مقایسه مدل اندازه‌گیری کلاسیک و مدل غیرپارامتریک سؤال ـ پاسخ از نظر ویژگی‌های سؤال

دکتر علی دلاور*
علی مقدم زاده**
سیده طیبه مطیعی لنگرودی***

چکیده

هدف پژوهش حاضر، مقایسه دو نظریه کلاسیک اندازه‌گیری و مدل سؤال ـ پاسخ غیرپارامتریک از نظر ویژگی‌های سؤال بوده است. روش تحقیق از نوع کاربردی ـ توصیفی بوده و برای دستیابی به این هدف، پژوهش توصیفی انجام شده است. در بررسی عملی، برای آزمون اختصاصی ریاضی، از پاسخنامه‌های داوطلبان رشته ریاضی ـ فیزیک در آزمون ورودی دانشگاه‌های کشور در سال ۱۳۸٤ استفاده شد. از بین کلیه داوطلبان رشته ریاضی ـ فیزیک شرکت‌کننده در آزمون سال ۱۳۸٤، به روش نمونه‌گیری سیستماتیک، یک گروه نمونه ۳۰۰۰ نفری انتخاب شد و سه سؤال این پژوهش، مورد بررسی قرار گرفت. برای تحلیل داده‌ها از روش‌های آماری مورد استفاده در مدل کلاسیک (شامل میانگین یا درجه دشواری سؤالات، واریانس سؤالات و همبستگی دورشته‌ای) استفاده شد. برای تحلیل داده‌ها در نظریه غیرپارامتریک سؤال ـ پاسخ، از روش آزمون **t** وابسته، ضریب همبستگی و آزمون معنی‌داری استفاده شد. نتایج تحقیق نشان داد که پارامترهای محاسبه شده سؤال در مدل غیرپارامتریک سؤال ـ پاسخ، برعکس مدل کلاسیک، وابسته به سؤال

* استاد دانشکده روان شناسی و علوم‌تربیتی دانشگاه علامه‌طباطبایی_(delavarali@yahoo.com)
** کارشناس ارشد سنجش و اندازه‌گیری دانشگاه علامه‌طباطبایی و دفتر آزمون‌سازی و روان‌سنجی سازمان سنجش آموزش کشور(Irsoyali_s2000@yahoo.com)
*** کارشناس ارشد تحقیقات آموزشی دانشگاه تهران و دفتر آزمون‌سازی و روان‌سنجی سازمان سنجش آموزش کشور (stml90@yahoo.com)

Archive of SID
www.SID.ir

است نه به آزمودنی. بنابراین، می‌توان نتیجه گرفت که نظریه سؤال ـ پاسخ غیرپارامتریک، هم از لحاظ نظری و هم از لحاظ عملی بر نظریه کلاسیک مزیت و برتری دارد.

کلید واژگان: مدل کلاسیک اندازه‌گیری، مدل غیرپارامتریک سؤال ـ پاسخ، ویژگی‌های سؤال.

مقدمه

به‌قول ثرندایک[1] (۱۹۸۲) تاریخ اندازه‌گیری‌های روانی و تربیتی در قرن بیستم ، در واقع تاریخ کشف و اختراع ابزارها و روش‌های اندازه‌گیری است که به طریقی استاندارد و تحت شرایط یکسان، رفتارهایی را که منعکس کننده خصیصه‌های افراد است، آشکار می‌کند و مورد سنجش قرار می‌دهد. ارائه نظریه‌های نوین اندازه‌گیری در قرن حاضر، به پیشرفت فنون و ابزارهای استاندارد شده‌ای انجامیده است که اندازه‌گیری و تبدیل توانش‌های فردی به مقیاس‌های قابل قبول برای توصیف، تفسیر و برآورد تفاوت‌های فردی را امکان‌پذیر می‌سازد (افروز و هومن، ۱۳۷۵). تلاش‌های نخستین برای تکوین نظریه کلاسیک اندازه‌گیری[2] در دهه ۱۸۹۰ آغاز شد. این نظریه، از روش‌های دیرینه برای ساخت و توسعه آزمون‌ها در حوزه علوم انسانی است که از اوایل دهه ۱۹۰۰ برای توسعه ابزارهای اندازه‌گیری و تعیین میزان همخوانی آزمون‌ها با نظریه و نمره‌گذاری امتحانات استفاده شده است. اوج تکامل این نظریه را بعد از اسپیرمن[3]، می‌توان در دو کتاب: (۱) «مبانی نظری آزمون‌های روانی[4]» گالیکسن[5] (۱۹۵۰) و «تئوری‌های آماری نمرات آزمون‌های روانی[6]» لرد و ناویک[7] (۱۹۶۸) مشاهده کرد (همبلتون و واندر لیندن[8]، ۱۹۸۲).

گرچه نظریه کلاسیک اندازه‌گیری مدت زمان طولانی به جامعه روان سنجی خدمت کرده است؛ اما برخی مطالعات، محدودیت‌هایی را در این نظریه (از جمله گالیکسن، ۱۹۵۰؛ لرد و ناویک، ۱۹۶۸؛ همبلتون و سوامیناتان و راجرز[9]، ۱۹۹۱) و در آزمون‌های ساخته شده بر اساس آن نشان می‌دهد (رک: لرد، ۱۹۸۰ و همبلتون، ۱۹۸۹).

زمینه ارائه نظریه‌های جدید اندازه‌گیری، از اوایل نیمه دوم قرن بیستم و به وسیله افرادی چون لرد (در سال‌های ۱۹۵۲تا ۱۹۵۳)، راش[10] (۱۹۵۸ تا ۱۹۶۸)، رایت[11] (۱۹۶۸)، همبلتون (۱۹۷۹ و ۱۹۸۳) و ... فراهم شد. نظریات جدید اندازه‌گیری، چه از لحاظ روش‌های آماری و به کارگیری توابع و مدل‌های ریاضی و چه از جهت مفروضه‌های نظری و نتایج کاربردی، تفاوت‌های چشمگیری با نظریه کلاسیک اندازه‌گیری دارد. این نظریه در دهه ۱۹۵۰ به‌عنوان یک جایگزین برای نظریه کلاسیک اندازه‌گیری معرفی شد.

