده پاسخ یا باز پاسخ است، در اینگونه سؤال‌ها آزمایش شونده موظف به پیروی از الگوی مشخصی نیست و آزادی کامل دارد. این نوع سؤال برای تمامی اهداف آموزشی در حوزه شناختی قابل استفاده است.

مقایسه سؤال‌های عینی و انشائی

سؤال‌های عینی	سؤال‌های انشائی
برای اندازه‌گیری همه اهداف و برآوردهای یادگیری به غیر از ترکیب و ارزشیابی مناسب است.	برای اندازه‌گیری ترکیب و ارزشیابی وسیله‌ای منحصر به فرد است.
نمونه گسترده‌ای از اهداف یادگیری را ارزیابی می‌کند.	نمونه محدودی از برآوردها یادگیری را ارزیابی می‌کند، در نتیجه ممکن است به روایی محتوایی لطمه بخورد.
عواملی نظیر تسط بر بیان مطلب، توانایی نویسندگی و زیبانویسی در ارزشیابی آن بی‌اثر است.	عواملی نظیر تسط بر بیان مطلب، توانایی نویسندگی و زیبانویسی در ارزشیابی آن مؤثر است.
عامل حدس و گمان در یافتن پاسخ درست مؤثر است.	عامل حدس و گمان در پاسخ به سؤال‌ها بی‌تأثیر است.
نمره گذاری و ارزشیابی عینی است.	نمره گذاری و ارزشیابی پاسخ‌ها ذهنی است.
نمره‌های امتحان از ثبات و دقت بیشتر برخوردار است.	نمره‌ها ثبات و دقت کمتری دارد.
تهیه و تنظیم سؤال‌ها و آماده کردن آزمون وقت گیر و دشوار است.	تهیه و تنظیم سؤال‌ها و آماده کردن آزمون با صرف وقت کمتر امکان پذیر است.

سؤال‌های عینی	سؤال‌های انشائی
تصحیح و نمره‌گذاری امتحان به سرعت و سهولت انجام می‌شود.	تصحیح و نمره‌گذاری امتحان مستلزم صرف وقت زیاد است.
عادات صحیح مطالعه را در زمینه یادآوری، تجزیه و تحلیل، تعبیر و تفسیر تقویت می‌کند.	عادات صحیح مطالعه را در همه زمینه‌ها به ویژه در زمینه ترکیب و ارزشیابی تقویت می‌کند.

نمونه سؤال کنگور دکتری روان‌شناسی تربیتی

۱۳- برای دستیابی به اهداف زیر، به ترتیب، کدام نوع سؤالات مناسب‌تر است؟ (سال ۱۴۰۲)

الف) طراحی و تهیه سریع و آسان آزمون

ب) طراحی آزمون با روایی محتوایی مطلوب

ج) اندازه‌گیری تفکر واگرا

د) تأکید بر شیوه‌ها و عادات مطلوب مطالعه

۱) تشریحی، چندگزینه‌ای، تشریحی، تشریحی

۲) تشریحی، چندگزینه‌ای، چندگزینه‌ای، تشریحی

۳) چندگزینه‌ای، تشریحی، تشریحی، چندگزینه‌ای

۴) چندگزینه‌ای، تشریحی، چندگزینه‌ای، چندگزینه‌ای

۱۴- کدام مورد درباره استفاده از آزمون‌های تشریحی در اندازه‌گیری پیشرفت تحصیلی درس ریاضیات درست است؟ (سال ۱۳۹۶)

۱) تهیه آن برای معلمان ریاضی دشوار است.

۲) بر روش مطالعه ریاضی دانش‌آموزان اثر مثبت دارد.

۳) نمونه معرفی از محتوای درس ریاضیات را اندازه می‌گیرد.

۴) از معلمان سایر دروس برای نمره‌گذاری آن می‌توان استفاده کرد.

۱۵- در کدام نوع از سؤال‌های جورکردنی، تعداد پرسش‌ها می‌تواند بیشتر از تعداد پاسخ‌ها باشد؟ (سال ۱۳۹۸)

(۱) خبری (۲) معمولی (۳) استفهامی (۴) رده‌بندی

۱۶- در کدام شرایط دادن حق انتخاب چند سؤال از میان تعدادی سؤال بیشتر توجیه پذیر است؟ (سال ۱۳۹۷)

(۱) قصد ایجاد مهارت تصمیم‌گیری در یادگیرندگان مدنظر است.

(۲) هدف مقایسه کیفیت و کمیت پاسخ‌های یادگیرندگان باهم باشد.

(۳) مشخص کردن یادگیرندگان قوی‌تر از یادگیرندگان ضعیف در نظر باشد.

(۴) هدف اندازه‌گیری مهارت یادگیرندگان در نوشتن و پروراندن مطالب باشد.

تحلیل سؤال‌های آزمون

یکی از مسائل مرتبط با ساخت آزمون‌ها انتخاب سؤال‌های مناسب برای باقی ماندن در آزمون است. هدف از تحلیل سؤال‌های آزمون حذف پرسش‌های نامناسب یا تجدید نظر در آنهاست. در نظریه کلاسیک اندازه‌گیری برای تحلیل سؤال‌های آزمون دو شاخص وجود دارد: (۱) درجه دشواری سؤال، (۲) قدرت تشخیص (ضریب تمیز)

سطح دشواری سؤال

سطح دشواری در نظریه کلاسیک اندازه‌گیری نسبت افرادی است که به سؤال پاسخ درست داده‌اند.

مثال: از ۱۲۵ آزمودنی، ۷۵ نفر به یک سؤال پاسخ درست داده‌اند، سطح دشواری سؤال چقدر است؟

$$P = \frac{75}{125} = 0.6$$

در تحلیل سؤال‌های آزمون یک روش تقسیم افراد به گروه قوی و ضعیف است. بدین صورت که پس از نمره-گذاری پاسخنامه‌ها و مرتب کردن آنها از بیشترین نمره به کمترین نمره، ۲۷ درصد بالا به عنوان گروه قوی (U) و

۲۷ درصد پایین به عنوان گروه ضعیف (L) انتخاب می‌شوند. محاسبه سطح دشواری بر اساس گروه قوی و ضعیف به صورت زیر است:

$$P = \frac{Nu + Nl}{u + L}$$

Nu = افرادی از گروه قوی که به سؤال پاسخ درست داده‌اند.

N_L = افرادی از گروه ضعیف که به سؤال پاسخ درست داده‌اند.

U = تعداد افراد گروه قوی

L = تعداد افراد گروه ضعیف

در حالتی که تعداد افراد دو گروه برابر است

$$P = \frac{Pu + Pl}{2}$$

دامنه سطح دشواری بین صفر و یک قرار دارد، هر چه سؤال آسان‌تر باشد سطح دشواری به یک نزدیک می‌شود، و هر چه سخت‌تر به صفر نزدیک می‌شود. یک سؤال با درجه دشواری $0/3$ سخت‌تر از سؤالی با درجه دشواری $0/8$ است. (در آزمون‌های شخصیت نسبت کسانی را منعکس می‌کند که گزینه خاصی را انتخاب کرده‌اند). اگر درجه دشواری نزدیک صفر باشد، یعنی هیچ آزمودنی به سؤال پاسخ درست نداده و اگر نزدیک یک باشد یعنی همه آزمودنی‌ها به آن پاسخ درست داده‌اند، سؤال‌های خیلی آسان و خیلی دشوار برای باقی ماندن در آزمون مناسب نیستند. در سطح دشواری $0/5$ واریانس سؤال (pq) حداکثر است، و حداکثر اطلاعات برای تفاوت‌های بین آزمودنی‌ها را فراهم می‌کند. به طور کلی درجات دشواری بین $0/3$ تا $0/7$ مناسب تلقی می‌شود.

میزان سطح دشواری خوب به هدف و کاربرد آزمون بستگی دارد. اما قاعده‌ای برای تعیین سطح دشواری مطلوب (بهینه) وجود دارد که از میانگین 100 درصد پاسخ‌های صحیح (سطح دشواری 1) و احتمال پاسخ تصادفی (شانس و تصادف) به دست می‌آید.

احتمال پاسخ تصادفی در انواع آزمون‌ها:

سؤال صحیح و غلط ($\frac{1}{2} = 0.5$) سؤال سه گزینه‌ای ($\frac{1}{3} = 0.33$)

سؤال چهار گزینه‌ای ($\frac{1}{4} = 0.25$) سؤال پنج گزینه‌ای ($\frac{1}{5} = 0.2$)

سطح دشواری مطلوب (بهینه):

$$P_{opt} = \frac{1 + P_r}{2}$$

مثال: سطح دشواری مطلوب در آزمون پنج گزینه‌ای چقدر است؟

$$P_{opt} = \frac{1 + 0.2}{2} = 0.6$$

به همین ترتیب سطح دشواری مطلوب در آزمون تشریحی برابر با ۰/۵، در آزمون دو گزینه‌ای برابر با ۰/۷۵، در سه گزینه‌ای برابر با ۰/۶۷ و در چهار گزینه‌ای برابر با ۰/۶۲۵ است.

برای اینکه سطح دشواری سؤال‌ها در یک آزمون در حد مطلوب (بهینه) باشد لازم است بین سطح دشواری سؤال و سطح بهینه تفاوت معنادار نباشد. برای پیدا کردن حدود اطمینان برای سطح اطمینان بهینه از فرمول زیر استفاده می‌شود.

$$P = P_{opt} \pm Z \sqrt{\frac{P_{opt}(1 - P_{opt})}{N}}$$

N برابر با مجموع افراد گروه قوی و ضعیف است.

سطح دشواری در سؤال‌های باز پاسخ

در سؤال‌های باز پاسخ از تقسیم میانگین نمره سؤال در دو گروه بر بارم سؤال سطح دشواری به دست می‌آید.

$$P_o = \frac{\frac{\sum S_u + \sum S_L}{n}}{b}$$

اصلاح سطح دشواری برای عامل حدس و گمان

یکی از محدودیت‌های آزمون‌های چندگزینه‌ای آن است که آزمودنی می‌تواند برای یافتن پاسخ درست از حدس و گمان استفاده کند. به همین منظور لازم است سطح دشواری سؤال تصحیح شود تأثیر این عامل کنترل و خنثی شود.

$$P_c = \frac{R - \frac{W}{K-1}}{N - HR}$$

R = تعداد پاسخ‌های درست

W = تعداد پاسخ‌های غلط

K = تعداد گزینه‌های سؤال

N = تعداد کل آزمودنی‌ها

HR = تعداد کسانی که به سؤال پاسخ نداده‌اند

مثال: اگر از یک گروه ۳۰۰ نفری، ۱۵۰ نفر به یک سؤال ۵ گزینه‌ای پاسخ درست، ۱۲۰ نفر پاسخ غلط و ۳۰ نفر باقیمانده سؤال را بی‌جواب گذاشته باشند، سطح دشواری اصلاح شده سؤال چقدر است؟

$$P_c = \frac{150 - \frac{120}{5-1}}{300 - 30} = 0.44$$

اصلاح نمره برای حدس زدن و پاسخ ندادن (مرتبط با مبحث نرم و هنجار)

یکی از روش‌های تبدیل نمره‌ها می‌تواند برای تعدیل تأثیر حدس و پاسخ ندادن به سؤال به کار برده شود. این تبدیل در تفسیر عملکرد هر یک از آزمودنی‌ها و مقایسه این عملکرد در بین آزمودنی‌های مختلف مفید است. تبدیلی که برای تعدیل اثر حدس یا پاسخ ندادن به سؤال به کار برده می‌شود، به طور سنتی نمره‌های فرمولی (formula score) نامیده می‌شود.

$$R - \frac{W}{K-1} = \text{نمره اصلاح شده}$$

مثال: فرض کنید در یک آزمون ۴۰ سؤالی ۵ گزینه‌ای فردی به ۱۶ سؤال پاسخ درست و به ۲۴ سؤال پاسخ غلط داده باشد. نمره اصلاح شده فرد برابر است با: (یعنی ده سؤال را بدون حدس پاسخ داده است)

$$10 = 16 - \frac{24}{5-1} = \text{نمره اصلاح شده}$$

گاهی آزمودنی‌ها به همه سؤال‌ها پاسخ نمی‌دهند. فرمولی که سؤال‌های بی‌پاسخ را نیز لحاظ می‌کند به شکل زیر است. B در این فرمول به معنای تعداد سؤال‌های بی‌پاسخ است.

$$R + \frac{B}{k}$$

در این فرمول حالتی فرض می‌شود که به جای سؤال بی‌پاسخ یک حدس انجام گیرد. برای مثال چنانچه در یک آزمون ۴ گزینه‌ای فردی به ۱۰ سؤال پاسخ صحیح و ۲ سؤال را بدون پاسخ بگذارد، نمره فرمولی وی عبارت خواهد بود از $10.5 = 10 + \frac{2}{4}$.

نکات:

در سؤالاتی که نمره گذاری آن صفر و یک است:

- میانگین نمره های خام برابر با $\sum P$ است.
- واریانس سؤال برابر با pq است (q نسبت افرادی که به سؤال پاسخ غلط داده‌اند برابر با $1-p$)
- حداکثر واریانس سؤال برای سطح دشواری ۰/۵ و برابر با ۰/۲۵ است.
- حداقل واریانس سؤال زمانی است که سطح دشواری سؤال صفر یا یک باشد.

ضریب تشخیص

ضریب تشخیص سؤال نشان می‌دهد که سؤال تا چه اندازه می‌تواند بین افراد گروه قوی و ضعیف تفکیک قائل شود. ضریب تشخیص برآوردی از همبستگی سؤال با نمره کل آزمون را به دست می‌دهد. برای برآورد ضریب تشخیص روش‌های متعددی وجود.

