مقایسه دو روش اندازه‌گیری کلاسیک و سوال- پاسخ از نظر تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سوال و بالعکس

سیده طیبه مطیعی لنگرودی [1]
رضا پیروی [2]
ضیاء تاج‌الدین [3]
علی مقدم زاده [4]

چکیده

پژوهش حاضر به‌منظور مقایسه دو نظریه کلاسیک اندازه‌گیری (CTT) و نظریه سوال ـ پاسخ (IRT) از نظر تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سوال و بالعکس انجام شده است. روش تحقیق از نوع کاربردی ـ توصیفی بود. از روش کتابخانه‌ای به‌منظور بررسی جنبه‌های نظری و از روش توصیفی به‌منظور بررسی جنبه‌های عملی استفاده شد. در بررسی عملی، پاسخنامه‌های داوطلبان ورود به دانشگاه‌های کشور در رشته ریاضی- فیزیک در آزمون اختصاصی درس ریاضی و در رشته علوم تجربی در آزمون اختصاصی درس زیست‌شناسی

۱- کارشناس ارشد تحقیقات آموزشی دانشگاه تهران و کارشناس دفتر آزمون سازی و روانسنجی سازمان سنجش آموزش کشور

۲- کارشناس ارشد سنجش و اندازه گیری دانشگاه علامه طباطبایی و کارشناس دفتر آزمون سازی و روانسنجی سازمان سنجش آموزش کشور

۳- استادیار دانشکده ادبیات فارسی و زبان های خارجی، دانشگاه علامه طباطبایی

٤- کارشناس ارشد سنجش و اندازه گیری دانشگاه علامه طباطبایی

مورد استفاده قرار گرفت. از بین کلیه داوطلبان در گروه آزمایشی ریاضی- فیزیک، ابتدا به روش نمونه‌گیری سیستماتیک یک گروه ۳۰۰۰ نفری انتخاب شد و سپس با توجه به نرم‌افزار مورد استفاده و حجم نمونه لازم برای پاسخگویی به سوالات تحقیق در روش کلاسیک و سوال - پاسخ، چندین گروه نمونه انتخاب گردید. از بین داوطلبان گروه علوم تجربی نیز یک گروه ۱۰۰۰ نفری برای بررسی سوال تحقیق انتخاب شدند. دو سوال تحقیقی مورد مطالعه در این مقاله، همه از لحاظ نظری بررسی شد و در هر کدام، مزیت‌ها و برتری‌های نظریه سوال - پاسخ بر نظریه کلاسیک مورد بحث قرار گرفت. از جنبه عملی نیز، سوالات تحقیق بررسی شد. به‌منظور تحلیل داده‌ها از روش‌های آماری مورد استفاده برای تحلیل سوالات بر پایه مدل کلاسیک اندازه‌گیری، شامل میانگین یا درجه دشواری سوالات، واریانس سوالات، ضریب همبستگی دو رشته‌ای (r_b) و ضریب همبستگی دو رشته‌ای نقطه‌ای (r_{pb}) استفاده شد. برای آزمون‌ها در این مدل نیز از پایایی به روش ضریب آلفا (کودر- ریچاردسون ۲۰)، توزیع فراوانی و نمودار نمرات آزمون‌ها استفاده شد. به‌منظور تحلیل داده‌ها در نظریه سوال- پاسخ، از روش‌های آماری مانند آزمون t وابسته، ضریب همبستگی پیرسون، جذر میانگین خطاها[1] (RMSE)، و آزمون‌های معنی‌داری آنها استفاده شد و مقایسه بین دو روش انجام گرفت. نتایج تحقیق نشان داد که ویژگی‌های دو گروه نمونه که از نظر توانایی با یکدیگر تفاوت دارند، بر برآورد شاخص‌های سوال در نظریه CTT تأثیر می‌گذارد ولی بر برآورد پارامترهای سوال در نظریه IRT بی‌تأثیر است. همچنین بررسی‌ها نشان داد که در CTT برآورد پارامتر

1-Root Mean Square Error

توانایی تحت‌تأثیر ویژگی‌های سوالات قرار گرفته و متفاوت است، اما در IRT، ویژگی‌های سوالات در برآورد پارامتر توانایی آزمودنی‌ها بی‌تأثیر بوده و پارامتر توانایی همچون پارامترهای سوال، یک پارامتر نامتغیر و بدون اریب است.

واژگان کلیدی: نظریه کلاسیک اندازه‌گیری؛ نظریه سوال- پاسخ؛ خصوصیات آزمودنی‌ها؛ ویژگی‌های سؤال.