---

1. Thorndike
2. classical test theory (CTT)
3. Spearman
4. theory of mental tests
5. Gulliksen
6. statistical theories of mental test scores
7. Lord & Novick
8. Hambleton & Vander Linden
9. Ragerz
10. Rasch
11. Wright

Archive of SID

www.SID.ir

نظریه سؤال- پاسخ، یک نظریه جامع آماری درباره عملکرد سؤال آزمون و آزمودنی و چگونگی سنجش توانایی‌هایی است. مقیاس سؤال – پاسخ‌ها می‌تواند گسسته یا پیوسته باشد؛ می تواند دوارزشی و یا چندارزشی نمره‌گذاری شود؛ طبقات نمره سؤال می‌تواند منظم و یا نامنظم باشد؛ یک توانایی یا چند توانایی می‌تواند در آزمون مستتر باشد؛ هم‌چنین در ارتباط بین سؤال – پاسخ‌ها و توانایی یا توانایی‌هایی که می‌تواند مشخص شود، چند روش (یا مدل) وجود دارد (همبلتون و جونز[1]، ۱۹۹۳).

این نظریه بر نظریه یا الگوی صفت مکنون استوار است. نظریه سؤال- پاسخ با استفاده از مدل‌های ریاضی پیچیده‌تر از آنچه در نظریه کلاسیک اندازه‌گیری به‌کار می‌رود، یک تابع ریاضی به‌دست می‌دهد که با استفاده از آن می‌توان احتمال پاسخ درست به یک سؤال آزمون را به‌عنوان تابعی از «توانایی[2]» آزمودنی و هم‌چنین برخی ویژگی‌های سؤال معرفی کرد. به سخن دیگر، در نظریه سؤال – پاسخ، فرض بر این است که مثلاً احتمال پاسخ درست دادن به یک سؤال جبر با افزایش دانش جبر آزمودنی افزایش می‌یابد و این، بالقوه یک بیان منطقی است (سیف، ۱۳۸۰).

مدل‌های نظریه سؤال ـ پاسخ را می‌توان به دو نوع پارامتریک و ناپارامتریک تقسیم کرد. مدل‌های ناپارامتریک سؤال ـ پاسخ، اجازه می‌دهد که پارامترهای هر سؤال، مثل ضریب دشواری و ضریب تمیز سؤال را با رتبه‌بندی پاسخ‌دهندگان بر اساس نمره آن‌ها (تعداد پاسخ‌های درست به‌علاوه خطای تصادفی) که روی θ مرتب شده‌است، برآورد کنیم.

موفقیت در کاربرد مدل‌های غیرپارامتریک سؤال- پاسخ (IRT) بر فرض های معینی استواراست که به ماهیت پاسخ سؤالات بر می گردد. به‌عبارت دیگر، مدل غیرپارامتریک سؤال- پاسخ بر پایه فرض‌هایی بنا شده است.این مدل مبتنی بر سه فرض اساسی است ؛اولین فرض با ساختار ابعاد داده‌های آزمون؛ دومی با شکل ریاضی تابع ویژگی سؤال یا منحنی ویژه سؤال، که با ICC نشان داده می‌شود (همبلتون و جونز، ۱۹۹۳) و سومی با یکنواختی توابع سؤال – پاسخ‌ها ارتباط دارد.

الف – تک‌بعدی بودن

اولین فرض، تک‌بعدی بودن است. یعنی، همه سؤالات آزمون یک صفت مکنون مشابه را اندازه می‌گیرند (سیجت‌سما و مولنار۲۰۰۲). به تعبیر دیگر، تک‌بعدی بودن آزمون ـ یعنی احتمال عملکرد موفقیت‌آمیز آزمودنی در مجموعه‌ای از سؤالات ـ را می‌توان به‌صورت یک مدل ریاضی که فقط یک پارامتر توانایی دارد، ارائه کرد (درانز و کینگستن[3]، ۱۹۸۵). البته همبلتون (۱۹۸۹)

1. Hambleton & Jones 2. ability 3. Dorans & Kingston

Archive of SID

www.SID.ir

می‌گوید که فرض تک‌بعدی به‌صورت کامل صادق نخواهد بود، زیرا همیشه یک‌سری عوامل شناختی، شخصیتی، اجرایی و ... وجود دارد که – حداقل تا اندازه‌ای – عملکرد در آزمون را تحت‌تأثیر قرار می‌دهد. همبلتون وکوک (۱۹۷۷) به‌نقل از لرد در ۱۹٦۸ می‌نویسند که فرض تک‌بعدی بودن در مورد مجموعه سؤالات آزمون‌های چندگزینه‌ای، به‌طور کامل برای بسیاری از آزمون‌ها مصداق ندارد؛ هر چند لرد اضافه می‌کند که در بعضی موارد با تقریب خوبی قابل قبول است.

ب– استقلال موضعی

فرض استقلال موضعی به این معنی است که پاسخ فرد به یک سؤال، تحت‌تأثیر پاسخ‌های او به سؤالات دیگر آزمون نیست. برای مثال، استقلال موضعی ممکن است با یادگیری ای که بر اثر تمرین به‌وجود آمده نقض شود (ذوالفقار نسب،۱۳۸۵). این امر در طول اجرای آزمون، هنگامی که نمره‌گذاری روی صفت مکنون صورت می‌گیرد، رخ می‌دهد. در صورتی‌که احتمال پاسخ به سؤال معینی برای آزمودنی A، ۰/۲ و برای آزمودنی B، ۰/۹ باشد و چنانچه پاسخ‌های آزمودنی‌ها به‌صورت موضعی مستقل از یکدیگر باشد، احتمال این‌که هر دوی آنها پاسخ درست بدهند، مساوی با ۰/۱۸= (۰/۹)(۰/۲) است (آلن و ین، ۱۹۷۹ / ترجمه علی دلاور، ۱۳۷٤).