یکی از روش های برآورد ضریب تشخیص روش کلی است. در این روش مقدار D ، یعنی تفاوت سطح دشواری سؤال در افراد گروه قوی و ضعیف به دست می آید. بدین ترتیب که اوراق امتحانی به ترتیب از بیشترین به کمترین نمره مرتب می شود. ۲۷٪ برگه های بالا و ۲۷٪ برگه های پایین انتخاب و به عنوان افراد گروه قوی و ضعیف در نظر گرفته می شوند، سطح دشواری در گروه قوی و ضعیف محاسبه شده و از هم کم می شود، مقدار باقیمانده برابر با ضریب تشخیص سؤال است.

$$D = P_U - P_L$$

مثال: از افراد گروه قوی یک نمونه ۲۰۰ نفری ۴۰ نفر، و از افراد ضعیف ۳۵ نفر به یک سؤال پاسخ کلید داده اند، ضریب تشخیص سؤال چقدر است؟ آیا این سؤال را در مجموعه سؤال های آزمون نگه داریم؟

تعداد نفرات گروه قوی و ضعیف: $54 = 200 \times 27\%$

$$P_U = \frac{40}{54} = 0.74$$

$$P_L = \frac{35}{54} = 0.65$$

$$D = 0.74 - 0.65 = 0.09$$

(از آنجایی که ضریب تشخیص سؤال نزدیک به صفر است احتمالاً سؤال مناسبی نیست و باید حذف شود)

دامنه تغییر D بین +۱ تا -۱ است، و معناداری آن با آزمون χ^2 بررسی می شود.

$$\chi^2 = \frac{N \times D^2}{4pq}$$

یادآوری ضرایب همبستگی:

نوع مقیاس اندازه گیری	ضریب همبستگی
هر دو متغیر پیوسته (فاصله ای یا نسبی)	گشتاوری پیرسون
هر دو متغیر رتبه ای	اسپیرمن
هر دو متغیر اسمی دو ارزشی واقعی	فای
هر دو متغیر اسمی دو ارزشی ساختگی	تراکوریچ
یک متغیر پیوسته و متغیر دیگر دو ارزشی واقعی	دو رشته ای نقطه ای
یک متغیر پیوسته و متغیر دیگر دو ارزشی ساختگی	دو رشته ای

یکی دیگر از روش‌های برآورد ضریب تشخیص استفاده از جدول فلنگان است، که ضریب تشخیص را براساس گروه‌های انتهایی توزیع (۲۷٪ بالا و پایین) با کمترین خطا به دست می‌دهد.

ضریب همبستگی دورشته‌ای نقطه‌ای

یکی دیگر از تعاریف ضریب تشخیص یک سؤال، همبستگی سؤال با نمره کل آزمون است. بدین منظور همبستگی عملکرد آزمودنی در هر سؤال دو ارزشی با عملکرد وی در کل آزمون با استفاده از ضریب همبستگی دو رشته‌ای نقطه‌ای به دست می‌آید:

$$r_{pbis} = \frac{M_p - M_t}{S_t} \sqrt{\frac{p}{q}}$$

M_p = میانگین نمره کل افرادی که به سؤال پاسخ درست داده‌اند.

M_t = میانگین نمره کل افراد

S_t = انحراف استاندارد کل نمره‌ها

p = ضریب دشواری سؤال

یا

$$r_{pbis} = \frac{M_p - M_q}{S_t} \sqrt{pq}$$

مثال: میانگین و انحراف معیار نمره‌های خام یک آزمون ۷۰ سؤالی در یک گروه ۵۰ نفری به ترتیب برابر با ۵۴ و ۹ است، و میانگین ۲۵ نفر که به یک سؤال آزمون پاسخ درست داده‌اند برابر ۵۷ است. ضریب تشخیص سؤال را محاسبه کنید.

$$r_{pbis} = \frac{57 - 54}{9} \sqrt{\frac{\frac{25}{50}}{(1 - \frac{25}{50})}} = 0.33$$

ضریب همبستگی دو رشته‌ای

این ضریب نیز رابطه بین یک متغیر پیوسته با یک متغیر دو ارزشی را نشان می‌دهد. اما توزیع زیر بنایی متغیر دو ارزشی در جامعه نرمال است.

$$r_{bis} = \frac{M_p - M_t}{S_t} \times \frac{p}{Y}$$

در این فرمول Y عرض نقطه تفکیک معادل P در منحنی توزیع نرمال است.

رابطه ضریب همبستگی دو رشته‌ای نقطه‌ای و ضریب همبستگی دو رشته‌ای:

$$r_{pbis} \times \sqrt{pq} = r_{bis} \times Y$$

نکته: از ضرب ضریب همبستگی دو رشته‌ای نقطه در انحراف استاندارد سؤال ($r_{pbis} \times \sqrt{pq}$) شاخص اعتبار (پایایی) به دست می‌آید.

ضریب فای

یکی از کاربردهای ضریب فای، شاخصی برای بررسی همبستگی هر سؤال با سؤال دیگر است، رابطه بین دو متغیر دو ارزشی.

$$\phi = \frac{P_c - P_x P_y}{\sqrt{P_x(1 - P_x)P_y(1 - P_y)}}$$

P_c = نسبت افرادی که به هر دو سؤال پاسخ درست داده‌اند.

P_x = نسبت افرادی که به سؤال X پاسخ مثبت داده‌اند.

P_y = نسبت افرادی که به سؤال Y پاسخ مثبت داده‌اند.

ضریب تتراکوریک (چهارخانه‌ای)

شاخصی برای بررسی همبستگی دو متغیر دو ارزشی در شرایطی که توزیع زیربنایی هر دو متغیر در جامعه نرمال فرض شود.

نمونه سؤال کنگور دکتري روان شناسی تربیتی

۱۷- ضریب دشواری یک سؤال ۰/۲ و همبستگی دو رشته‌ای نقطه‌ای آن با نمره کل ۰/۳۲ است. اگر میانگین نمره کل افرادی که به سؤال پاسخ درست داده‌اند، ۲۸ و نمره کل افرادی که به سؤال پاسخ نادرست داده‌اند، ۲۰ باشد، انحراف استاندارد نمره کل آزمون کدام است؟ (سال ۱۴۰۲)

۵(۱) ۸(۲) ۱۰(۳) ۱۴(۴)

۱۸- سطح بهینه دشواری کدام سؤال چند گزینه‌ای ۰/۶ است؟ (سال ۱۴۰۱)

(۱) دو گزینه‌ای (۲) سه گزینه‌ای (۳) چهار گزینه‌ای (۴) پنج گزینه‌ای

۱۹- آزمونی را برای ۳۰ دانش آموز اجرا کرده‌ایم، پس از آن یک سوم برگه‌هایی که بالاترین نمره و یک سوم برگه‌هایی که پایین‌ترین نمره را کسب کرده بودند انتخاب و مشخص شد که ۹ نفر از گروه قوی و ۴ نفر از گروه ضعیف به سؤال A پاسخ درست داده‌اند، ضریب تمیز این سؤال برابر کدام است؟ (سال ۱۴۰۰)

۰/۲(۱) ۰/۳(۲) ۰/۴(۳) ۰/۵(۴)

۲۰- انحراف استاندارد یک سؤال دو ارزشی در یک نمونه ۴۰ نفری برابر با ۰/۴ است، ضریب دشواری آن کدام است؟ (سال ۱۳۹۹)

۰/۱(۱) ۰/۲(۲) ۰/۴(۳) ۰/۵(۴)

۲۱- اگر نمره اصلاح شده (با احتساب نمره منفی) فردی در یک آزمون ۴ گزینه‌ای برابر با ۶ باشد و تعداد پاسخ‌های نادرست او برابر با ۱۸ باشد، به چند سؤال آزمون پاسخ درست داده است؟ (سال ۱۳۹۷)

۶(۱) ۱۲(۲) ۱۸(۳) ۲۴(۴)

۲۲- در یک آزمون تشریحی ۱۰ سؤالی نمره هر سؤال ۲ است، در صورتی که میانگین نمره سؤال اول آن ۱/۶ باشد، ضریب دشواری آن کدام است؟ (سال ۱۳۹۶)

۰/۱۶(۱) ۰/۸۴(۲) ۰/۸(۳) ۰/۲۰(۴)

۲۳- اگر نمره های افراد در دو متغیر X و Y رتبه‌بندی و سپس همبستگی بین رتبه‌ها با استفاده از روش پیرسون محاسبه شود، کدام شاخص همبستگی به دست می‌آید؟ (سال ۱۳۹۶)

(۱) اسپیرمن (۲) چهارخانه‌ای (۳) تائو کندال (۴) دو رشته‌ای

۲۴- اصلی ترین هدف از تبدیل نمره‌های خام (row score) به نمره های فرمولی (formula score) کدام است؟ (سال ۱۳۹۶)

(۱) اعمال جریمه برای حدس تصادفی (۲) اعمال تشویق برای حدس آگاهانه

(۳) محاسبه نمره کل برای محاسبه رتبه (۴) محاسبه نمره کل برای مقایسه پذیری افراد

نظریه کلاسیک اندازه‌گیری

نظریه آزمون یا الگوی آزمون، نمایش نمادی عوامل مؤثر بر نمره‌های مشاهده شده است که با مفروضات آن توصیف می‌شود. به طور کلی نظریه‌های روان‌سنجی را می‌توان به دو دسته: نظریه‌های مبتنی بر نمونه‌گیری تصادفی (شامل نظریه کلاسیک اندازه‌گیری (CTT) و نظریه تعمیم پذیری (GT))، و نظریه‌های جدید اندازه‌گیری طبقه‌بندی کرد.

اسپیرمن در ۱۹۰۴ بیان کرد که هر آزمونی خطا می‌کند و تا سال ۱۹۷۹ روان‌سنجی بر روی همین نظریه (نظریه کلاسیک (یا ضعیف) نمره واقعی) استوار بود. نظریه کلاسیک اندازه‌گیری نمره مشاهده شده در آزمون را به عنوان نمره‌ای در نظر می‌گیرد که دارای دو مؤلفه یا دو جزء اصلی جمع‌پذیر، نمره «واقعی» و نمره «خطا» است. خطای تصادفی به عنوان عاملی ناهمبسته با نمره واقعی و ناهمبسته با خطای ناشی از اندازه دیگری از همان خصیصه تعریف می‌گردد.

مفروضه های بنیادی نظریه کلاسیک اندازه‌گیری

(۱) نمره مشاهده شده فرد شامل نمره واقعی و نمره خطاست.

$$X_0 = X_t + X_e$$

(۲) بین نمره خطا و نمره واقعی همبستگی وجود ندارد. به عبارت دیگر در جامعه اصلی خطا مستقل از نمره واقعی است.

$$r_{te} = 0$$

(۳) خطایی که در یک آزمون مرتکب می‌شویم ارتباط با خطای دیگر آزمون‌ها ندارد.

$$r_{eé} = 0$$

تعریف نمره واقعی در نظریه کلاسیک اندازه‌گیری:

اگر یک نفر را بی‌نهایت بار آزمون کنیم، یا بی‌نهایت نفر را تست کنیم، میانگین این بی‌نهایت بار برابر با نمره واقعی فرد است. و جمع خطاها در این بی‌نهایت بار برابر با صفر است. به عبارت دیگر مقدار مورد انتظار یا امید ریاضی برای نمرات برابر با نمره واقعی $\varepsilon(X)=T$ و امید ریاضی (میانگین جامعه) نمره‌های خطا برابر با صفر است $\varepsilon(E)=0$.
برخی نتایج استخراج شده از نظریه کلاسیک اندازه‌گیری:

(۱) واریانس نمره مشاهده شده برابر با واریانس نمره واقعی و واریانس خطاست.

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

(۲) مجذور همبستگی بین نمره‌های مشاهده شده و نمره واقعی برابر نسبت واریانس نمره‌های واقعی به نمره‌های مشاهده شده است، که نشان دهنده اعتبار (پایایی) آزمون است.

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_O^2}$$

(۳) همبستگی بین نمره‌های دو آزمون موازی، با نسبت واریانس نمره واقعی به واریانس نمره مشاهده شده برابر است، که مثل مورد دوم نشان دهنده ضریب اعتبار (پایایی) آزمون است.

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_O^2}$$

۴) مجذور همبستگی بین نمره‌های مشاهده شده و نمره خطا برابر نسبت واریانس نمره‌های خطا به نمره‌های مشاهده شده است.

$$\rho_{XE}^2 = \frac{\sigma_E^2}{\sigma_O^2}$$

۵) معادله‌های ضریب اعتبار (پایایی) براساس نظریه کلاسیک اندازه‌گیری

$$\rho_{XX} = \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_O^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2} = 1 - \rho_{XE}^2$$

۶) خطای استاندارد اندازه‌گیری برابر است با

$$\sigma_E = \sigma_O \sqrt{1 - \rho_{XX}}$$

۷) همبستگی نمره واقعی و نمره مشاهده شده برابر با جذر اعتبار (پایایی) است

$$\rho_{XT} = \sqrt{\rho_{XX}}$$

مدل‌های اندازه‌گیری مبتنی بر نظریه کلاسیک

مدل اندازه‌ها یا آزمون‌های موازی (Parallel model)

چنانچه برای هر یک از آزمودنی‌های جامعه نمره‌های واقعی دو آزمون برابر باشد ($T_1 = T_2$) و واریانس خطای دو آزمون ($\sigma_{E1}^2 = \sigma_{E2}^2$) نیز برابر باشد، این دو آزمون موازی نامیده می‌شود. واریانس و میانگین نمره‌های مشاهده شده آزمون‌های موازی مساوی‌اند. همبستگی هر یک از آنها با آزمون‌های دیگر برابر است. اعتبار (پایایی) دو آزمون نیز برابر است.

مدل اندازه‌ها یا آزمون‌های معادل τ (Tau equivalent)

آزمون‌های معادل تائو می‌توانند واریانس خطای نامساوی داشته باشند. اما نمره واقعی آنها با یکدیگر برابرند ($T_1 = T_2$). میانگین نمره مشاهده شده آنها برابر، اما واریانس نمره مشاهده شده نابرابر است. همبستگی هر یک از آنها با آزمون‌های دیگر، و اعتبار (پایایی) دو آزمون نیز نابرابر است.

مدل اندازه‌ها یا آزمون‌های اساساً معادل τ (essentially Tau equivalent)

آزمون‌های اساساً معادل تائو می‌توانند واریانس خطای نامساوی داشته باشند. نمره واقعی آنها با یک مقدار اضافی ثابت برابرند $(T_1 = T_2 + C)$. میانگین و واریانس نمره مشاهده شده آنها نابرابر است. همبستگی هر یک از آنها با آزمون‌های دیگر، و اعتبار (پایایی) دو آزمون نیز نابرابر است.

مدل اندازه‌های متجانس (congeneric model)

در این مدلها واریانس خطا و نمره واقعی دو آزمون نابرابر است. میانگین و واریانس نمره مشاهده شده آنها نابرابر است. همبستگی هر یک از آنها با آزمون‌های دیگر، و اعتبار (پایایی) دو آزمون نیز نابرابر است.

نمونه سؤال دکتري روان‌شناسي تربيتي

۲۵- همبستگی میان نمره واقعی و نمره مشاهده شده، کدام است؟ (۱۴۰۳)

$$\begin{matrix} r_{tt} & (1) & \sqrt{r_{tt}} & (2) & 1-r_{tt} & (3) & \sqrt{1-r_{tt}} & (4) \end{matrix}$$

۲۶- کدام مدل اندازه‌گیری منعکس کننده بیشترین شباهت در سؤال‌های یک آزمون است؟ (سال ۱۴۰۱)

Parallel (1) congeneric (۲)

Tau equivalent (۳) essentially Tau equivalent (۴)

۲۷- کدام مورد از عبارت $X_{pf}-T_p$ استنباط می‌شود؟ (p نماد فرد، f نماد فرم آزمون) (سال ۱۴۰۰)

(۱) خطای اندازه‌گیری تابع نمره واقعی نیست.