مقدمه

تاریخ اندازه‌گیری‌های روانی و تربیتی در قرن بیستم، به‌قول ثرندایک[1] (۱۹۸۲ /ترجمه هومن، ۱۳۷۵)، در واقع تاریخ کشف و اختراع ابزارها و روش‌های اندازه‌گیری است که به‌طریقی استاندارد و تحت شرایط یکسان، رفتارهایی را که منعکس‌کننده خصیصه‌های افراد است، آشکار کرده و مورد سنجش قرار می‌دهد. سنجش‌های روانی و تربیتی بدون تردید یکی از مهم‌ترین کمک‌های علوم رفتاری به جامعه است که در زمینه‌های اداری، صنعتی، فنی ـ حرفه‌ای و تربیتی، تغییرات اساسی و قابل توجهی به‌وجود آورده و می‌توان گفت که امر آموزش و استخدام را به‌کلی دگرگون ساخته است.

ارائه نظریه‌های نوین اندازه‌گیری در قرن بیستم، به پیشرفت فنون و ابزارهای استاندارد شده‌ای انجامیده است که اندازه‌گیری و تبدیل توانش‌های فردی را به مقیاس‌های قابل قبول برای توصیف، تفسیر و برآورد تفاوت‌های فردی امکان‌پذیر می‌سازد (افروز و هومن، ۱۳۷۵). نظریه‌های اندازه‌گیری و مدل‌های مربوط به آن برای عمل اندازه‌گیری در روانشناسی و تعلیم و

1-Thorndike

تربیت مهم هستند، زیرا چارچوبی را برای موضوعات و معرفی مسایل فنی (مانند خطای اندازه‌گیری) فراهم می‌سازند. این نکته نیز که توانایی‌ها یا صفاتی که درصدد اندازه‌گیری آن هستیم، معمولاً به‌طور مستقیم قابل اندازه‌گیری نیستند، خود دلیل دیگری بر نیاز به نظریه‌های اندازه‌گیری می‌باشد.

اغلب نظریه‌های اندازه‌گیری دارای مدل‌هایی هستند که ساختار منسجمی را برای مفاهیم کلی نظریه ایجاد می‌کنند. هر یک از این مدل‌ها، بخش‌هایی از نظریه را شرح می‌دهند و به‌طور سیستماتیک با پدیده‌های قابل مشاهده سروکار دارند. این مدل‌ها استدلال‌های منطقی را در مورد روابط بین پدیده‌ها ارائه می‌دهند. نظریه‌های علمی حقایق را به‌طور مطلق توصیف و تبیین نمی‌کنند، ولی مدل‌های مربوط به آن‌ها در زمینه‌های مختلف از دقت نسبی برخوردارند. از آن‌جایی‌که مدل‌های مطرح شده در نظریه‌های اندازه‌گیری بر ریاضیات مبتنی هستند، با مدل‌های کلامی تفاوت‌هایی دارند. این مدل‌ها در یک سیستم ریاضی تعریف شده‌اند و از دقت بالایی برخوردارند (لرد و ناویک[1]، ۱۹۶۸). یک نظریه یا مدل اندازه‌گیری خوب، همچنین چارچوب مرجعی را برای طراحی آزمون و یا حل دیگر مشکلات عملی فراهم می‌سازد. یک نظریه خوب می‌تواند رابطه مابین سؤالات آزمون و نمرات توانایی را مشخص سازد، همان‌طور که طراحی دقیق آزمون می‌تواند توزیع نمرات آزمون درخواست شده و حجم خطاهای جایز را نشان دهد (همبلتون، ۱۹۸۹). همبلتون و واندر لیندن[2] (۱۹۸۲) نظریه‌های اندازه‌گیری را به‌دو

1-Lord & Novick

2-Hambleton & Van der Linden

دسته عمده تقسیم کرده‌اند: (الف) نظریه کلاسیک اندازه‌گیری[1] (CTT)، (ب) نظریه‌های جدید اندازه‌گیری یا نظریه سوال ـ پاسخ[2] (IRT).

نظریه کلاسیک اندازه‌گیری که بیشتر از نیم قرن بر نهضت آزمون‌سازی حکومت کرد، دارای مفروضاتی راجع به داده‌های حاصل از آزمون است که بر اساس سه مولفه اصلی: (الف) نمره مشاهده شده[3]، (ب) نمره حقیقی[4]، (ج) نمره خطای تصادفی[5] به تبیین و تفسیر داده‌ها می‌پردازد. مدل کلاسیک اندازه‌گیری مدلی ضعیف[6] است که برای اندازه‌گیری ویژگی‌های رفتاری به‌کار می‌رود و متأسفانه دشواری‌های عملی و مسایل تفسیری و تحلیلی متعددی را موجب می‌شود که قادر به پاسخگویی و حل آن‌ها نمی‌باشد. دو ضعف عمده این نظریه عبارتند از: (۱) وابستگی پارامترهای سوال به نمونه آزمودنی. (۲) وابستگی پارامتر توانایی افراد به نمونه سوالات آزمون.