ج– یکنواختی توابع سؤال پاسخ‌ها

فرض بعدی این است که احتمال شرطی $p_i(\theta)$ به‌طور یکنواخت روی θ بدون کاهش است. این فرض در معادله۱ نشان داده شده است. به‌روشنی می‌توان دید وقتی هم که احتمال $X_i = 0$ است، با تابع سؤال ـ پاسخ قابل توصیف است.

$$p(X_i = 0|\theta) = 1 - p(X_i = 1|\theta)$$ معادله (۱)

مدل‌های غیرپارامتریک سؤال ـ پاسخ، دو نوع است: (الف) مدل همگنی یکنواختی[1] و (ب) مدل همگنی یکنواختی جفتی[2]. مدل همگنی یکنواختی بر مبنای فرض تک‌بعدی بودن، استقلال موضعی و یکنواختی قرار گرفته است. این مدل به‌صورت همگون ایجاد شده‌است. توابع سؤال ـ پاسخ آن به‌صورت یکنواخت به صفت مکنون ارتباط دارد. اهمیت کاربردی مدل همگنی یکنواختی این است که می‌توان پاسخ‌دهندگان را با نمره‌کل، روی مقیاس θ رتبه‌بندی کرد. بنابراین، مدلی است که در آن، افراد براساس یک مقیاس ترتیبی مورد اندازه‌گیری قرار می‌گیرند. اگر c را مقداری ثابت درنظر بگیریم و این مقدار ثابت را با s و t جمع کنیم، به شرط برقراری رابطه $o \le s \le t \le k$ خواهیم داشت:

1. monotone homogeneity model
2. double homogeneity model

Archive of SID

www.SID.ir

$$p(\theta > c|X_+ = s) \le p(\theta > c|X_+ = t)$$ معادله (۲)

معادله (۲) به این معنی است که افراد را می‌توان با $\mathbf{X}_+$ به‌صورت احتمالی روی $\boldsymbol{\theta}$ رتبه‌بندی کرد.

دومین مدل غیرپارامتریک سؤال- پاسخ برای سؤالات دوارزشی، مدل همگنی یکنواختی جفتی (DMM) است. این مدل در سه فرض با اولین مدل مشترک است. برای فهم بهتر این مطلب، اول به سؤال‌های دوارزشی می‌پردازیم که به‌صورت o و ۱ نمره‌گذاری می‌شوند. نمره شرطی موردانتظار هر سؤال برابر است با مقادیر تابع سؤال ـ پاسخ آن. یعنی:

معادله (۳)

$$E(X_i|\theta) = 0 \times P(X_i = 0|\theta) + 1 \times P(X_i = 1|\theta) = P_i(\theta)$$

به‌عبارت دیگر، تغییرناپذیری ترتیب بر این دلالت دارد که احتمال پاسخگویی به سؤال i کوچک‌تر مساوی سؤال j است: $P_j \le P_i$

برای تحلیل داده‌ها در مدل غیرپارامتریک سؤال– پاسخ، دو بررسی صورت می‌گیرد.

الف – بررسی پذیرش سؤال (P_i)

در نظریه سؤال ـ پاسخ، عموماً و مدل غیرپارامتریک آن خصوصاً، ویژگی‌های سؤالات، مستقل از ویژگی‌های آزمودنی‌های نمونه معین و درواقع بر اساس جامعه آزمودنی‌ها برآورد می‌شود (سیجت‌سما و مولنار۲۰۰۲). این پارامتر نشان می‌دهد که منحنی ویژگی سؤال در کجای مقیاس توانایی قرار دارد و معادل دشواری سؤال در نظریه سؤال ـ پاسخ در رابطه با سطح صفت زیربنایی θ است (مولنار، ۱۹۹۷).

ب– بررسی مقیاس‌پذیری تک‌تک سؤال‌ها

این پارامتر که با نماد H_i نشان داده می‌شود، قدرت سؤال را در تمایزگذاری یا تشخیص آزمودنی‌ها در سطوح مختلف توانایی نشان می‌دهد و تعیین می‌کند که هر سؤال آزمون تا چه حد می تواند آزمودنی‌ها را در سطوح مختلف توانایی تفکیک کند. در واقع، این پارامتر نشان‌دهنده همسویی و هماهنگی سؤال با کل آزمون است و یا بیانگر آن است که آیا سؤال همان ویژگی موردنظر آزمون را اندازه‌گیری می‌کند یا نه؟ هر چه قدر قدرت تشخیص سؤال بیش تر باشد، نشان‌دهنده این است که آزمودنی‌هایی با توانایی پایین، که در کل آزمون عملکرد پایینی داشته‌اند، به احتمال زیاد به سؤال موردنظر غلط پاسخ می دهند و آزمودنی‌هایی با سطوح بالای توانایی، که در کل آزمون عملکرد بالایی داشته‌اند، به احتمال زیاد سؤال مزبور را به‌صورت صحیح پاسخ

Archive of SID

www.SID.ir

می‌دهند. لرد (۱۹۸۰) بیان می‌کند که میزان این پارامتر نشان می‌دهد که هر چه به سمت بالای خصیصه مکنون مورد سنجش برویم، احتمال پاسخ صحیح به سؤال افزایش می‌یابد. چنانچه نمره فرد (a) در یک آزمون معین را با X و پاسخ وی به هر سؤال را با u_i نشان دهیم ، پاسخ دو ارزشی و شامل صفر و یک است و نمره هر فرد در یک آزمون n سؤالی: $T_a = \varepsilon(X_a) = \sum_{i=1}^{n} \varepsilon(u_{ia})$. در صورتی‌که فرد در سطح توانایی مشخص باشد، عبارت فوق به‌صورت یک عبارت شرطی روی θ تعریف می‌شود:

$$T_A = \sum_{I=1}^{N} \varepsilon\left(U_{IA} \middle| \theta\right)$$

از طرف دیگر، طبق آزمایش برنولی[1] برای متغیرهای دوجمله‌ای، احتمال شرطی وقوع یک حادثه به‌صورت زیر است:

$$\begin{aligned}\varepsilon\left(U_{IA} \middle| \theta\right) &= 1.P_{IA}\left(U_{IA} = 1 \middle| \theta\right) + 0.P_{IA}\left(U_{IA} = 0 \middle| \theta\right) \\ &= P_{IA}\left(U_{IA} = 1 \middle| \theta\right) \\ &= P_{IA}\left(\theta\right)\end{aligned}$$

(لرد و ناویک، ۱۹٦۸).