(۲) نمره مشاهده شده در فرم‌های مختلف تغییر می‌کند.

(۳) نمره واقعی در فرم‌های مختلف متفاوت است.

(۴) خطای نمره واقعی بیشتر از نمره مشاهده شده است.

۲۸- کدام عبارت بر اساس مدل کلاسیک $(O=T+E)$ درست نیست؟ (سال ۱۴۰۰)

$$\text{CoV}(O,E)=0 \quad (1) \quad \text{CoV}(T,E)=0 \quad (2)$$

$$E(T)=E(O) \quad (3) \quad \text{CoV}(T,O) \neq 0 \quad (4)$$

۲۹- مطابق با نظریه کلاسیک اندازه گیری، میانگین نظری اندازه های مستقل و مکرر یک آزمودنی با یک آزمون یکسان برابر با کدام مورد است؟ (سال ۱۳۹۹)

(۱) روایی (۲) پایایی (۳) نمره خطا (۴) نمره واقعی

۳۰- اگر در کل فرایند آزمون هیچ خطایی وجود نداشته باشد، نمره ای که آزمودنی کسب می کند برابر با کدام است؟ (سال ۱۳۹۷)

(۱) نمره واقعی (T) (۲) نمره خام (۳) نمره خطا (۴) نمره مشاهده شده (X)

۳۱- اگر در یک آزمون، انحراف معیار خطا $S_e = 2$ و واریانس مشاهده شده برابر با $S_o^2 = 8$ باشد. ضریب پایایی آزمون کدام است؟ (سال ۱۴۰۰)

(۱) ۰/۸۰ (۲) ۰/۷۵ (۳) ۰/۵ (۴) ۰/۲۵

ویژگی های روان سنجی آزمون

هر گاه بخواهیم از آزمون به منظور فراهم کردن اطلاعات مورد نیاز برای کمک به تصمیم درباره اشخاص استفاده کنیم، باید برخی ویژگی ها را مورد توجه قرار دهیم شامل: عملی بودن (Practicality)، پایایی (اعتبار) (Reliability) و روایی (Validity).

عملی بودن

هر گاه بخواهیم آزمونی را به کار ببریم باید برخی ملاحظات عملی را در نظر بگیریم از جمله این موارد نقش عوامل اقتصادی، سهولت اجرا، و سهولت تفسیر و استفاده صحیح از نمره هاست.



اعتبار (پایایی) (Reliability)

برای تعریف و تفسیر اعتبار (پایایی) راه‌های متعددی وجود دارد. برای مثال یک آزمون زمانی دارای اعتبار (پایایی) است که نمره‌های مشاهده شده و نمره‌های واقعی آن دارای همبستگی بالایی باشند. مجذور همبستگی بین این نمره‌ها (ρ_{XT}^2) ضریب اعتبار (پایایی) آزمون نامیده می‌شود. یا در آزمون‌های موازی می‌توان اعتبار (پایایی) را به عنوان ضریب همبستگی بین نمره‌های مشاهده شده بیان کرد. چنانچه برای جامعه‌ای از آزمودنی‌ها، دو آزمون موازی اجرا شود و نمره‌های مشاهده شده به دست آمده همبسته باشند به این همبستگی ضریب اعتبار (پایایی) گفته می‌شود (ρ_{XX}). در تعریف سنتی، پایایی آزمون یعنی اگر خصیصه مورد نظر را با همان وسیله (یا وسیله مشابه و قابل مقایسه با آن) در شرایط مشابه دوبار اندازه بگیریم، نتایج حاصل مشابه، دقیق، پیش‌بینی‌پذیر و قابل اعتماد باشد. اعتبار (پایایی) آزمون به دقت اندازه‌گیری و ثبات آن مربوط است.

اعتبار (پایایی) آزمون را می‌توان به عنوان خارج قسمت واریانس نمره واقعی بر واریانس نمره مشاهده شده تعریف کرد.

$$r_{11} = \frac{\sigma_T^2}{\sigma_O^2}$$

برای آزمونی که کاملاً معتبرست، واریانس خطا برابر صفر است و تمام واریانس نمره‌های مشاهده شده مربوط به نمره واقعی است و اندازه‌گیری بدون خطا صورت گرفته است. هر چقدر اعتبار (پایایی) آزمون افزایش یابد واریانس نمره خطا به همان نسبت کوچکتر می‌شود.

نکته: از آنجایی که مجذور همبستگی نمره مشاهده شده و نمره واقعی بیانگر ضریب اعتبار (پایایی) است، در نتیجه همبستگی بین نمره مشاهده شده و نمره واقعی برابر با جذر ضریب اعتبار (پایایی) است $\rho_{XT} = \sqrt{\rho_{XX}}$. در تعریف ضریب روایی داریم که همبستگی آزمون با یک ملاک بیرونی است، از آنجایی همبستگی یک آزمون با خودش همواره بزرگتر از همبستگی آزمون با هر متغیر دیگری است، در نتیجه حد بالای روایی برابر با جذر اعتبار (پایایی) $\sqrt{\rho_{XX}}$ است. ضریب روایی نمی‌تواند از جذر ضریب اعتبار (پایایی) بیشتر باشد، اما می‌تواند بزرگتر از خود ضریب اعتبار (پایایی) باشد.

اگر ضریب اعتبار (پایایی) برابر یک ($\rho_{XX} = 1$) باشد نتایج زیر حاصل می‌شود:

- (۱) اندازه‌گیری بدون خطا انجام شده ($E=0$)
- (۲) برای همه آزمودنی‌ها $X=T$ است.
- (۳) کل واریانس نمره مشاهده شده منعکس کننده واریانس نمره واقعی است ($\sigma_O^2 = \sigma_T^2$).
- (۴) تمام اختلاف بین نمره‌های مشاهده شده منعکس کننده اختلاف‌های نمره‌های واقعی است.
- (۵) همبستگی نمره‌های مشاهده شده و نمره‌های واقعی مساوی یک است ($\rho_{XT} = 1$)
- (۶) همبستگی نمره‌های مشاهده شده و نمره‌های خطا برابر با صفر است ($\rho_{XE} = 0$)

اگر ضریب اعتبار (پایایی) برابر صفر ($\rho_{XX} = 0$) باشد نتایج زیر حاصل می‌شود:

- (۱) نمره آزمون تنها خطای اندازه‌گیری را منعکس می‌کند.
- (۲) برای همه آزمودنی‌ها $X=E$ است.
- (۳) کل واریانس نمره مشاهده شده منعکس کننده واریانس خطاست ($\sigma_O^2 = \sigma_E^2$).
- (۴) تمام اختلاف بین نمره‌های مشاهده شده نشان دهنده خطای اندازه‌گیری است.
- (۵) همبستگی نمره‌های مشاهده شده و نمره‌های واقعی مساوی صفر است ($\rho_{XT} = 0$).
- (۶) همبستگی نمره‌های مشاهده شده و نمره‌های خطا برابر با یک است ($\rho_{XE} = 1$).

از آنجایی که نمره‌های واقعی را در بیشتر موارد نمی‌توان به دست آورد و تأیید موازی بودن آزمون‌ها امکان پذیر نیست، بنابراین از روش‌های دیگری برای برآورد اعتبار (پایایی) استفاده می‌شود. سه روش عمده برآورد اعتبار (پایایی) عبارت است از: بازآزمایی، استفاده از فرم‌های هم‌تا و ثبات (همسانی) درونی.

بازآزمایی

در این روش آزمون واحد را درباره گروه نمونه واحدی از آزمودنی‌ها در دو زمان متفاوت اجرا و ضریب همبستگی بین نمره‌های حاصل از دوبار اجرای آزمون محاسبه می‌شود. ضریب همبستگی که از این روش به دست می‌آید حالت خاصی از ضریب اعتبار (پایایی) است که در اصلاح ضریب ثبات یا پایایی گفته می‌شود. بازآزمایی می‌تواند با فاصله زمانی یا بدون فاصله زمانی انجام شود.

اشکالات روش بازآزمایی:

(۱) اثر انتقال: آشنایی آزمودنی‌ها با سؤال آزمون در آزمایش اول بر آزمایش دوم اثر می‌گذارد. همچنین تغییر در نگرش‌ها یا میزان اطلاعات آزمودنی‌ها می‌تواند علت دیگر برای انتقال باشد.

(۲) مدت زمان بین دو اجرا نیز مشکل ساز است، اگر فاصله زمانی کوتاه باشد به دلیل وجود حافظه، **تمرین** موجب اثر انتقال می‌شود. فاصله زمانی زیاد احتمال تأثیرات ناشی از تغییر اطلاعات یا تغییرات خلقی را بیشتر می‌کند.

در نتیجه روش بازآزمایی برای آن دسته ویژگی‌هایی مناسب است که نسبت به اثر انتقال حساس نیستند و در فاصله زمانی ثابت‌اند، مثل آزمون‌های ادراک بصری یا شنوایی.

آزمون‌های همتا

در بعضی از موارد آزمون دارای دو فرم همتا است، در این روش دو فرم را یکی پس از دیگری یا با فاصله زمانی اجرا کرده و همبستگی نمره‌های حاصل از دو آزمون محاسبه می‌شود. فرم‌های همتا، دو فرم هر آزمونی است که در ساخته شدن آنها سعی می‌شود آنها را موازی کنند و ممکن است میانگین و واریانس‌های نمره‌های مشاهده شده و همچنین همبستگی هریک از آنها با اندازه‌های دیگر مساوی (یا بسیار شبیه) باشد. استفاده از فرم همتا تا حدی اثر انتقال ناشی از بازآزمایی را از بین می‌برد.

برآورد اعتبار (پایایی) ثبات (همسانی) درونی

در برخی موارد امکان اجرای دوبار آزمون وجود ندارد. در این حالت می‌توان برای برآورد اعتبار (پایایی) از روش‌های همسانی درونی که مستلزم یک بار اجرای آزمون است استفاده کرد، این روش‌ها شامل دو نیمه کردن، و همسانی در حالت کلی است.

• روش دو نیمه کردن:

در این روش ابتدا آزمون در مورد گروه نمونه‌ای از آزمودنی‌ها اجرا می‌شود، سپس آزمون به دو قسمت که هر کدام فرم‌های موازی یکدیگرند تقسیم می‌شوند. هر نیمه آن جداگانه نمره گذاری می‌شود، ضریب همبستگی دو نیمه آزمون شاخصی از اعتبار (پایایی) هر یک از دو نیمه است. از آنجایی که با افزایش تعداد سؤال‌های آزمون ضریب

اعتبار (پایایی) آن نیز افزایش می‌یابد و ضریب همبستگی دو نیمه آزمون از اعتبار (پایایی) کل آزمون کمتر است. برای برآورد اعتبار (پایایی) کل آزمون از فرمول اسپیرمن - برآون استفاده می‌شود.

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

مثال: اگر ضریب همبستگی دو نیمه آزمون ۰/۶ باشد، ضریب اعتبار (پایایی) کل آزمون را برآورد کنید.

$$r_{11} = \frac{2 \times 0.6}{1 + 0.6} = 0.75$$

نکات:

✓ چنانچه نیمه‌های آزمون موازی باشند، اعتبار (پایایی) کل آزمون از فرمول اسپیرمن - براون برآورد می‌شود. در صورتی که نیمه‌های آزمون اساساً معادل تائو باشد، ضریب آلفا برای برآورد ضریب اعتبار (پایایی) به کاربرده می‌شود. در صورتی که دو نیمه حداقل اساساً معادل تائو نباشد (یعنی متجانس باشد) ضریب آلفا، اعتبار (پایایی) را کمتر از اندازه واقعی آن برآورد می‌کند.

فرمول آلفا برای دو نیمه آزمون درحالتی که واریانس دو نیمه یکسان نیست.

$$\alpha = \frac{2(S_t^2 - (S_X^2 + S_Y^2))}{S_t^2}$$

✓ روش دونیمه کردن مستلزم یکسانی واریانس دونیمه آزمون است، یک روش جایگزین برای زمانی که شرط

یکسانی واریانس دو نیمه برقرار نیست روش رولون است. فرمول رولون: $1 - \frac{S_d^2}{S_t^2}$

✓ برای آزمون‌های سرعت نباید برآورد اعتبار (پایایی) دو نیمه‌ای محاسبه شود.

سه روش برای دونیمه کردن آزمون به کار برده می‌شود:

روش یکم: روش زوج و فرد است، در این روش سؤال‌ها براساس فرد یا زوج بودن آنها به دو دسته تقسیم می‌شوند.

در روش دوم: نیمه‌ها براساس ترتیب شکل می‌گیرد، نیمه اول و نیمه دوم آزمون.

روش سوم: معروف به «زیرمجموعه‌های جور شده تصادفی است»، سؤال‌ها براساس درجه دشواری و ضریب تشخیص به دو دسته تقسیم می‌شوند.

به طور کلی تعداد دو نیمه‌های ممکن برابر است با

$$\frac{(n_1 + n_2)!}{n_1! \times n_2!} \times \frac{1}{2}$$

مثال در یک آزمون ۶ سؤالی به چند طریق می‌توان آزمون را به دو بخش تقسیم کرد؟

$$\frac{(6)!}{3! \times 3!} \times \frac{1}{2} = 10$$

نمونه سؤال گنگور دگتری روان‌شناسی تربیتی

۳۲- کدام مورد، جزو شرایط و پیش نیازهای آلفای کرونباخ نیست؟ (سال ۱۴۰۲)

(۱) تک بعدی بودن (۲) مستقل بودن خطا

(۳) اساساً تائو معادل بودن نشانگرها (۴) موازی بودن سؤال‌ها

۳۳- در یک آزمون ۴ سؤالی به چند طریق می‌توان سؤال‌ها را به دو نیمه تقسیم و همبستگی دو نیمه را محاسبه کرد؟ (سال ۱۴۰۰)

(۱) ۲ (۲) ۳ (۳) ۶ (۴) ۸

۳۴- پایایی یک آزمون ۲۰ سؤالی با استفاده از روش دو نیمه کردن برابر با ۰/۷۵ است. همبستگی دو نیمه آزمون کدام است؟ (سال ۱۳۹۹)

(۱) ۰/۳۷ (۲) ۰/۴۸ (۳) ۰/۵ (۴) ۰/۶

• ثبات (همسانی) درونی در حالت کلی

اشکال روش دو نیمه کردن در این است که انتخاب بخش بندی آزمون روی اعتبار (پایایی) اثر می گذارد. روش دو نیمه کردن را می توان به تقسیم آزمون به بیش از دو جزء تعمیم داد. با فرض اینکه آزمون را به π جزء تقسیم کنیم، از واریانس نمره های این π جزء (تعداد سؤالات) و واریانس نمره کل برای برآورد ضریب اعتبار (پایایی) استفاده می شود.

(۱) روش کودر ریچاردسون

با روش های کودر ریچاردسون می توان متوسط ضریب اعتبار (پایایی) را، که براساس ضرایب همبستگی تمام دو نیمه های ممکن آزمون به دست می آید برآورد کرد. دو فرمول KR20 و KR21 استفاده می شود. زمانی که نمره گذاری آزمون به صورت دو ارزشی صفر و یک باشد از فرمول کودر ریچاردسون استفاده می شود.

$$KR20 = \frac{k}{k-1} \left(\frac{S^2 - \sum pq}{S^2} \right)$$

k = تعداد سؤال ها

p = نسبت آزمودنی هایی که به سؤال پاسخ درست داده اند.

q = نسبت آزمودنی هایی که به سؤال پاسخ غلط داده اند ($1-p$)

pq = واریانس هر سؤال

S^2 = واریانس نمره کل آزمون

زمانی که نمره هر سؤال به صورت صفر و یک باشد، و سطح دشواری تمام سؤال ها یکسان باشد، از فرمول کودر ریچاردسون ۲۱ می توان استفاده کرد.

$$KR21 = \frac{k}{k-1} \left[1 - \frac{\bar{X} - \left(1 - \frac{\bar{X}}{k}\right)}{S^2} \right]$$

۲) ضریب آلفای کرونباخ

معمولاً در پاره‌ای از آزمون‌ها مانند آزمون‌های شخصیت و یا مقیاس‌های نگرش سنج برای سؤال‌های آزمون پاسخ درست و غلط وجود ندارد. مثلاً در یک آزمون نگرش سنج از آزمودنی می‌خواهند درجه موافقت یا مخالفت خود را نسبت به یک عقیده اعلام کنند. در این موارد از ضریب آلفای کرونباخ برای برآورد همسانی درونی (اعتبار، پایایی) آزمون استفاده می‌شود.

$$\alpha = \frac{k}{k-1} \left(\frac{S_t^2 - \sum S_i^2}{S_t^2} \right)$$

k = تعداد سؤال‌ها

S_i^2 = واریانس هر سؤال

S_t^2 = واریانس نمره کل آزمون

نکته: در فرمول بالا $(S_t^2 - \sum S_i^2)$ برابر با دو برابر مجموع کوواریانس‌های سؤال‌های آزمون است.

(واریانس نمره مرکب: واریانس نمره مرکب برابر با مجموع واریانس تک تک متغیرها، به علاوه دو برابر مجموع کوواریانس دو به دو بین متغیرهاست. یعنی در اینجا واریانس کل آزمون از مجموع واریانس سؤال‌ها بیشتر است و کوواریانس‌ها را نیز در برمی‌گیرد)

نکته: مقادیر ثبات یا همسانی درونی زمانی بالاست که همبستگی درونی اجزاء آزمون بالا باشد. یعنی اگر همبستگی بین سؤال‌ها (به همان نسبت کوواریانس بین سؤال‌ها) برابر با صفر باشد، ضریب اعتبار (پایایی) نیز صفر خواهد بود. این حالت واریانس کل آزمون با مجموع واریانس سؤال‌ها برابر خواهد بود.

سایر فرمول‌های برآورد ضریب آلفا:

برآورد آلفای کرونباخ براساس ماتریس واریانس و کوواریانس

$$\alpha = \frac{K\bar{C}}{\bar{V} + (K-1)\bar{C}}$$

فرمول آلفای استاندارد شده (براساس میانگین ضرایب همبستگی)

$$\alpha = \frac{K\bar{r}}{1 + (K - 1)\bar{r}}$$

فرمول آلفای طبقه‌ای

زمانی که بتوان سؤال‌ها را براساس حوزه محتوایی یا خرده آزمون‌ها گروه بندی کرد (آزمون ساختار چندبعدی داشته باشد). در این حالت ابتدا آلفا را برای هر خرده آزمون برآورد و سپس با فرمول زیر ترکیب می‌کنیم.

$$\text{آلفای طبقه‌ای} = 1 - \frac{\sum s_i^2(1-\alpha_i)}{s_t^2}$$

اعتبار (پایایی) آزمون‌های باز پاسخ

ضریب کاپا

برای آزمون‌هایی که نمره گذاری یا درجه بندی آزمودنی‌ها تا حدی ذهنی است، مثل آزمون‌های انشایی یا آزمون‌هایی نظیر گودیناف و بندرگشتالت برای برآورد اعتبار (پایایی) می‌توان بین نمره‌های چند ارزیاب (داور) همبستگی به دست آورد. ضریب کاپا به عنوان شاخص توافق بین داوران به کار می‌رود.

$$K = \frac{P_o - P_e}{1 - P_e}$$

P_o = نسبت مواردی که بین ارزیاب‌ها توافق وجود دارد.

P_e = نسبت توافق مورد انتظار

خطای معیار (استاندارد) اندازه‌گیری

خطای معیار اندازه‌گیری شاخصی است که، متوسط اندازه خطا در اندازه‌گیری با یک آزمون را نشان می‌دهد. هر چقدر خطای اندازه‌گیری آزمون بیشتر باشد، اطمینان ما نسبت به دقت خصیصه مورد اندازه‌گیری کمتر است. بین خطای معیار اندازه‌گیری و ضریب اعتبار (پایایی) آزمون رابطه معکوس وجود دارد.

$$\sigma_E = \sigma_O \sqrt{1 - \rho_{XX}}$$

مثال: انحراف معیار و ضریب اعتبار (پایایی) یک آزمون هوش به ترتیب ۱۵ و ۰/۹۱ است. اگر نمره مشاهده شده آزمودنی در این آزمون برابر با ۱۰۸ باشد، نمره واقعی او را در سطح اطمینان ۹۵ درصد برآورد کنید.

$$\sigma_E = 15\sqrt{1 - 0.91} = 4.5$$

$$108 \pm (1.96 \times 4.5) \cong (99 \text{ و } 117)$$

نمره واقعی فرد با ۹۵ درصد اطمینان حدوداً بین ۹۹ و ۱۱۷ قرار دارد.

نمونه سؤال کنگور دکتري روان‌شناسی تربیتی

۳۵- در یک آزمون ۶ سؤالی، مقدار آلفای استاندارد شده برابر با ۰/۸ است. میانگین همبستگی درونی بین سؤالات آزمون کدام است؟ (سال ۱۴۰۲)

- (۱) ۰/۲ (۲) ۰/۴ (۳) ۰/۶ (۴) ۰/۸

۳۶- اگر تعداد سؤال‌های یک آزمون برابر $k=10$ ، واریانس نمره خام برابر $\sigma_Y^2=30$ و میانگین کوواریانس سؤال‌ها $\sigma_{ij} = 0.2$ باشد، میزان آلفا برابر چند است (سال ۱۴۰۱)

- (۱) ۰/۵ (۲) ۰/۳۳ (۳) ۰/۴ (۴) ۰/۶۷

۳۷- براساس ماتریس واریانس - کوواریانس زیر، پایایی آزمون کدام است؟ (سال ۱۴۰۱)

$$\begin{bmatrix} Cov_{11} = 1 & & \\ Cov_{21} = 0.5 & Cov_{22} = 2 & \\ Cov_{31} = 1 & Cov_{32} = 0.5 & Cov_{33} = 0.5 \end{bmatrix}$$

- (۱) ۰/۵ (۲) ۰/۶ (۳) ۰/۷ (۴) ۰/۸

روش‌های افزایش اعتبار (پایایی) آزمون

(۱) افزایش تعداد سؤال‌ها

(۲) حذف سؤال‌های ناهمساز (سؤال‌هایی که با کل آزمون همبستگی پایینی دارند، یا سؤال‌های خیلی دشوار یا خیلی آسان)

(۳) گروه‌های ناهمگون (ضریب اعتبار (پایایی) گروه‌های ناهمگون بیش از گروه‌های همگون است)

افزایش تعداد سؤال‌های آزمون

با استفاده از فرمول زیر می‌توان اثر افزایش سؤال در اعتبار (پایایی) آزمون را به دست آورد، این فرمول شکل کلی فرمول اسپیرمن - برآون است.

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}$$

نکته: استفاده از این فرمول بر این فرض استوار است که اجزای اضافه شده باید موازی با سؤال‌های موجود در آزمون باشند.

مثال: فرض کنید آزمونی ۲۰ سؤال دارد، و ضریب اعتبار (پایایی) آن ۰/۷ است، اگر ۴۰ سؤال به آن اضافه کنیم ضریب اعتبار (پایایی) جدید را برآورد کنید:

(با اضافه کردن ۴۰ سؤال تعداد سؤال‌های آزمون به ۶۰ افزایش می‌یابد، به عبارت دیگر آزمون ۳ برابر شده، پس ضریب افزایش در فرمول برابر با ۳ خواهد بود)

$$r_{nn} = \frac{3 \times 0.7}{1 + (3 - 1)0.7} = 0.875$$

برخی مواقع پژوهشگر به دنبال آن است که بداند با چه مقدار افزایش می‌توان اعتبار (پایایی) آزمون را به سطح مطلوب (r_d) رساند:

$$n = \frac{r_d(1 - r_o)}{r_o(1 - r_d)}$$

مثال: آزمونی ۲۰ سؤال دارد و اعتبار (پایایی) آن برابر با ۰/۷۶ است. می‌خواهیم اعتبار (پایایی) آزمون را به ۰/۸۵ افزایش دهیم، تعداد سؤال باید چند برابر شود؟

$$n = \frac{0.85(1 - 0.76)}{0.76(1 - 0.85)} = 1.79$$

- یعنی برای رسیدن به اعتبار (پایایی) مطلوب، تعداد سؤال‌ها باید تقریباً دو برابر شود و با ۴۰ سؤال می‌توان به اعتبار (پایایی) مطلوب دست یافت، به عبارت دیگر در این مثال باید ۲۰ سؤال به سؤال‌های آزمون اضافه شود.
اعتبار (پایایی) نمره‌های تفاوت

گاهی رابطه بین دو نمره برای ما اهمیت دارد. یعنی تفاوت دو نمره تا چه اندازه مهم است. مثلاً تفاوت نمره هوش کلامی با هوش غیر کلامی آزمودنی در آزمون و کسلر تا چه اندازه معتبر است.

$$r_{dif} = \frac{\frac{(r_{xx} + r_{yy})}{2} - r_{xy}}{1 - r_{xy}}$$

مثال: اگر اعتبار (پایایی) بخش کلامی آزمون هوش و کسلر برابر با ۰/۹۱ و اعتبار (پایایی) بخش غیر کلامی آن برابر با ۰/۸۴ باشد. با فرض اینکه همبستگی بین بخش کلامی و غیر کلامی ۰/۸ است، اعتبار (پایایی) تفاوت دو نمره چقدر است؟

$$r_{dif} = \frac{\frac{(0.91 + 0.84)}{2} - 0.8}{1 - 0.8} = 0.375$$

نکته: اعتبار (پایایی) تفاوت دو نمره همواره کمتر از اعتبار (پایایی) دو تست است.

خطای استاندارد تفاوت:

$$S_{Edif} = \sqrt{S_{Ex}^2 + S_{Ey}^2}$$

در مثال قبل خطای استاندارد تفاوت دو نمره به شرح زیر است:

$$S_{E_x} = 15\sqrt{1 - 0.91} = 4.5$$

$$S_{E_y} = 15\sqrt{1 - 0.84} = 6$$

$$S_{E_{dif}} = \sqrt{4.5^2 + 6^2} = 7.5$$

نکته: خطای استاندارد تفاوت همواره بزرگتر از خطای استاندارد دو تست است.

نمونه سؤال کنگور دکتري روان شناسي تربيتي

۳۸- پایایی آزمون X به عنوان پیش آزمون، ۰/۷ و پایایی آزمون Y به عنوان پس آزمون، ۰/۹ است. اگر همبستگی بین نمره‌های دو آزمون ۰/۵ باشد، پایایی نمره‌های اختلافی کدام است؟ (سال ۱۴۰۲)

۰/۴(۱) ۰/۵(۲) ۰/۶(۳) ۰/۸(۴)

۳۹- ضریب پایایی یک آزمون ۳۰ سؤالی برابر با ۰/۶ است. اگر ۱۵ سؤال به آزمون اضافه شود، ضریب پایایی آزمون جدید چقدر است؟ (سال ۱۴۰۰)

۰/۶(۱) ۰/۶۹(۲) ۰/۶۷(۳) ۰/۷۵(۴)

نظریه تعمیم پذیری (GT)

کرانباخ و همکارانش مفهوم تعمیم پذیری را به عنوان تلاشی برای افزایش دقت تفسیر آزمون، ابداع کرده‌اند. نظریه تعمیم پذیری تلاشی برای برآورد اعتبار (پایایی) آزمون از طریق تحلیل واریانس است. خطا در نظریه کلاسیک (CTT) یک مفهوم یکپارچه و کلی است و اجزای آن مشخص نیست و روش کنترل خطا در نظریه کلاسیک انتخاب تصادفی است. در نظریه تعمیم پذیری می‌توان سهم منابع خطا را شناسایی و مؤلفه‌های خطا را به حداقل برسانیم و در نتیجه اعتبار (پایایی) آزمون را بهبود بخشیم. عواملی مثل نوع ارزیاب، روش و ... می‌تواند به عنوان خطا در نظر گرفته شود یا جزئی از پژوهش باشد.

مطالعات تعمیم‌پذیری به عنوان راهبردی برای شناسایی منابع احتمالی خطا در نمرات استفاده می‌شوند. تحلیل تعمیم‌پذیری بر مبنای اصول تحلیل واریانس با استفاده از جنبه‌های مختلف آزمون به عنوان عوامل تحلیل انجام می‌شود. مطالعات تعمیم‌پذیری می‌توانند برآوردی از واریانس ناشی از عملکرد ارزیاب‌ها، تکالیف، روش‌ها و سایر جنبه‌های سنجش ارائه دهند. در طول فرایند ساخت آزمون، متخصصان آزمون‌سازی می‌توانند از مطالعات تعمیم‌پذیری برای تعیین تعداد بهینه سؤال‌ها، ارزیاب‌ها و روش‌های لازم برای به حداقل رساندن خطاها با توجه به مقدار زمان موجود در آزمون استفاده کنند. در نظریه تعمیم‌پذیری می‌خواهیم بینم آیا نمره آزمون در این مجموعه سؤال، مکان و زمان و ارزیاب مشخص، قابل تعمیم به هر مکان، زمان و ارزیاب ممکن دیگر هست.