احتمال بروز یک پیشامد ناسازگار عبارت است از مجموع احتمال‌های هر دو آن ها که مساوی یک خواهد بود. از طرفی، در توزیع دو جمله‌ای که در آن با متغیر دوارزشی گسسته سروکار داریم، در صورت بروز یک پیشامد، پیشامد دیگر غیرممکن خواهد بود. بنابراین، در واقع آنچه باقی می‌ماند، احتمال پیشامد موردنظر است. در این جا منظور، احتمال پاسخ درست فرد a با سطح توانایی θ به سؤال i (یعنی $P_{ia}(\theta)$) است. با توجه به مطالب بیان شده، پس:

$$\mathbf{T_A = \sum_{I=1}^{N} P_{IA}(\theta)}$$

یعنی، نمره حقیقی آزمودنی a در یک آزمون n سؤالی، برابر با مجموع احتمال‌های شرطی پاسخ‌های او به سؤالات آزمون مذکور خواهد بود. در صورتی که احتمال شرطی برای همه سؤالات و آزمودنی‌ها محاسبه و جمع شود، نمره حقیقی آزمون به‌دست می‌آید؛ در معادله بالا، تأثیر تعداد سؤالات آزمون بر نمره حقیقی آزمون و آزمودنی به‌وضوح مشاهده می‌شود. هر قدر تعداد سؤالات آزمون بیشتر شود، تعداد $p_{ia}(\theta)$ ها افزایش می‌یابد و به‌تبع آن T هم افزایش خواهد

1 - Bernoly

Archive of SID

www.SID.ir

داشت. از طرف دیگر، احتمال پاسخگویی صحیح به سؤالات دشوار کم تر از سؤالات ساده است؛ بنابرین در صورت نامساوی بودن درجه دشواری سؤالات دو آزمون ناموازی، برآورد نمره حقیقی یک آزمودنی در آن آزمون متفاوت خواهد بود و به‌عبارت دیگر، برآورد نمره حقیقی آزمودنی، صرف‌نظر از عوامل دیگر، وابسته به آزمون، نوسانات نمونه‌برداری و تعداد سؤالات موجود در آزمون خواهد بود (از معایب نظریه کلاسیک اندازه‌گیری که قبلا به آن اشاره شد). البته همان‌طور که لرد و ناویک (۱۹۶۸) فرمولی را ارائه نموده‌اند، با در دست داشتن شاخص‌های کلاسیک سؤال و فرد می‌توان نمره واقعی را به‌طور تقریبی به‌دست آورد.

$$R\left(T_a \middle| X_a\right) = r_{xx} X_a + \left(1 - r_{xx}\right)\overline{X}$$

برآورد رگرسیون نمره حقیقی آزمودنی a از روی نمره مشاهده شده وی، نیازمند داشتن برآوردی از اعتبار آزمون، نمره مشاهده شده آزمودنی a و میانگین آزمون است. نتیجه، برآوردی از نمره حقیقی فرد a از روی نمره مشاهده شده او خواهد بود. نکته دیگر آن‌که، اگر احتمال‌های شرطی معادله فوق معدل‌گیری شود و میانگین پاسخ صحیح در هر سطح برای سؤال‌ها به‌دست آید، از طریق آن می‌توان منحنی ویژه آزمون (TCC) را رسم کرد.

$$T = \overline{P}(\theta) = \frac{\sum_{i=1} p_{i(\theta)}}{n}$$

احتمال‌های شرطی پاسخ همه آزمودنی‌های سطح معینی از θ به سؤالات آزمون، جمع شده، بر تعداد سؤالات تقسیم می‌شود. نتیجه این محاسبه، متوسط احتمال پاسخ صحیح به سؤالات آزمون در آن سطح توانایی خواهد بود. با ادامه محاسبه فوق برای θهای مختلف می‌توان (TCC) را رسم کرد که در آن، محور افقی شامل توانایی یا θ و محور عمودی یا y شامل متوسط احتمال محاسبه شده یا $p(\theta)$ است.

با گذشت بیش از ۶۰ سال از کار نظریه سؤال ـ پاسخ و گسترش پرشتاب مبانی نظری آن در تمام سال‌های دهه ۱۹۸۰ و اوایل سال‌های دهه ۱۹۹۰، متأسفانه در کشور ما این نظریه هنوز آن‌چنان که باید و شاید شناخته شده نیست و تعداد تحقیقات در این زمینه بسیار اندک است و هنوز نظریه سؤال ـ پاسخ به‌طور اعم و مدل ناپارامتریک سؤال ـ پاسخ به‌طور اخص، در اندازه‌گیری توانایی‌های داوطلبان در پاسخگویی به سؤالات آزمون‌های ورود به مراکز آموزش عالی و دیگر آزمون‌ها راه نیافته است؛ در حالی‌که الزامات جامعه ما، به‌ویژه در سال‌های اخیر و با توجه به خیل عظیم داوطلبان دانشگاه‌ها و مراکز و موسسات آموزش عالی ایجاب می‌کند که افراد، هرچه دقیق‌تر و درست‌تر، برمبنای توانایی‌های ذهنی خود از یکدیگر متمایز و برای تحصیل در