ضریب تعمیم‌پذیری (اعتبار، پایایی) نسبتی از واریانس نمره جهان به نمره مشاهده شده است.

$$\text{ضریب تعمیم‌پذیری از طریق میانگین مجزورات} = \frac{MS_p - MS_{PI}}{MS_p}$$

MS_p = میانگین مجزورات شخص

MS_{PI} = میانگین مجزورات تعامل شخص و سؤال یا میانگین مجزورات پسماند (باقیمانده)

$$\text{ضریب تعمیم‌پذیری از طریق مؤلفه‌های واریانس} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n_i}}$$

نمونه سؤال گنگور، دکتري روان‌شناسی تربیتی

۴۰- کدام رویکرد مطالعه اعتبار، قابلیت بررسی منابع چندگانه خطا را دارد؟ (۱۴۰۳)

(۱) تعمیم‌پذیری (۲) کلاسیک (۳) پرسش - پاسخ (۴) داده بنیاد

۴۱- در صورتی که در یک آزمون ۱۰ سؤالی میانگین مجزورات ناشی از آزمودنی‌ها برابر با $MS_p = ۱۲۰$ و میانگین مجزورات ناشی از سؤال‌ها $MS_I = ۷۰$ و میانگین مجزورات ناشی از تعامل آزمودنی و سؤال $MS_{PI} = ۳۰$ باشد، برآورد اعتبار (پایایی) آزمون از طریق تحلیل واریانس کدام است؟ (سال ۱۴۰۱)

(۱) ۰/۸۳ (۲) ۰/۸ (۳) ۰/۷۵ (۴) ۰/۴۵

۴۲- اگر $MS_{person}=24$ و $MS_{residual}=0/8$ باشد، پایایی داوران برابر چند است؟ (سال ۱۴۰۲)

۰/۹۷ (۴)

۰/۷۵ (۳)

۰/۲۷ (۲)

۰/۰۳ (۱)

روایی (Validity)

در تعریف سنتی، روایی به معنای هماهنگی بین نمره‌های آزمون با صفت یا خصیصه‌ای است که آزمون برای اندازه‌گیری آن ساخته شده است. در تعریف دیگر مقصود از روایی مناسب بودن، با معنا بودن و مفید بودن استنباط‌های خاصی است که از روی نمره آزمون به عمل می‌آید. رواسازی آزمون به معنای گردآوری شواهدی برای تفسیر و استنباط درباره نمره آزمون است. گردآوری شواهد در سه دسته امکان پذیر است: شواهد وابسته به محتوا، شواهد وابسته به ملاک و شواهد وابسته به سازه.

شواهد وابسته به محتوا

روایی محتوایی به تحلیل منطقی محتوایی یک آزمون بستگی دارد و تعیین آن براساس قضاوت ذهنی و فردی است. دو نوع شواهد مربوط به محتوا وجود دارد: روایی صوری و روایی منطقی (محتوایی)

• روایی صوری یا ظاهری آزمون (روایی ذهنی) (face validity)

روایی صوری نوعی از شواهد وابسته به محتوا است و ویژگی آزمون محسوب می‌شود. اینکه آزمون تا چه حد در ظاهر شبیه موضوعی است که برای اندازه‌گیری آن تهیه شده، درجه‌ای از منطقی بودن یک آزمون به نظر کسانی که به آن پاسخ می‌دهند را نشان می‌دهد. روایی صوری از لحاظ قابلیت پذیرش و معقول بودن آزمون برای آزمون شوندگان دارای اهمیت است. هنگامی این روایی برقرار می‌شود که فردی آزمونی را بررسی بکند و نتیجه بگیرد که این آزمون صفت مورد نظر را اندازه‌گیری می‌کند. فردی که این بررسی را انجام می‌دهد، می‌تواند یک آزمودنی یا یک متخصص باشد.

نکته: هر چند روایی صوری ویژگی مطلوب آزمون است، با این حال در برخی آزمون‌ها این ویژگی نه تنها ضروری نیست، بلکه بهتر است وجود نداشته باشد، مثل آزمون‌های تشخیص بیماری روانی که وجود این نوع روایی می‌تواند سبب شود افراد به شکل عمدی خود را مطلوب‌تر از آنچه هستند جلوه دهند.

• **روایی محتوایی، روایی منطقی (Content Validity)**

روایی محتوایی به این مطلب اشاره دارد که نمونه سؤال‌های مورد استفاده در آزمون تا چه حد، معرف کل جامعه سؤال‌های ممکن است، که می‌توان از موضوع و محتوای مورد نظر تهیه کرد. ارزشیابی رابطه بین اجزای آزمون و مجموعه کلی اغلب مبتنی بر داوری تخصصی و تجارب حرفه‌ای است.

نکته: ارزیابی روایی محتوایی به عهده متخصصان موضوعی و صاحب‌نظران است، اما در روایی صوری بیشتر بر اساس قضاوت شرکت‌کنندگان در آزمون و در کنار آن از نظر متخصصان نیز استفاده می‌شود.

برای ارزیابی روایی محتوایی علاوه بر نظر متخصصان موضوعی و صاحب‌نظران دو روش کمی نیز وجود دارد:

(۱) نسبت روایی محتوایی (CVR) (ضریب لاوشه)

این نسبت توسط لاوشه طراحی شده است. برای محاسبه آن از نظر کارشناسان متخصص در زمینه محتوای مورد نظر استفاده می‌شود. بدین صورت که هر متخصص سؤال‌ها را براساس یک طیف سه درجه‌ای (ضروری است، مفید اما ضروری نیست، ضروری نیست) طبقه‌بندی می‌کنند.

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

n_e = تعداد متخصصینی است که گزینه ضروری را انتخاب کرده‌اند.

N = تعداد کل متخصصان است.

(۲) شاخص روایی محتوایی (CVI)

شاخص روایی محتوایی نیز برای سنجش روایی محتوایی استفاده می‌شود، توسط والتز و باسل تهیه شده است. در این روش از متخصصان خواسته می‌شود میزان مرتبط بودن هر گویه را در طیف چهاردرجه‌ای (غیرمرتبط، نیاز به بازبینی اساسی، مرتبط اما نیاز به بازبینی، کاملاً مرتبط) درجه‌بندی کنند.

در این حالت تعداد متخصصینی که گزینه ۳ و ۴ (مرتبط اما نیاز به بازبینی، کاملاً مرتبط) را انتخاب کرده‌اند بر تعداد کل متخصصین تقسیم می‌شود، اگر این مقدار کمتر از ۰/۷ باشد، گویه مناسب نیست، بین ۰/۷ تا ۰/۷۹ گویه نیاز به بازبینی دارد، و بیشتر از آن گویه مناسب است.

نمونه سؤال کنگور دکتری روان‌شناسی تربیتی

۴۳- کدام مورد به ترتیب در خصوص روایی محتوایی و روایی صوری، درست است؟ (سال ۱۴۰۲)

(۱) داوری متخصصان- داوری آزمون شوندگان (۲) داوری آزمون شوندگان- داوری متخصصان

(۳) داوری آزمون شوندگان- داوری آزمون شوندگان (۴) داوری متخصصان- داوری متخصصان

۴۴- برای بررسی روایی محتوایی سؤالات یک آزمون استعداد، از ۱۰ متخصص خواسته شده است که مرتبط بودن سؤال‌ها را ارزیابی کنند. در صورتی که نسبت روایی محتوایی (CVR) برای سؤال اول ۰/۶ باشد، چند متخصص سؤال را ضروری تشخیص داده‌اند؟ (سال ۱۴۰۱)

۹ (۴)

۸ (۳)

۶ (۲)

۴ (۱)

۴۵- معلمی به منظور ارزشیابی تراکمی در پایان سال تحصیلی، قصد طراحی آزمون پیشرفت تحصیلی دارد. اگر به جای سؤالات چهارگزینه‌ای از سؤالات تشریحی برای این آزمون استفاده کند، کدام مورد درست است؟ (سال ۱۴۰۲)

(۱) روایی محتوایی آزمون کاهش می‌یابد.

(۲) اعتبار (پایایی) نمره‌های آزمون افزایش می‌یابد.

(۳) هزینه‌های تصحیح و نمره‌گذاری آزمون کاهش می‌یابد.

(۴) تعداد و گستره هدف‌های آموزشی افزایش می‌یابد.

روایی وابسته به ملاک (Criterion-related validity)

منظور از روایی ملاکی میزان ارتباط نمره‌های حاصل از یک آزمون با نمره‌های حاصل از یک آزمون یا وسیله اندازه‌گیری دیگر است. آزمون دوم که عملکرد فرد در آن پیش‌بینی می‌شود، ملاک نام دارد. شواهد وابسته به ملاک نشان می‌دهد که نمره‌های آزمون به گونه‌ای نظام‌دار با یک یا چند ملاک ارتباط دارد. رابطه بین آزمون و متغیر ملاک به صورت ضریب همبستگی بیان می‌شود و آن را ضریب روایی می‌نامند. اندازه ملاک دست کم باید دارای چهار ویژگی مطلوب باشد. این ویژگی‌ها به ترتیب اهمیت عبارت است از: (۱) مربوط بودن، (۲) بدون تورش (سوگیری) بودن، (۳) معتبر بودن، و (۴) در دسترس بودن.

برای به دست آوردن شواهد وابسته به ملاک می‌توان به دو نوع مطالعه پیش‌بینی و همزمان اشاره کرد:

- **روایی پیش‌بینی:**

زمانی که نمره‌های ملاک پس از گذشت فاصله زمانی از اجرای آزمون اول گردآوری می‌شود، به آن روایی پیش‌بینی گفته می‌شود. قدرت پیش‌بینی آزمون در عمل به روایی ملاکی پیش‌بین بستگی دارد. به عبارت دیگر همبستگی بین نمره‌های آزمون با ملاکی که در آینده به دست می‌آید مد نظر است. به عنوان مثال همبستگی بین نمره کنکور با نمره‌های دوسال بعد در دانشگاه (برای پیش‌بینی موفقیت تحصیلی).

- **روایی همزمان:**

یکی دیگر از انواع روایی وابسته به ملاک، روایی همزمان است. زمانی که نمره‌های آزمون ملاک همزمان با نمره‌های آزمون اصلی به دست می‌آید و رابطه همزمان آزمون و ملاک را مشخص می‌کند. به عنوان مثال همبستگی بین نمره فرد در یک پرسشنامه تشخیص بیماری روانی و نتایج درجه‌بندی متخصصان به صورت همزمان.

ضریب نابستگی

در همبستگی نمره آزمون (X) با نمره ملاک (Y)، نسبتی از واریانس نمره ملاک که توسط نمره آزمون تبیین نمی شود را اصلاً ضریب نابستگی گویند، که از طریق معادله زیر به دست می آید:

$$\sqrt{1 - r_{xy}^2} = \text{ضریب نابستگی}$$

تصحیح برای کاهش

اعتبار (پایایی) یک آزمون بر روایی آن اثر می گذارد. ضریب همبستگی نمره آزمون (X) و نمره ملاک (Y) تحت تأثیر اعتبار (پایایی) آنها قرار دارد. برآورد همبستگی دو نمره برای حالتی که هر دو کاملاً پایا و معتبر باشند، از طریق فرمول تصحیح برای کاهش به دست می آید. این معادله همبستگی بین نمره واقعی X و نمره واقعی Y را نشان می دهد.

$$\rho_{XY_T} = \frac{\rho_{XY}}{\sqrt{\rho_{XX}\rho_{YY}}}$$

ρ_{XY} = همبستگی نمره X و Y

ρ_{XX} = اعتبار (پایایی) آزمون X

ρ_{YY} = اعتبار (پایایی) ملاک Y

اگر یکی از نمره ها کاملاً پایا و دیگری ناپایا باشد، معادله به شکل زیر تغییر می کند.

$$\rho_{XY_T} = \frac{\rho_{XY}}{\sqrt{\rho_{YY}}}$$

گزینش و تصمیم گیری در آزمون های وابسته به ملاک

غالباً آزمون های وابسته به ملاک برای گزینش در طبقه های دو ارزشی (به عنوان مثال قبول یا مردود) به کار برده می شوند. این روش مستلزم برقراری یک نمره برش است که بر اساس آن افرادی که نمره آنها مساوی و یا بالاتر از نمره برش است در یک طبقه و بقیه در طبقه دیگر قرار می گیرند.

نتایج احتمالی در یک موقعیت گزینش دو ارزشی مشابه شکل زیر است.

نتایج پیش بینی شده			
درست		نادرست	
نتایج حقیقی	درست	صحیح	خطا
	نادرست	خطا	صحیح

یکی از موارد مرتبط با این آزمون ها قدرت آزمون در پیش بینی درست و ارزش تشخیصی آن است. دو خصوصیت عمده آزمون شامل حساسیت (Sensitivity) و ویژگی (Specificity) است.

حساسیت: توانایی آزمون برای پیدا کردن موارد درست است که آزمون آنها را به درستی تشخیص می دهد.

(مثبت حقیقی + منفی کاذب) / مثبت حقیقی = حساسیت آزمون

ویژگی (اختصاصی بودن): توانایی آزمون در پیدا کردن موارد منفی که آزمون آن را به درستی منفی تشخیص می دهد.

(منفی حقیقی + مثبت کاذب) / منفی حقیقی = ویژگی آزمون

نمونه سؤال گفتاری روان شناسی تربیتی

۴۶- مهم ترین ویژگی متغیر ملاک در مطالعه روایی آزمون، کدام است؟ (۱۴۰۳)

- (۱) سوگیری نداشتن (۲) دقت داشتن (۳) مربوط بودن (۴) در دسترس بودن

۴۷- اگر بررسی روایی (validity) یک آزمون، از ضریب همبستگی استفاده شده باشد، به احتمال زیاد، این همبستگی بین کدام موارد است؟ (سال ۱۴۰۲)

(۱) نمرات دو نیمه آزمون (۲) نمرات آزمون با یک ملاک بیرونی

(۳) نمرات بین دو فرم موازی (۴) نمرات یک گروه از آزمودنی ها در دو زمان مختلف

۴۸- برای اندازه گیری عملکرد شغلی کارکنان یک سازمان از پرسشنامه خود گزارش دهی عملکرد شغلی و همچنین درجه بندی سرپرست از عملکرد شغلی کارکنان استفاده شده است. اگر همبستگی بین نمره های دو اندازه گیری بالا باشد. کدام روایی پرسشنامه خود گزارش دهی حمایت می شود؟ (سال ۱۳۹۹)

(۱) واگرا (۲) تشخیص (۳) پیش بین (۴) همزمان

۴۹- همبستگی یک ملاک با متغیر پیش بین ۰/۴۲ است. اگر پایایی آزمون پیش بین و ملاک به ترتیب ۰/۸۱ و ۰/۴۹ باشد، ضریب همبستگی برآورد شده بین نمره های مشاهده شده آزمون پیش بین و نمره های واقعی آزمون ملاک، کدام است؟ (سال ۱۳۹۹)

(۱) ۰/۴ (۲) ۰/۴۶ (۳) ۰/۵۶ (۴) ۰/۶

۵۰- چنانچه پایایی دو آزمون حداکثر ۰/۷ و همبستگی بین آنها نیز ۰/۵ باشد، مقدار روایی آزمون حداکثر چقدر است؟ (سال ۱۳۹۷)

(۱) ۰/۲ (۲) ۰/۳۵ (۳) ۰/۵ (۴) ۰/۷۱

۵۱- اگر همبستگی نمره آزمون با متغیر ملاک برابر با ۰/۸ باشد، ضریب نایستگی کدام است؟ (سال ۱۴۰۰)

(۱) ۰/۲ (۲) ۰/۰۴ (۳) ۰/۶ (۴) ۰/۳۶

۵۲- کدام مورد به ترتیب در خصوص حساسیت و اختصاصی بودن ابزار سنجش، درست است؟ (سال ۱۴۰۲)

(۱) درصد مثبت های درست- درصد منفی های غلط

(۲) درصد مثبت های غلط که ابزار شناسایی کرده- درصد منفی های غلط

(۳) درصد مثبت های درست که ابزار شناسایی کرده- درصد منفی های درست

(۴) درصد مثبت های غلط که ابزار شناسایی کرده- درصد منفی های درست

۵۳- یک آزمون تشخیص اختلال یادگیری در بین ۵۰ دانش آموز سالم و ۵۰ دانش آموز مبتلا به اختلال یادگیری اجرا شد. از بین دانش آموزان سالم ۵ نفر و از بین دانش آموزان مبتلا به اختلال یادگیری ۴۰ نفر بالاتر از نقطه برش آزمون است، ویژگی آزمون (specificity) کدام است؟ (سال ۱۳۹۶)

(۱) ۰/۱۰ (۲) ۰/۴۵ (۳) ۰/۸۰ (۴) ۰/۹۰

روایی سازه (Construct Validity)

روایی سازه یک آزمون عبارت است از میزان صحت آن در اندازه گیری سازه نظری یا ویژگی مورد نظر. یک آزمون در صورتی دارای روایی سازه است که نمره های حاصل از اجرای آن، به مفاهیم یا سازه های نظریه مورد نظر مربوط باشد. روایی سازه می گوید معنای نمره چیست.

نکته: روایی سازه تمام انواع روایی را در بر می گیرد.

روش ها متعددی برای بررسی روایی سازه آزمون وجود دارد، شامل:

• روایی چند صفتی - چند روشی

روایی چند صفتی - چند روشی جنبه ای از روایی سازه است و زمانی به کار برده می شود که دو یا چند صفت با دو یا چند روش اندازه گیری می شود. مثال دو صفت درون گرایی و روان رنجوری با دو روش پرسش های صحیح - غلط و چند گزینه ای اندازه گیری شود. شکل ماتریس همبستگی این موارد در جدول زیر نمایش داده شده است.

روان رنجوری	روان رنجوری	درون گرایی	درون گرایی	
چند گزینه ای	صحیح - غلط	چند گزینه ای	صحیح - غلط	
۰/۱۷	۰/۲	۰/۷۵	۰/۸	درون گرایی
				صحیح - غلط
۰/۲۱	۰/۱۵	۰/۸۳	۰/۷۵	درون گرایی
				چند گزینه ای
۰/۷۱	۰/۸۵	۰/۱۵	۰/۲	روان رنجوری
				صحیح - غلط
۰/۹۱	۰/۷۵	۰/۲۱	۰/۱۷	روان رنجوری
				چند گزینه ای

قطر ماتریس بالا همبستگی هر آزمون با خودش را نشان می‌دهد و بیانگر اعتبار (پایایی) هر یک از آزمون‌هاست. به عنوان مثال اعتبار (پایایی) آزمون درون‌گرایی با روش صحیح غلط برابر با $0/8$ است.

در این ماتریس دو نوع روایی عمده نیز نشان داده می‌شود: روایی همگرا و روایی واگرا (تمیز)

- **روایی همگرا:** شواهد روایی همگرا از طریق همبستگی نمره‌های آزمون با نمره‌های دیگر آزمون‌هایی که همان سازه را اندازه می‌گیرند، جمع‌آوری می‌شود. بنابراین در ماتریس بالا، روایی همگرا از طریق همبستگی بالا بین نمره‌های آزمون‌هایی نشان داده می‌شود، که به وسیله روش‌های مختلف ویژگی‌های یکسانی را اندازه‌گیری می‌کنند. به عنوان مثال همبستگی بین درون‌گرایی به شیوه صحیح - غلط و به شیوه چندگزینه‌ای برابر با $0/75$ و بیانگر روایی همگرای این دو آزمون است.

- **روایی واگرا:** از طریق همبستگی پایین بین نمره‌های آزمون‌هایی که ویژگی‌های مختلفی را اندازه می‌گیرند، به ویژه زمانی که روش‌های یکسانی به کار برده می‌شود نشان داده می‌شود. به عبارت دیگر همبستگی ضعیف بین نمره‌های آزمون با نمره‌های آزمون دیگری که سازه متفاوتی را اندازه می‌گیرد. به عنوان مثال همبستگی بین روان‌رنجوری با روش صحیح - غلط و درون‌گرایی با روش صحیح - غلط برابر با $0/2$ است که مقدار پایین را نشان می‌دهد و بیانگر روایی واگرا است (روایی تشخیصی).

- **روایی عاملی**

روایی عاملی شکلی از روایی سازه است که از طریق تحلیل عاملی انجام می‌شود. تحلیل عاملی نشان دهنده شیوه‌های ریاضی گوناگون برای تحلیل همبستگی‌های درونی بین مجموعه‌ای از متغیرها (سؤال‌ها) و تبیین این همبستگی‌ها بر حسب تعداد محدودی از متغیرهاست که عوامل نامیده می‌شوند. به عبارت دیگر تحلیل عاملی تکنیکی است که کاهش تعداد زیادی از متغیرها را به صورت تعداد کوچکتري از ابعاد پنهان یا مکنون (عامل) امکان پذیر می‌سازد.

دو نوع تحلیل عاملی مفروض است: تحلیل عامل اکتشافی و تحلیل عامل تأییدی.

(۱) تحلیل عامل اکتشافی زمانی به کار می‌رود که از قبل هیچ اطلاعی در خصوص نحوه خوشه‌بندی سؤال‌ها وجود ندارد، و پژوهشگر قصد دارد ببیند که ماده‌های تشکیل دهنده یک آزمون از چند عامل تشکیل شده‌اند و هر ماده به کدام عامل تعلق دارد.

۲) تحلیل عامل تأییدی به آزمون فرضیه می‌پردازد. در این روش ساختار عاملی داده‌ها از پیش مشخص است و بررسی می‌شود که ساختار مفروض تا چه اندازه با داده‌ها برازش دارد.

در مورد تحلیل عاملی دو رویکرد وجود دارد. تحلیل عامل مشترک و تحلیل مؤلفه‌های اصلی. در رویکرد تحلیل مؤلفه‌های اصلی فرض بر این است که متغیرهای تحلیل، اطلاعات موجود در ماتریس همبستگی را به طور کامل تبیین می‌کنند، بنابراین اشتراک هر متغیر با عامل‌ها برابر با یک و اندازه‌گیری بدون خطاست. در این رویکرد در قطر ماتریس همبستگی عدد یک قرار می‌گیرد. در رویکرد تحلیل عاملی مشترک، واریانس مشترک متغیرها برابر یک نخواهد بود و مقداری واریانس اختصاصی نیز وجود دارد.

در تحلیل عاملی، واریانس متغیر (سؤال) به چند بخش افراز می‌شود:

واریانس کل = واریانس مشترک + واریانس اختصاصی + واریانس خطا

واریانس یگانه = واریانس اختصاصی + واریانس خطا

اعتبار (پایایی) = واریانس مشترک + واریانس اختصاصی

برخی اصطلاحات تحلیل عاملی:

عامل: یک متغیر فرضی است که روی یک یا چند متغیر مورد مشاهده تأثیر می‌گذارد.

بارعاملی: بار عاملی عددی است که از نظر معنی و اندازه خیلی شبیه به همبستگی است. هنگامی که عوامل ناهمبسته (متعامد) باشند، بارعاملی بیانگر همبستگی بین متغیر (سؤال) و عوامل است. سؤال‌هایی که روی یک عامل قرار می‌گیرند، دارای وزن عاملی یا بار عاملی در آن عامل هستند.

ماتریس عاملی: یک ماتریس متشکل از متغیرها و عامل‌هاست. نشان می‌دهد هر عامل از کدام متغیرها (سؤال‌ها) تشکیل شده است. معمولاً نقطه برش ۰/۳ برای تعلق یک متغیر به عامل در نظر گرفته می‌شود.

واریانس مشترک: مجذور بار عاملی هر متغیر روی هر عامل، واریانس مشترک متغیر و عامل را نشان می‌دهد. اگر عامل‌ها مستقل از هم باشند و بین آنها همبستگی نباشد (متعامد باشند)، مجموع مجذور بارعاملی هر متغیر روی عامل‌ها میزان اشتراک متغیر را با h^2 (میزان اشتراک) نشان می‌دهد.

چرخش عامل‌ها: چرخش وضعیت عامل‌ها را نسبت به متغیرها (سؤال‌ها) تغییر می‌دهد تا تفسیر راه حل آسان‌تر باشد. در تحلیل عاملی دو نوع چرخش برای عامل‌ها وجود دارد: چرخش متعامد زمانی که عامل‌ها مستقل از یکدیگرند و همبستگی ندارند و چرخش متمایل زمانی که عامل‌ها مستقل از هم نیستند و همبستگی دارند.

عوامل تهدید کننده روایی آزمون

- کوچک نمایی سازه (کم بازنمایی سازه): بدین معنا که آزمون محدود شده و تمام جنبه‌ها و ابعاد سازه را در بر نمی‌گیرد.
- واریانس بی ارتباط به سازه: بدین معنا که آزمون چیزی بیشتر از سازه مورد نظر را اندازه می‌گیرد. به عبارت دیگر آزمون دارای واریانس مازاد است و آلوده شده است. زمانی که در آزمون ویژگی وجود دارد که به سازه هدف مرتبط نیست، به عبارت دیگر آزمون دچار سوگیری شده است (بحث نامتغیر بودن اندازه‌گیری و کارکرد افتراقی سؤال (DIF) از موارد مربوط به سوگیری سؤال‌ها و مرتبط با بحث روایی است)

نمونه سؤال گنگور دگتری روان‌شناسی تربیتی

۵۴- در چه صورتی ماتریس بارهای عاملی (وزهای رگرسیون)، با همبستگی متغیرها و عامل‌ها برابرند؟ (سال ۱۴۰۱)

(۱) متغیرها نرمال باشند. (۲) عامل‌ها متعامد باشند.

(۳) متغیرها پیوسته باشند. (۴) اندازه نمونه ۱۰ برابر تعداد سؤال‌ها باشد.

۵۵- اگر واریانس متغیر ۱ برابر با ۲/۵ و بارعاملی آن روی عامل اول و دوم به ترتیب ۰/۵ و ۰/۶ باشد، واریانس خطای آن برابر با چند است؟ (سال ۱۴۰۱)

(۱) ۱/۳۱ (۲) ۱/۴۹ (۳) ۱/۵۲ (۴) ۱/۸۹

۵۶- اگر بارعاملی سؤال‌های ۱ و ۲ روی دو عامل به ترتیب $\alpha_{i1} = 0/6$ و $\alpha_{i2} = 0/4$ ، $\alpha_{j1} = 0/3$ و $\alpha_{j2} = 0/5$ باشد، کوواریانس دو سؤال برابر چند است؟ (سال ۱۴۰۱)

(۱) ۰/۳۸ (۲) ۰/۵۵ (۳) ۰/۸۱ (۴) ۱/۸۰

نرم یا هنجار

نرم یا هنجار متوسط عملکرد گروه نمونه‌ای از آزمودنی‌هاست که به روش تصادفی از یک جامعه تعریف شده انتخاب می‌شوند. نرم یک چهارچوب داوری است که نمره خام آزمودنی براساس آن تفسیر و درباره وی قضاوت می‌شود. به بیان دیگر نرم ضابطه یا ملاک و مقیاسی است که مبنای مقایسه و سنجش قرار می‌گیرد و به کمک آن می‌توانیم جایگاه فرد را مشخص کنیم، نمره‌های فرد را در آزمون‌های مختلف مقایسه کنیم و برای فرد نیمرخ روانی رسم کنیم. پس از تصحیح برگه‌ها، برای هر فرد نمره‌ای به نام نمره خام به دست می‌آید که به خودی خود قابل تفسیر نیست و باید در رابطه با آن اقداماتی انجام گیرد. به طور معمول نمره‌های خام تبدیل می‌شوند. تبدیل نمره‌های خام به دو صورت اساسی خطی و غیر خطی صورت می‌گیرد. در تبدیل خطی نمره‌های خام شکل توزیع نمره‌های تبدیل شده همانند توزیع نمره‌های خام است. همچنین تبدیل خطی اندازه همبستگی را تغییر نمی‌دهد. اگر برای تفسیر نمره خام آن را با میانگین یک گروه مرجع مقایسه کنیم، به آن گروه مرجع، گروه هنجار می‌گویند.

طبقه بندی انواع نرم

نرم یا هنجار از دیدگاه‌های مختلف قابل طبقه‌بندی است:

انواع نرم از لحاظ وسعت جامعه هدف:

- نرم ملی یا کشوری: اگر جامعه هدف به صورت همه افراد یک کشور که در یک یا چند صفت مشترکند باشد و افراد گروه نرم از این جامعه انتخاب شوند، نرم ملی یا کشوری نامیده می‌شود. در این نرم چون جامعه هدف گسترده و ناهمگون است، از نمونه‌گیری خوشه‌ای و طبقه‌ای استفاده می‌شود.
- نرم منطقه‌ای: اگر جامعه هدف به صورت همه افراد یک استان یا منطقه از یک کشور (شامل چند استان همجوار) باشد. جدول نرم به دست آمده نرم منطقه‌ای خواهد بود.
- نرم محلی: اگر جامعه از یک شهر، یک بخش، قسمتی از یک شهر یا حتی دانش‌آموزان یک مدرسه، یا دانشجویان یک دانشگاه باشد، جدول نرم به دست آمده، نرم محلی خواهد بود.