Archive of SID

www.SID.ir

دانشگاه‌ها و مؤسسات آموزش عالی انتخاب شوند. با توجه به مطالب فوق، این پرسش مطرح است که: مدل اندازه‌گیری کلاسیک و مدل غیرپارامتریک سؤال- پاسخ از نظر ویژگی‌های سؤال، چه تفاوتی با یکدیگر دارند؟

به‌منظور پاسخگویی به پرسش یاد شده، با استفاده از داده‌های حاصل از اجرای این پژوهش، پرسش‌های زیر موردنظر قرار گرفت:

۱ـ نظریه کلاسیک اندازه‌گیری و مدل ناپارامتریک سؤال ـ پاسخ، با توجه به عملکرد افراد در آزمون ریاضی، از نظر برآورد شاخص یا پارامتر سطح دشواری سؤالات آزمون، چه تفاوتی دارند؟

۲ـ نظریه کلاسیک اندازه‌گیری و مدل ناپارامتریک سؤال ـ پاسخ، با توجه به عملکرد افراد در آزمون ریاضی، از نظر برآورد شاخص یا پارامتر قدرت تشخیص سؤالات آزمون، چه تفاوتی دارند؟

روش

جامعه

به‌منظور پاسخ به پرسش‌های پژوهشی موردنظر، جامعه ای از کلیه داوطلبان ورود به دانشگاه‌های کشور در رشته ریاضی ـ فیزیک در سال ۱۳۸٤انتخاب شد. جدول ۱، تعداد داوطلبان شرکت کننده در کنکور سراسری ۱۳۸٤ را در رشته ریاضی- فیزیک نشان می‌دهد.

جدول ۱. تعداد داوطلبان شرکت‌کننده در کنکور سراسری ۱۳۸۴ در رشته ریاضی ـ فیزیک

رشته	زن	درصد	مرد	درصد	کل
ریاضی- فیزیک	۱۳۸۸٦۸	٤٥/۹۱	۱٦۳٦٤۳	٥٤/۰۹	۳۰۲٥۱۱

نمونه‌گیری بر اساس دسترسی به فهرست تصادفی داوطلبان و با توجه به برنامه کامپیوتری طراحی شده در سازمان سنجش آموزش کشور، با استفاده از روش نمونه‌گیری منظم انجام شد. از بین کلیه داوطلبان شرکت‌کننده در گروه آزمایشی ریاضی ـ فیزیک، یک گروه ۳۰۰۰ نفری به‌طور تصادفی انتخاب شد و پس از حذف آزمودنی‌هایی که به‌درستی به سؤالات پاسخ نگفته بودند (منظور، افرادی هستند که پایین‌تر از $\frac{1}{4}$ نمره کل یا حد شانس را کسب کرده بودند)، از بین افراد باقی مانده چند نمونه‌گیری صورت گرفت.

برطبق نظر طرفداران نظریه کلاسیک، یکی از مزایای این نظریه آن است که حجم نمونه در آن نسبت به نظریه مدل پارامتریک سؤال- پاسخ و مدل غیر پارامتریک سؤال – پاسخ، کم تر

Archive of SID

www.SID.ir

است. مطالعات اندکی وجود دارد که تأثیر حجم نمونه را بر برآوردهایی که برای پایایی پارامترها صورت می‌گیرد، به‌طور منظم بررسی کرده باشد. حداقل حجم نمونه پیشنهادی برای به‌کارگیری موثر CTT در دامنه‌ای از حدود ۳۰۰ تا ۵۰۰ متغیر است. البته ترجیح داده می‌شود که حجم نمونه ۱۰۰۰ باشد (تروسکوسکی، ۱۹۹۹ به‌نقل از نانالی، ۱۹۶۷). مطابق پژوهش‌های موجود در زمینه آزمون‌سازی، کاربرد موفق مدل‌های پارامتریک سؤال – پاسخ و مدل های غیر پارامتریک سؤال – پاسخ مستلزم استفاده از سؤالات و آزمودنی‌هایی با حجم بزرگ است تا بتوان به‌طور همزمان صفت مکنون و پارامترهای سؤال را برآورد کرد (لرد، ۱۹۶۸ و همبلتون و سوآمیناتان، ۱۹۸۵). بنابراین در پژوهش حاضر، بر اساس تحقیقات انجام شده در هر دو نظریه و با توجه به نرم‌افزار کامپیوتری به‌کار رفته، از بین داوطلبان گروه آزمایشی ریاضی- فیزیک و از بین گروه‌های نمونه مختلف، در نهایت دو نمونه ۱۰۰۰ نفری، که در آن‌ها تفاوت توانایی آزمودنی‌ها بیش تر بود، انتخاب شد.

ابزار و روش اجرا و تحلیل داده‌ها

در آزمون ورودی دانشگاه‌های ایران از آزمون‌های پیشرفت تحصیلی استفاده می‌شود. این آزمون‌ها در سازمان سنجش و توسط استادان و طراحان زبده و کارکشته کشور طراحی و تهیه می‌شود. ابزار مورد استفاده برای جمع‌آوری داده‌های پژوهش، پاسخنامه‌های ۵۵ سؤالی داوطلبان در آزمون اختصاصی درس ریاضی بود.

در مورد تعداد سؤالات در مقایسه مدل‌ها، صاحب‌نظران، حداقل حجم نمونه سؤال را ۱۵ ذکر کرده‌اند. لیکن مانند آزمودنی‌ها، تعداد سؤالات خرده‌آزمون‌ها نیز از تعداد پارامترهای مدل تأثیر می‌پذیرد. (همبلتون، ۱۹۸۹).