انواع نرم از لحاظ نوع گروه نرم:

■ نرم سنی

نرم سنی برای سنجش خصایصی به کار می‌رود که با افزایش سن رابطه دارد، مثل هوش و برخی از استعدادهای شناختی. اگر صفت مورد مطالعه با بالارفتن سن افزایش نسبتاً یکنواخت و مداومی نشان دهد، ممکن است مناسب باشد که نمره مربوط به آن را به یک معادل سنی تبدیل کرد، و نرم هر سن به عنوان ارزش متوسط صفت مورد نظر برای افرادی که دارای آن سن هستند، تعریف می‌شود. نرم سنی نمره میانی افراد گروه سنی معین در یک آزمون است. به عنوان مثال: اگر بدانیم متوسط قد در پسران ۱۲ ساله ۱۵۰ سانتی متر، و متوسط قد در پسران ۸ ساله ۱۲۵ سانتی متر است. می‌توان قد سایر کودکان را با این ملاک مقایسه کرد. اگر قد یک پسر ۱۰ ساله ۱۵۰ باشد، می‌توان گفت نسبت به سنش قد بلندتر و مطابق با بچه‌های ۱۲ ساله است، و یا اگر قد یک پسر ۱۰ ساله ۱۲۵ باشد، می‌توان گفت قد وی از همسالانش کوتاه‌تر و مطابق با پسران ۸ ساله است.

■ نرم کلاسی

برای هر صفتی که از یک کلاس به کلاس دیگر افزایشی نسبتاً یکنواخت داشته باشد می‌توان نرم کلاسی تهیه کرد. این نرم یا هنجار مشابه نرم سنی است با این تفاوت که در نرم سنی گروه‌های مرجع گروه‌های سنی هستند، و در نرم کلاسی گروه‌های مرجع گروه‌های کلاسی هستند. با به دست آوردن میانگین نمره کلاس‌های مختلف می‌توان معادل یا هنجار کلاسی را به دست آورد و افراد را با آن مقایسه کرد.

یکی از محدودیت‌های این دو نوع نرم این است که واحدهای مقیاس این دو نوع نرم در طول مقیاس برابر نیستند.

نرم سن عقلی

سن عقلی یک آزمودنی معین برابر با سن تقویمی آن دسته از آزمودنی‌های هم سال او در یک گروه هنجاریابی است که نمره میانی آنها با نمره آزمودنی مورد نظر برابر است.

نسبت سن عقلی (MA) را بر سن تقویمی (CA) ضربدر ۱۰۰، هوشبهر (IQ) می‌نامند.

$$IQ = \frac{MA}{CA} \times 100$$

انواع نرم از لحاظ مقیاس:

○ نرم‌های درصدی

اگر نمره آزمون یک فرد را به جای سنجش با متوسط عملکرد گروه‌های سنی یا کلاسی با گروهی بسنجیم که وی می‌تواند از لحاظ منطقی عضو آن باشد، نرم درصدی را به کار برده‌ایم. با نرم درصدی می‌توان نمره افراد شرکت کننده در یک آزمون را نسبت به یکدیگر مقایسه کرد.

رتبه درصدی: وضعیت نسبی فرد را در گروه برحسب کسانی که نمره پایین‌تر از او گرفته‌اند مشخص می‌کند. هر نمره خام دارای یک رتبه درصدی است.

محدودیت‌ها:

- ۱) لازم است گروه مرجع گروهی باشد که فرد منطقاً عضو آن است. بنابراین نیاز به جداول نرم متعدد دارد.
- ۲) عدم تساوی واحدها در طول مقیاس، مثلاً اختلاف ۵۰ با ۵۵ با اختلاف ۹۰ و ۹۵ مساوی نیست.
- ۳) شکل توزیع درصدها در گروه هنجار راست گوشه است (مستطیل شکل) نه به شکل منحنی طبیعی، در نتیجه برای استفاده از روش‌های آماری باید از به کار بردن درصدها خودداری کرد.

○ نمره‌های تراز شده

برخلاف نرم درصدی، نمره‌های تراز شده از نوع مقیاس فاصله‌ای هستند، که اندازه‌های آنها در سرتاسر مقیاس برابرند. نمره‌های تراز شده، نمره‌های تبدیل شده‌ای هستند که می‌توان آنها را بر حسب میانگین و انحراف معیار دلخواه محاسبه کرد.

❖ نمره Z

در عمل همه نمره‌های تراز شده بر مبنای نمره استاندارد Z محاسبه می‌شوند. با تبدیل نمره خام به نمره Z توزیعی به دست می‌آید، که میانگین آن برابر با صفر و انحراف استاندارد آن برابر با یک است.

$$Z = \frac{X - \bar{X}}{S}$$

X = نمره خام

$\bar{X}$ = میانگین نمره‌های خام

S = انحراف معیار نمره‌های خام

چون نمره Z دارای ارقام اعشاری است و نیمی از نمره‌ها در توزیع نرمال منفی است. بنابراین کار کردن با آن دشوار و معمولاً این نمره به نمرات استاندارد دیگری نظیر نمره استاندارد T، IQ، CEEB، نمره‌های نه بخشی و ... تبدیل می‌شود.

❖ نمره استاندارد T

نمره‌ای تراز شده با میانگین ۵۰ و انحراف معیار ۱۰ است.

$$T = 10Z + 50$$

❖ نمره هوشبهر انحرافی و کسلر

نمره‌ای تراز شده با میانگین ۱۰۰ و انحراف معیار ۱۵ است.

$$IQ = 15Z + 100$$

❖ نمره‌های تراز شده CEEB

نمره‌های آزمون‌های ورودی دانشگاه با میانگین ۵۰۰ و انحراف معیار ۱۰۰ است.

$$CEEB = 100Z + 500$$

❖ نمره های نه بخشی Stanine

نرمی که برای درجه بندی افراد در نیروی هوایی آمریکا به کار رفت و بعدها برای سایر مقاصد آموزشی و ارزشیابی مورد استفاده قرار گرفت. مقیاس نمره های تراز شده با میانگین ۵ و انحراف معیار ۲ است.

$$\text{نمره نه بخشی} = 2Z + 5$$

❖ نمره AGCT

نرمی است که ارتش آمریکا برای نمرات آزمون طبقه بندی عمومی ارتش به کار می برد. میانگین آن ۱۰۰ و انحراف معیار آن ۲۰ است.

$$AGCT = 20Z + 100$$

مثال: نمره دانش آموزی که در آزمونی با میانگین ۵۶ و واریانس ۱۶، نمره ۵۸ گرفته است را به نمره های تراز شده T، IQ و نمره های نه بخشی تبدیل کنید.

(نکته: برای محاسبه نمره استاندارد Z، احتیاج به انحراف معیار آزمون داریم، انحراف معیار برابر با جذر واریانس است، در نتیجه انحراف معیار این آزمون برابر با ۴ است.)

$$Z = \frac{58 - 56}{4} = 0.5$$

$$T = 10 \times 0.5 + 50 = 55$$

$$IQ = 15 \times 0.5 + 100 = 107.5$$

$$Stnine = 2 \times 0.5 + 5 = 6$$

نیمرخ روانی

انواع نرم‌های مطرح شده وسیله‌ای فراهم می‌سازند که نمره‌های آزمون‌های کاملاً متفاوت طوری به صورت واحدهای مشترک بیان شوند که بتوان آنها را مستقیماً با یکدیگر مقایسه نمود. نیمرخ روانی را می‌توان به صورت نمودار ستونی یا نمودار چندضلعی ترسیم کرد. برای رسم نیمرخ بر روی محور افقی خصایص مورد اندازه‌گیری (مثلاً خرده‌مقیاس - های آزمون و کسلر) و بر روی محور عمودی نرم تراز شده قرار می‌گیرد. برای رسم نیم رخ روانی می‌توان از رتبه‌های درصدی و انواع نمره‌های تراز شده استفاده کرد.

نمونه سؤال کنگور دگتری روان‌شناسی تربیتی

۵۷- واریانس متغیر تصادفی X برابر با $\sigma^2 = 225$ است. در صورتی که نمره استاندارد معادل Z با $X=76$ برابر با ۲- باشد میانگین متغیر X کدام است؟ (سال ۱۴۰۱)

(۱) ۷۳ (۲) ۹۱ (۳) ۱۰۶ (۴) ۱۱۴

۵۸- با تبدیل $Y = 2\left(\frac{X - \bar{X}}{S}\right) + 5$ کدام نمره استاندارد به دست می‌آید؟ (سال ۱۳۹۹)

(۱) نه بخشی (۲) رتبه درصدی (۳) هوشبهر انحرافی (۴) معادل بهنجار

۵۹- کدام مورد در تبدیل خطی نمره‌های آزمون تغییر می‌کند؟ (سال ۱۳۹۸)

(۱) کجی و کشیدگی (۲) پایایی و روایی

(۳) جایگاه افراد روی توزیع (۴) میانگین و انحراف استاندارد

مقیاس‌های سنجش نگرش

نگرش نوعی حالت آمادگی و در واقع تمایل به کنش یا واکنشی در برابر محرکی خاص است. برای اندازه‌گیری نگرش روش‌های مختلفی وجود دارد:

مقیاس ترستون

مقیاس ترستون با دو روش مقایسه‌های زوجی و فاصله‌های یکسان‌نما تهیه می‌شود. در روش فاصله‌های یکسان‌نما پژوهشگر ابتدا تعداد زیادی عبارت درباره یک موضوع یا مفهوم خاص را انتخاب می‌کند. اظهارنظرها در اختیار گروه بزرگی از افراد قرار می‌گیرد تا برحسب یک مقیاس یازده درجه‌ای بین دو حد نگرش مثبت تا منفی مرتب شود. سپس ارزش مقیاسی محاسبه می‌شود، که برابر نمره میانی داده شده به هر عبارت توسط داوران است.

روش لیکرت

در این روش تعداد زیادی از اظهارنظرهای مساعد و نامساعد در رابطه با نگرش نسبت به یک موضوع، درباره گروه بزرگی از آزمودنی‌ها اجرا و نظر خود را به صورت طیفی (از کاملاً موافقم تا کاملاً مخالفم) در برابر هر یک از اظهارنظرها مشخص می‌کنند.

مقیاس گاتمن

یک مقیاس رتبه‌ای یا ترتیبی است، که تک بعد یا عامل را اندازه می‌گیرد. در مقیاس گاتمن عبارت‌ها به ترتیب درجه دشواری از آسان‌ترین به دشوارترین تنظیم می‌شود. اگر کسی در یک سؤال شکست بخورد نمی‌تواند به سؤال بعدی پاسخ دهد و اگر کسی به سؤال پاسخ درست بدهد به همه سؤال‌های پیش از آن پاسخ درست می‌دهد.

نظریه صف مکنون (نظریه سؤال - پاسخ) (IRT)

در نظریه سؤال - پاسخ فرض بر این است که صفت عمده‌ای (نمره واقعی یا خصیصه مکنون) بر عملکرد تمام سؤال‌ها اثر می‌گذارد و به آن نظریه کارآمد نمره واقعی گفته می‌شود.

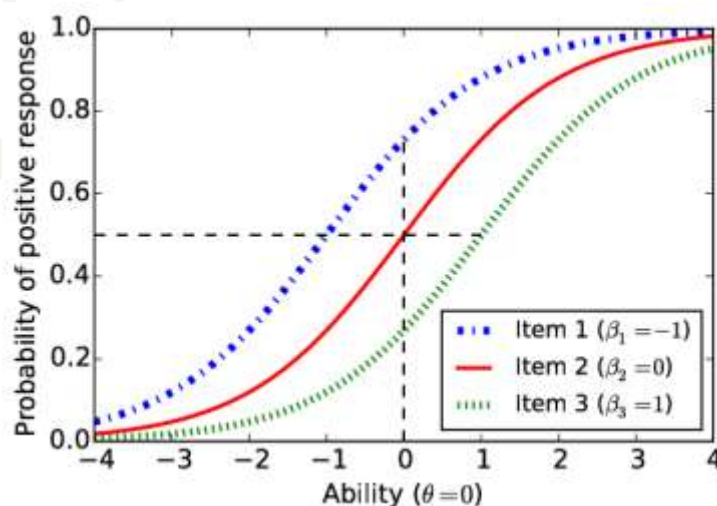
نظریه سؤال - پاسخ بر دو مفروضه اساسی استوار است:

(۱) عملکرد آزمودنی روی یک سؤال می‌تواند به وسیله مجموعه‌ای عوامل که، صفات مکنون یا توانایی نامیده می‌شوند، پیش‌بینی شود.

(۲) رابطه میان عملکرد آزمودنی در سؤال و مجموعه از صفات زیربنایی به وسیله یک تابع افزایشی یکنواخت که تابع ویژه سؤال یا خم ویژه سؤال (ICC) خوانده می‌شود، قابل توصیف است.

تابع ویژه سؤال (خم ویژه سؤال):

خطی است احتمال پاسخ درست به سؤال را به عنوان تابعی از توانایی آزمودنی در صفت مکنون (θ) و پارامترهای سؤال یعنی سطح دشواری (b)، ضریب تشخیص (a) و عامل حدس و گمان (c) نمایش می‌دهد. به عبارت دیگر در نظریه سؤال - پاسخ احتمال اینکه شخصی به سؤال پاسخ درست (پاسخ مدنظر) را بدهد به توانایی وی و پارامترهای سؤال بستگی دارد.



خم ویژه سؤال سه پارامتر برای سؤال در نظر می گیرد:

- سطح دشواری سؤال یا پارامتر مکانی: نقطه‌ای روی مقیاس توانایی که احتمال پاسخ صحیح به آن ۰/۵ است. یعنی جایی که $\theta=b$ باشد، هر آزمودنی احتمال ۰/۵ برای پاسخ درست به سؤال را دارد.
- ضریب تشخیص یا شیب خم: ضریب تشخیص سؤال که با (a) یا آلفا نمایش داده می‌شود، شیب منحنی را در نقطه مکانی نشان می‌دهد. بدین معنا که سؤال تا چه اندازه افراد دارای توانایی بالا را از توانایی پایین تفکیک می‌کند.
- C یا پارامتر حدس: احتمال اینکه فردی با توانایی پایین به سؤال پاسخ درست دهد با C نشان داده می‌شود. توابع ویژگی سؤال در این موارد سه پارامتری است.

به طور کلی مدل‌های IRT براساس تعداد پارامترهای سؤال به سه دسته: مدل‌های یک پارامتری (مدل منطقی راش)، دوپارامتری و سه پارامتری تقسیم می‌شوند:

- (۱) مدل‌های یک پارامتری: مدل‌های یک پارامتری تنها از لحاظ درجه دشواری، یعنی جایگاهشان روی مقیاس توانایی با یکدیگر متفاوتند، در مدل یک پارامتری فرض بر این است که تنها دشواری سؤال بر عملکرد آزمودنی تأثیر می‌گذارد. اگر ضریب دشواری همه سؤال‌ها یک در نظر گرفته شود مدل، به مدل یک پارامتری معروف است. اما در الگوی منطقی راش فرض بر این است ضریب تشخیص برای تمام سؤال‌ها ثابت است، در نتیجه شیب همه نمودارها یکسان است.
- (۲) مدل‌های دو پارامتری: در مدل دو پارامتری فرض بر این است که علاوه بر دشواری سؤال، ضریب تشخیص نیز بر عملکرد آزمودنی تأثیر می‌گذارد.
- (۳) مدل‌های سه پارامتری: مدل دوپارامتری برای آزمودنی‌هایی که مقدار خصیصه مکنون در آنها پایین است، اما به سؤال‌ها پاسخ درست داده‌اند، مناسب نیست. این اتفاق زمانی می‌افتد که آزمودنی‌ها با توانایی پایین می‌توانند از طریق حدس و گمان به سؤال پاسخ درست دهند. در این حالت فرض می‌شود که علاوه بر دشواری سؤال و ضریب تشخیص، عامل حدس و گمان نیز بر عملکرد آزمودنی تأثیر می‌گذارد.

دو الگوی اصلی خصیصه مکنون بر اساس الگوی اجایو طبیعی و الگوی منطقی (لوجستیک) است، احتمال اینکه یک آزمودنی با مقدار خصیصه مکنون θ به یک سؤال پاسخ دهد، یک تابع تراکمی یا منطقی است.

تابع تراکمی

$$P(\theta) = \int_{-\infty}^{a(\theta-B)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

تابع منطقی (لجستیک)

$$P(\theta) = \frac{1}{1 + e^{-Da(\theta-b)}}$$

در این فرمول‌ها b درجه دشواری سؤال و a ضریب تشخیص است. این الگوها چون دارای دو پارامتر است به الگوی دو پارامتری معروف است. مقدار D عدد ثابت $1/7$ است، و برای اینکه نتایج تابع تراکمی و منطقی شبیه یکدیگر باشند، در ضریب تشخیص ضرب می‌شود (شیب تابع منطقی از شیب تابع تراکمی کمتر است و به این طریق اصلاح می‌شود).

برای برآورد خصیصه مکنون از روش‌های برآورد مختلفی استفاده می‌شود که معروفترین آن روش بیشینه (حداکثر) درست نمایی است.

مفروضه‌های نظریه سؤال - پاسخ

• تکنوا بودن

با افزایش توانایی (θ) احتمال پاسخ صحیح افزایش می‌یابد.

- بعدیت (تک بعدی بودن، همگنی)

بعدیت بدین معناست که فقط یک توانایی به وسیله مجموعه سؤال‌های آزمون اندازه‌گیری شود. یعنی حضور یک مؤلفه غالب که عملکرد در آزمون متأثر از آن است. آزمون‌های همگن از سؤال‌هایی تشکیل شده که تنها یک صفت یا ویژگی را اندازه‌گیری می‌کنند.

- استقلال موضعی (استقلال مکانی)

منظور از استقلال موضعی این است که، اگر توانایی‌هایی که عملکرد در آزمون را تحت تأثیر قرار می‌دهند ثابت نگه داشته شوند، پاسخ‌های آزمودنی‌ها به هر زوجی از سؤال‌ها به لحاظ آماری ناهمبسته و مستقل باشد. احتمال اینکه آزمودنی به دو سؤال آزمون پاسخ درست دهد برابر با حاصلضرب احتمال‌های پاسخگویی درست به هر یک از سؤال‌هاست. اگر عمل یا معلومات پاسخگویی به یک سؤال تحت تأثیر پاسخگویی به سؤال دیگر باشد آن سؤال استقلال موضعی ندارد. (احتمال اینکه آزمودنی بتواند به یک سؤال پاسخ درست بدهد، به پاسخ درست او به هیچ یک از سؤال‌های دیگر بستگی ندارد)

آگاهی آزمون (اطلاعات آزمون)

تابع آگاهی به دقت برآورد توانایی آزمودنی مربوط می‌شود (تحدودی مشابه مفهوم اعتبار (پایایی) در نظریه کلاسیک). هر سؤال مقداری آگاهی درباره سطح توانایی آزمودنی می‌دهد. مقدار این آگاهی بستگی به این دارد که سطح دشواری سؤال چقدر به توانایی آزمودنی نزدیک باشد. در مدل یک پارامتری حداکثر اندازه آگاهی جایی است که سطح توانایی آزمودنی با سطح دشواری سؤال برابر باشد.

تابع آگاهی در مدل یک پارامتری

$$I(\theta) = p(\theta) \times q(\theta)$$

تابع آگاهی در مدل دو پارامتری

$$I(\theta) = a^2 p(\theta) \times q(\theta)$$

خطای معیار اندازه‌گیری

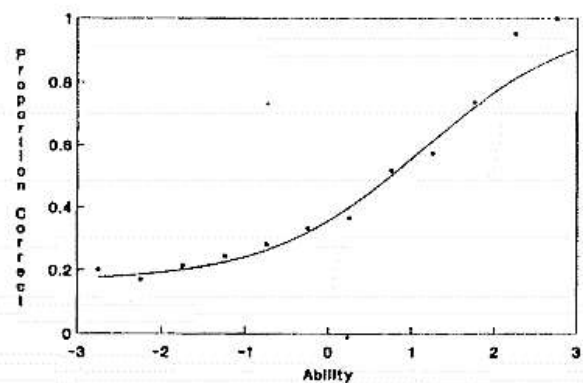
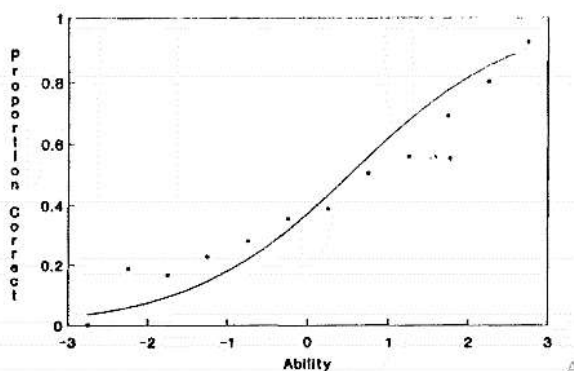
خطای معیار اندازه‌گیری با تابع آگاهی رابطه معکوس دارد، و برای هر سطح توانایی برابر با معکوس جذر تابع آگاهی سؤال در آن سطح توانایی است.

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

هر اندازه آگاهی سؤال در سطح توانایی مشخص بیشتر باشد، خطای اندازه‌گیری برای آن سطح توانایی کمتر است.

برازش داده‌ها با مدل

موفقیت روش‌های IRT تنها زمانی دست‌یافتی است که برازش میان داده‌ها و مدل IRT برقرار باشد. برای بررسی برازش مدل با داده‌ها می‌توان از روش‌های آماری و نموداری استفاده کرد. در روش‌های آماری از تحلیل باقیمانده‌ها استفاده می‌شود، به عنوان مثال از آماره Q_1 مجذور کا برای نشان دادن نیکویی برازش مدل استفاده می‌شود. در روش‌های نموداری از مقایسه نسبت‌های مشاهده شده و مورد انتظار استفاده می‌شود. در روش نموداری قضاوت درباره برازش داده‌ها با مدل، برپایه مقایسه خم ویژه سؤال (مقادیر موردانتظار) با مقادیر مشاهده شده صورت می‌گیرد. اگر نسبت‌های مشاهده شده، نزدیک خم ویژه باشد داده‌ها با مدل برازش دارد.



سنجش انطباقی

یکی از کاربردهای نظریه سؤال- پاسخ برای سنجش انطباقی است. شرایط تستی ایده آل آن است که به هر آزمودنی، آزمونی داده شود که با توانایی او برازش دارد یا سازگاری داشته باشد. در سنجش انطباقی تلاش می شود دشواری سؤال ها با توانایی آزمودنی منطبق و جور شود. در سنجش انطباقی نخست سؤال هایی به آزمودنی ارائه می شود که دارای سطح دشواری پایین و متوسط هستند. اگر آزمودنی در پاسخ دادن به این ماده ها موفق بود در مرحله بعدی می توان ماده های دشوارتری به او ارائه کرد و اگر پاسخ غلط داد، می توان در مرحله بعد سؤال های آسان تر به وی ارائه کرد و هر کس به آزمونی پاسخ می دهد که با توانایی او برازش دارد.

آزمون های طول ثابت برای اغلب آزمودنی ها به ویژه آزمودنی ها با توانایی بالا و پایین ناکارا است. در سنجش انطباقی تنها سؤالاتی انتخاب می شود که در سطح توانایی آزمودنی بیشترین آگاهی را دارند. در نتیجه بدون کاهش در دقت و اعتبار (پایایی) اندازه گیری، آزمون کوتاه تر می شود و نیازی نیست سؤال های ساده روی آزمودنی های توانمند و سؤال های سخت روی آزمودنی های ضعیف اجرا شود.

کنش افتراقی سؤال

یکی از عوامل مرتبط با روایی آزمون بحث سوگیری است. بدین معنا که آیا در آزمون ویژگی هایی وجود دارد که به سازه یا عملکرد هدف ارتباطی نداشته باشد. یکی از روش های بررسی سوگیری سؤال در نظریه کلاسیک اندازه- گیری «تغییرناپذیری اندازه گیری» نام دارد. یکی دیگر روش های بررسی سوگیری در سؤال های آزمون، در نظریه سؤال پاسخ است، «کنش افتراقی سؤال» نام دارد.

کارکرد افتراقی سؤال DIF به پدیده ای گفته می شود که در آن دو آزمودنی با سطح دانش، مهارت، توانایی یا استعداد یکسان، احتمالات متفاوتی در دادن پاسخ صحیح یا مطلوب به سؤالات داشته باشند. به عبارت دیگر در صورتی که افراد با توانایی یکسان ولی متعلق به گروه های متفاوت دارای احتمال یکسانی برای دادن پاسخ صحیح نباشند، سؤال DIF دارد. روش های مختلفی برای بررسی DIF سؤال وجود دارد. از جمله مقایسه پارامترهای خم ویژه سؤال در دو گروه اقلیت و اکثریت، مقایسه خم ویژه سؤال در دو گروه و اندازه گیری مساحت بین دو خم.

همترازسازی

قابل مقایسه بودن نمره‌ها در آزمون‌هایی که توانایی یکسانی را اندازه می‌گیرند، موضوعی با اهمیت و قابل ملاحظه برای متخصصان اندازه‌گیری است. اگر دو آزمودنی در آزمون‌های متفاوتی شرکت کنند چگونه نمره‌های آن‌ها را می‌توان مقایسه کرد؟

برای مقایسه نمراتی که از آزمون‌های X و Y به دست می‌آیند، فرایند همترازسازی (معادل سازی) نمره‌های دو آزمون باید انجام شود. از طریق این فرایند، رابطه‌ای میان نمره‌های X و Y برقرار می‌گردد و نمره‌های آزمون X به مقیاس نمرات آزمون Y تبدیل می‌شود.

در نظریه کلاسیک اندازه‌گیری دو روش عمده برای همترازسازی وجود دارد: همترازسازی صدک‌های یکسان و همترازسازی خطی.

(۱) همترازسازی صدک‌های یکسان: در این روش نمره‌های آزمون‌های X و Y زمانی همتراز در نظر گرفته می‌شوند که مرتبه‌های درصدی آن‌ها در هر دو گروه مفروض یکسان باشد.

(۲) همترازسازی خطی: در این روش فرض بر این است که نمره x در آزمون X و نمره y در آزمون Y به طور خطی با هم رابطه دارند:

$$y = ax + b$$

مقیاس‌سازی یا همترازسازی در نظریه IRT به معنای قرار گرفتن پارامترهای توانایی و پارامترهای سؤال (درجه دشواری و ضریب تشخیص) مبتنی بر دو آزمون بر روی مقیاس مشترک است.

مقایسه نظریه کلاسیک و IRT

نظریه کلاسیک اندازه‌گیری	نظریه سؤال - پاسخ
خطای معیار اندازه‌گیری برای همه نمره‌ها یکسان است.	خطای معیار اندازه‌گیری برای هر سطح توانایی متفاوت است.
اعتبار (پایایی) آزمون‌های بلند بیشتر از آزمون‌های کوتاه است	اعتبار (پایایی) آزمون کوتاه ممکن از آزمون بلند بیشتر باشد.
بر آورد بدون سوگیری پارامتر سؤال‌ها به معرف بودن گروه نمونه بستگی دارد.	بر آورد بدون سوگیری ویژگی‌های سؤال را می‌توان در مورد نمونه نامعرف نیز به دست آورد.

نظریه کلاسیک اندازه گیری	نظریه سؤال - پاسخ
نمره های آزمون زمانی معنا پیدا می کند که با گروه نرم مقایسه شود.	نمره های آزمون زمانی معنا پیدا می کند که فاصله آنها از ماده ها یا سؤال ها مقایسه شود.
ویژگی مقیاس فاصله ای زمانی توزیع نمره های آزمون نرمال باشد حاصل می شود.	ویژگی مقیاس فاصله ای زمانی تحقق می یابد که روش اندازه گیری توجه پذیر استفاده شود.

نمونه سؤال انگور دکتري روان شناسی تربیتی

۶۰- اگر میزان آگاهی یک سؤال در $\theta = 1/5$ برابر $0/25$ باشد، میزان خطای استاندارد در این سطح چقدر است؟ (۱۴۰۳)

(۱) $1/5$ (۲) $1/75$ (۳) $2/3$ (۴) تقریباً ۴

۶۱- عبارت «اصلاح نمره های افراد به خاطر دشواری فرم های مختلف یک آزمون» به کدام مورد اشاره دارد؟ (سال ۱۴۰۲)

(۱) استاندارد سازی (۲) همتراز سازی (۳) تغییر ناپذیری اندازه گیری (۴) کارکرد افتراقی سؤال

۶۲- منحنی ویژگی سؤال در چه صورت دارای برآزش خواهد بود؟ (سال ۱۳۹۶)

(۱) پراکندگی داده ها به توزیع نرمال نزدیک باشد.

(۲) توزیع نمونه گیری داده ها به توزیع نرمال نزدیک باشد.

(۳) مجذور فواصل بین داده های مشاهده شده کمترین باشد.

(۴) مجذور فواصل بین داده های مشاهده شده و نمودار کمترین باشد.

۶۳- کدام ویژگی سؤالات آزمون برای به کارگیری نظریه های نمره کارآمد واقعی و الگوهای خصیصه مکنون الزامی است؟ (سال ۱۳۹۶)

(۱) وابستگی موضوعی (۲) دشواری متوسط (۳) چندبعدی بودن (۴) همگنی