برای تجزیه و تحلیل داده‌ها از نرم‌افزارهای کامپیوتری SPSS و MSP استفاده شد. ابتدا تحلیل‌ها و مشخصه‌های کلاسیک سؤال‌ها و آزمون‌ها از طریق SPSS محاسبه شد و سپس داده‌ها با نرم‌افزار MSP تحلیل شد. روش‌ها و شاخص‌های مورد استفاده برای تحلیل سؤالات بر پایه مدل کلاسیک اندازه‌گیری، شامل میانگین یا درجه دشواری سؤالات، واریانس سؤالات، ضریب همبستگی دو رشته‌ای (r_{bis}) و ضریب همبستگی دو رشته‌ای نقطه‌ای (r_{pbis}) بود. در این مدل نیز برای آزمون‌ها از پایایی به روش ضریب آلفا (کودر- ریچاردسون ۲۰)، توزیع فراوانی و نمودار نمرات آزمون‌ها و ... استفاده شد.

به منظور تحلیل داده‌ها در نظریه مدل غیر پارامتریک سؤال- پاسخ، از روش‌های آماری مانند آزمون t وابسته، ضریب همبستگی پیرسون و آزمون‌های معنی‌داری آن ها استفاده شد و سپس مقایسه‌های موردنظر انجام گرفت.

Archive of SID

www.SID.ir

نتایج

با استفاده از آزمون اختصاصی ریاضی رشته ریاضی- فیزیک، سؤال اول تحقیق، یعنی تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سؤال، بررسی شد. در تحلیل سؤالات آزمون ریاضی به روش کلاسیک، ٦ سؤال که فاقد برازندگی بود، حذف شد و ٤٩ سؤال باقی ماند. بار دیگر با مدل غیرپارامتریک سؤال- پاسخ این سؤالات تحلیل شدند. به‌دنبال آن، با استفاده از ضریب مقیاس پذیری، آزمون برازندگی برای هر سؤال و کل آزمون محاسبه شد.سؤال هایی که ضریب مقیاس‌پذیری کم تر از ۰/۳ دارند، در حدود ۲۳ سؤال بودند. با توجه تعداد زیاد سؤال‌های حذف شدنی، آنها را ۵ تا ۵ تا کنار گذاشتیم. در مجموع، ۱۱ سؤال کنار گذاشته شد و ٤٤ سؤال برای تجزیه و تحلیل باقی ماند.

هم‌چنین همان‌طور که جدول ۲ نشان می‌دهد، ضرایب همبستگی برآورد شاخص‌های دشواری، قدرت تشخیص در نظریه کلاسیک (CTT) مقادیر بسیار پایینی را نشان می‌دهد؛ در حالی‌که در مدل غیرپارامتریک سؤال- پاسخ (NIRT) همبستگی بسیار بالایی بین مقادیر برآورد شده این دو پارامتر در دو گروه نمونه وجود دارد. این نکته بیانگر آن است که برآورد این شاخص‌ها در نظریه کلاسیک، یک برآورد متغیر و غیرثابت است؛ ولی در NIRT چنین نیست. لازم به ذکر است که در نرم‌افزار MSP، که برای تحلیل مدل غبرپارامتریک به‌کار می‌رود، محبوبیت سؤال (P_i) معادل ضریب دشواری و مقیاس‌پذیری (H_i) معادل ضریب تشخیص در نظر گرفته می‌شود.

جدول ۲. مقایسه ضرایب همبستگی شاخص‌های سؤال در CTT و NIRT

مورد مقایسه نظریه	تعداد سؤالات آزمون ریاضی دارای برازش	ضریب‌همبستگی برآورد شاخص یا پارامتر دشواری	سطح معنی‌داری	ضریب همبستگی برآورد شاخص یا پارامتر قدرت تشخیص	سطح معنی‌داری
CTT	٤٩	۰/۳۳۵	۰/۰۱۹	۰/۲۹۹	۰/۰۳۹
NIRT	٤٤	۰/۹۸۹	۰/۰۰۰	۰/۹٤۱	۰/۰۰۰

Archive of SID

www.SID.ir

مقایسه ضریب همبستگی: به‌منظور بررسی ثبات شاخص یا محبوبیت سؤال از ضریب همبستگی نقطه‌ای یا به‌عبارتی ضریب همبستگی پیرسون استفاده شد. نتایج مقایسه ضرایب همبستگی در این دو نظریه، در جدول ۳ آمده است.

جدول ۳. مقایسه ضرایب همبستگی شاخص یا محبوبیت سؤال (دشواری سؤال) در دو گروه نمونه با روش CTT و NIRT

روش مقایسه	تعداد سؤالات	ضریب همبستگی r_{pbis}	سطح معنی‌داری
CTT	۴۸	۰/۳۳۵	۰/۰۱۹
NIRT	۴۵	۰/۹۸۹	۰/۰۰۰

بررسی جدول فوق، نشان می‌دهد که همبستگی پارامتر محبوبیت سؤال (دشواری سؤال) در نظریه NIRT بسیار بالا بوده (r_{pbis}= ۰/۹۸۹) و تفاوت فاحشی بین این ضریب همبستگی در NIRT و CTT وجود دارد که بیانگر یکسان بودن مقادیر برآورده شده محبوبیت سؤال در NIRT است.

از آزمون t وابسته نیز برای نشان دادن تفاوت دو نظریه در برآورد شاخص یا محبوبیت سؤال استفاده شد. جدول ۳ مقایسه این دو نظریه را نشان می‌دهد.

جدول۴. مقایسه دو نظریه CTT و NIRT از نظر شاخص یا محبوبیت سؤالات با استفاده از آزمون t در دو گروه نمونه

روش مورد مقایسه	تعداد سؤالات	میانگین سطح دشواری		خطای استاندارد میانگین سطح دشواری		آزمون t	درجه آزادی (d.f)	سطح معنی‌دار
		گروه A	گروه B	گروه A	گروه B			
CTT	۴۹	۰/۳۸۰۲ **sd**=۰/۲۸	۰/۱۱۰۵ **sd**=۰/۱۵	۰/۰۳۷	۰/۰۱۷	۶/۳۴	۴۸	۰/۰۰۱
NIRT	۴۴	۱/۱۰۹۸ **sd**=۱/۱۳	۱/۱۰۴۵ **sd**=۱/۱۴	۰/۱۷	۰/۱۷	۰/۷۹	۴۳	/۳۸۸

Archive of SID

www.SID.ir

همان‌طور که ملاحظه می‌شود، میانگین سطح دشواری در دو گروه CTT با یکدیگر متفاوت بود؛ در حالی‌که این میزان در NIRT تقریباً یکسان است. خطای استاندارد میانگین سطح دشواری نیز در دو گروه در CTT متفاوت بوده است؛ در حالی‌که این میزان در NIRT بسیار نزدیک به یکدیگر است. آزمون t در CTT نشان می‌دهد که تفاوت بین دو برآورد سطح دشواری در دو گروه آزمودنی معنی‌دار است؛ یعنی مقادیر برآورد شده در دو گروه با یکدیگر تفاوت معنی‌دار دارند؛ در حالی‌که در NIRT آزمون t وابسته نشان می‌دهد که تفاوت بین دو گروه از نظر برآورد پارامتر دشواری، معنی‌دار نیست و مقادیر برآورد شده پارامتر دشواری در دو گروه متفاوت از آزمودنی‌ها در NIRT یکسان یا تقریباً یکسان است. این مسئله بیانگر تأثیرناپذیری برآورد محبوبیت سؤال از ویژگی‌های آزمودنی‌ها در نظریه غیرپارامتریک سؤال ـ پاسخ است.

از آن‌جا که در نظریه کلاسیک، شاخص قدرت تشخیص سؤال، ضریب همبستگی نقطه‌ای است، در NIRT مقیاس پذیری، معادل برآورد شاخص یا پارامتر قدرت تشخیص یا شیب سؤال (a_i) است. جدول ۵، ضریب همبستگی آزمون ریاضی را در برآورد شاخص قدرت تشخیص در دو گروه نمونه در CTT و NIRT نشان می‌دهد.

جدول ۵– مقایسه ضرایب همبستگی شاخص یا مقیاس پذیری (پارامتر قدرت تشخیص) سؤال در دو گروه نمونه با روش CTT و NIRT

روش مقایسه	تعداد سؤالات	ضریب همبستگی r_{pbis}	سطح معنی‌داری
CTT	٤٩	۰/۳۱۰	۰/۰۳٤
NIRT	٤٤	۰/۹٤۱	۰/۰۰۱

همان‌طور که ملاحظه می‌شود، ضریب همبستگی در CTT برابر ۰/۳۱۰ و در NIRT برابر ۰/۹٤۱ است. این میزان نشان‌دهنده همبستگی بسیار بالا در نظریه غیرپارمتریک سؤال ـ پاسخ (در مقایسه با نظریه کلاسیک) در برآورد شاخص یا پارامتر قدرت تشخیص است.

Archive of SID

www.SID.ir

جدول ۶ ــ مقایسه دو نظریه CTT و NIRT از نظر شاخص یا مقیاس‌پذیری (پارامتر قدرت تشخیص) سؤال با استفاده از آزمون t

روش مورد مقایسه	تعداد سؤالات	میانگین سطح دشواری		خطای استاندارد میانگین سطح دشواری		آزمون t	درجه آزادی (d.f)	سطح معنی‌دار
		گروه A	گروه B	گروه A	گروه B			
CTT	٤٩	۰/۲۷۷۷ sd=۰/۰۸	۰/۱٤٤۱ sd=۰/۰۵	۰/۰۱٤	۰/۰۰۹	۹/۳۵	٤۸	۰/۰۰۱
NIRT	٤٤	۰/۷۹۳۳ sd=۰/۲۳	۰/۷۹۳۲ sd=۰/۲۲	۰/۰۳۷	۰/۰۳٦	-۰/۱۲	٤۳	۰/۸۹۸

همان‌طور که در جدول ٦ ملاحظه می‌شود، در نظریه کلاسیک، میانگین شاخص قدرت تشخیص در دو گروه تفاوت زیادی داشته است؛ در حالی‌که در نظریه غیرپارامتریک سؤال ـ پاسخ، میانگین قدرت تشخیص و sd آن‌ها در دو گروه آزمودنی بسیار نزدیک به‌یکدیگر است. خطای استاندارد میانگین قدرت تشخیص نیز در CTT در دو گروه نمونه، متفاوت است؛ ولی در NIRT این مقادیر تقریباً یکسان است. آزمون t در CTT در سطح ۰/۰۰۱ معنی‌دار است؛ یعنی برآورد شاخص قدرت تشخیص در دو نمونه در CTT تفاوت معنی‌داری با یکدیگر دارد؛ اما در NIRT تفاوت بین دو گروه در برآورد این شاخص معنی‌دار نیست. به‌عبارتی در نظریه غیرپارامتریک سؤال ـ پاسخ ، ضریب مقیاس پذیری (پارامتر قدرت تشخیص) ثابت است و تحت‌تأثیر ویژگی‌های گروه آزمودنی‌ها قرار نمی‌گیرد و از یک گروه به گروه دیگر، مقادیر آن مشابه است؛ در نتیجه مقادیر برآوردشده در دو گروه، تفاوت معنی‌داری با یکدیگر ندارند؛ اما در نظریه CTT، تفاوت بین مقادیر برآوردشده شاخص‌های سؤال، زیاد و تفاوت بین دو گروه معنی‌دار است.

Archive of SID

www.SID.ir

بحث و نتیجه‌گیری

در بررسی اول، تفاوت دو نظریه از نظر پارامتر سطح دشواری در کلاسیک یا محبوبیت سؤال نظریه مدل غیرپارامتریک سؤال – پاسخ، مورد مطالعه قرار گرفت و نتایج زیر به دست آمد: در مدل کلاسیک، درجه دشواری سؤال (p_i) با انتخاب نمونه‌ای از آزمودنی‌ها تغییر می‌کند؛ یعنی یک سؤال برای گروه نمونه قوی‌تر،آسان و برای گروه نمونه ضعیف‌تر،دشوار خواهد بود؛ در حالی‌که بررسی‌ها نشان داد که در نظریه مدل غیرپارمتریک سؤال ـ پاسخ، ضریب مقیاس‌پذیری (H_i)، نامتغیر و تقریباً ثابت است و می‌توان نحوه پاسخگویی افرادی را که قبلاً با آن سؤال مواجه نشده‌اند، در این نظریه پیش‌بینی کرد. در ضمن، این شاخص تحت‌تأثیر ویژگی‌های آزمودنی‌ها یا سؤالات دیگر تغییر نمی‌کند.در این زمینه موکن (۱۹۹۷) به نتایج مشابه دست یافت. بر اساس میزان دشواری آزمون در نظریه کلاسیک، نسبت پاسخ‌های صحیح و نمرات افراد گروه نمونه تغییر می‌کند و برحسب میزان دشواری آزمون، توانایی افراد، بیشینه یا کمینه برآورد می‌شود؛ ولی در نظریه مدل غیرپارمتریک سؤال ـ پاسخ، ضریب مقیاس‌پذیری، بدون تأثیر از میزان توانایی افراد برآورد می‌شود و ویژگی‌های سؤالات آزمون، تأثیری بر مقدار برآورد شده ندارد. رمزی (۱۹۹۷) تحلیل‌هایی را با استفاده از مدل‌های همگنی یکنواختی و همگنی یکنواختی جفتی انجام داد و بررسی تحقیقی وی نیز مطالب فوق را تأیید کرد.

در بررسی سؤال دوم تحقیق، یعنی تفاوت دو نظریه از نظر شاخص قدرت تشخیص سؤال در کلاسیک و ضریب مقیاس‌پذیری، نتایج زیر به‌دست آمد: شاخص قدرت تمیز در نظریه کلاسیک اندازه‌گیری، مانند شاخص دشواری سؤال، از نمونه‌های مختلف آزمودنی تأثیرات متفاوتی را می‌پذیرد. این نظریه برآوردهایی وابسته به نمونه ارائه می‌کند و تحت‌تأثیر ناهمگونی گروه، این مشخصه افزایش می‌یابد؛ در حالی‌که در نظریه غیرپارامتریک سؤال ـ پاسخ، ضریب مقیاس‌پذیری، شاخصی است که میزان اطلاعی را که یک سؤال درباره سطح توانایی مورد سنجش ارائه می‌کند، نشان می‌دهد. ضریب مقیاس‌پذیری در نظریه NIRT، نامتغیر است و به گروه‌های مختلف آزمودنی‌ها قابل تعمیم است. تغییرناپذیری یکی از مهم‌ترین ویژگی‌های این نظریه است (مولنارو ۱۹۹۷ و ترسکوسکی ۱۹۹۹).

Archive of SID

www.SID.ir

منابع

آلن، مری‌جی؛ ین، وندی‌ام (۱۳۷٤). مقدمه‌ای بر نظریه‌های اندازه‌گیری (روان‌سنجی)، ترجمه علی دلاور، تهران: انتشارات سمت (۱۹۷۹).

افروز، غلامعلی؛ هومن، حیدرعلی (۱۳۷۵). روش تهیه آزمون هوش: هوش‌آزمای تهران ـ استنفورد ـ بینه (T.S.B)، تهران: موسسه انتشارات و چاپ دانشگاه تهران.

ثرندایک، رابرت (۱۳۷۵). روان‌سنجی کاربردی، ترجمه حیدرعلی هومن. تهران: انتشارات دانشگاه تهران.

سیف، علی‌اکبر (۱۳۸۰). روش‌های اندازه‌گیری و ارزشیابی آموزشی. نشر دوران.

مولن آیر، ایودبلیو؛ سیجت‌سما، کلاس (۱۳۸۵). مقدمه بر تئوری ناپارامتریک سؤال – پاسخ (به‌همراه نرم‌افزار)، ترجمه سلیمان ذوالفقاری‌نسب: مشهد: انتشارات کتابخانه‌ای رایانه‌ای (۲۰۰۲).

Gulliksen, H. (1950). Theory of Mental tests. Newyork: John Wiley & Sons.

Hambleton, R. K. (1989). Principles and selected applications of Item Response Theory. In R. Linn (Ed), Educational Measurement (3_{rd} end). Newyork: Memillan. PP: 147-200.

Hambleton, R. K. & Cook, L. L. (1977). Latent Trait models and their use in the analysis of Educational test data. Journal of Educational Measurement. 14(2), P: 75-94.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications, Boston: Kluwer.

Hambleton, R.K ; Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement: Issues and Practice. 12(3), 38-47.

Hambleton, R. K. & Vander Linden, Wim. J. (1982). Advance in Item Response Theory and Applications: An Introduction, Applied Psychplogical Measurement. 6(4), 373- 378.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of mental test scores.Reading, MA: Addison-Wesley.

Lord, F. M. (1980). Applications of Item Response Theory to Practice Testing Problems, Hillsdale, N.J: Lawrence Erlbaum.

Archive of SID

www.SID.ir

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. In: Hambleton, R.K. and Van der Linden, W.J. (Ed's). Handbook of Modern Item Response Theory. New York-Berlin: Springer-Verlag, pp. 351-367.

Molenaar, I. W. (1997). Nonparametric models for dichotomous responses. In: Hambleton, R.K. and Van der Linden, W.J. (ed's). Handbook of Modern Item Response Theory. New York-Berlin: Springer-Verlag, pp. 369-379

Ramsey, J . (1997). Nonparametric Models for Dichotomous Responses. In: Hambleton, R.K. and Van der Linden, W.J. (ed's). Handbook of Modern Item Response Theory. New York-Berlin: Springer-Verlag, pp. 369-379

Truskosky, D.M. (1999). An empirical examination of Classical Test Theory and Item Response Theory parameters: implications for research and practice in small- and large- sample assessment, Department of Psychology in the graduate school Southern Illinois university at Carbondale.

Wright, B.D. & Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.

Archive of SID

www.SID.ir