اما نظریات جدید اندازه‌گیری، چه از لحاظ روش‌های آماری و به‌کارگیری توابع و مدل‌های ریاضی و چه از جهت مفروضات نظری و نتایج کاربردی، تفاوت‌های چشمگیری با نظریه کلاسیک دارد و توانسته است چهارچوب مفیدی برای حل مسایل گسترده اندازه‌گیری خصوصیات روانی تربیتی فراهم آورد. تعداد روزافزونی از موسسات اندازه‌گیری که برای گزینش داوطلبان در مدارس و دانشگاه‌ها و صنایع، آزمون می‌سازند از این نظریه در ساخت

1-Classical Test Theory
2-Item Response Theory
3-Observed Score
4-True score
5-Random Error Score
6-Weak Model

آزمون به‌منظور تعیین سوگیری‌های ضمنی ممکن در سؤالات آزمون، ساختن آزمون‌های مختلف با فرم‌های همتا از یک آزمون، تفسیر دقیق نمرات آزمون و ... استفاده می‌کنند.

تلاش‌های نخستین تکوین نظریه کلاسیک اندازه‌گیری در دهه ۱۸۹۰ آغاز شد. این نظریه یکی از روش‌های دیرینه در ساخت و توسعه آزمون‌ها در حوزه علوم انسانی است که از اوایل دهه ۱۹۰۰ برای توسعه ابزارهای اندازه‌گیری و برای تعیین این‌که آیا آزمون‌ها با نظریه همخوانی دارد و برای نمره‌گذاری امتحانات. استفاده شده است. گرچه نظریه کلاسیک اندازه‌گیری مدت‌های طولانی به جامعه روانسنجی خدمت کرده است، با این‌حال دارای برخی محدودیت‌ها است که به‌وسیله محققین اندازه‌گیری مورد توجه قرار گرفته است (از جمله: گالیکسن[1]، ۱۹۵۰؛ همبلتون و سوامیناتان و راجرز[2]، ۱۹۹۱؛ لرد و ناویک، ۱۹۶۸). از جمله محدودیت‌های روش اندازه‌گیری کلاسیک، می‌توان به محدودیت در روش‌ها اشاره داشت.

علی‌رغم سادگی و ویژگی‌های جذب‌کننده CTT، در کاربرد آن با چندین مشکل مواجه هستیم. شاخص‌های مورد قضاوت در مورد تناسب سوالات و برآورد پارامترهای سوالات و برآورد پارامترهای آزمون، هم به آزمودنی‌ها و هم به آزمون وابسته است. در عمل، وابستگی به آزمون و نمونه ممکن است کاملاً برای سازندگان آزمون و استفاده‌کنندگان از آن در به‌کارگیری CTT مشکل‌زا شود، زیرا ممکن است به تصمیم‌گیری‌های نادرست در مورد گنجاندن یا

1-Gulliksen

2-Hambleton; Swaminathan & Rojers

حذف سوالاتی منجر شود که ممکن است برای اندازه‌گیری خصیصه مکنون[1] مناسب یا نامناسب باشد. به‌عبارتی، یکی از معایب CTT، وابستگی آماره‌های سوال و نمرات ویژگی نهفته[2] به نمونه آزمودنی و سوالات آزمون است. از طرف دیگر، برآوردهای CTT از پارامتر قدرت تشخیص[3] (r_{bis})و ضریب پایایی همسانی درونی[4] (آلفای کرونباخ[5])، به واریانس کوواریانس همان سوالات وابسته است و واریانس سؤالات به نمونه آزمودنی‌هایی که پارامترها از بررسی آن‌ها برآورد می‌شود، وابسته است. این وابستگی چرخشی در زیر، به‌خوبی بیان می‌شود:

« هم یک سوال دشوار و یا آسان به توانایی آزمودنی‌های مورد اندازه‌گیری وابسته است و هم توانایی آزمودنی‌ها به این وابسته است که آیا سوالات آزمون دشوار یا آسان است. ضریب تمیز و پایایی نمره آزمون و روایی برحسب یک گروه مشخص از آزمودنی‌ها تعریف می‌شود. آماره‌های آزمون و سوال تغییر می‌یابد، همان‌طور که نمونه آزمودنی‌ها تغییر می‌کند؛ و ویژگی‌های آزمودنی‌ها تغییر می‌یابد. همان‌طور که محتوا یا مضمون سوالات تغییر می‌کند (همبلتون و سوامیناتان و راجرز، ۱۹۹۱).

برای توضیح بیشتر در این رابطه لازم است گفته شود که:

---

۱ -در نظریه‌های خصیصه مکنون (Latent Trait). فرض بر این است که مهمترین جنبه‌های عملکرد آزمون می‌تواند با تعیین وضعیت فرد در یک خصیصه مکنون ـ یک ویژگی فرضی و مشاهده نشده یا خصیصه، مثل توانایی کلامی، معلومات تاریخی یا برونگرایی ـ توصیف شود. الگوهای نظریه‌های خصیصه مکنون به این منظور طراحی شده‌اند تا نحوه تأثیرگذاری خصیصه مکنون را بر عملکرد هر یک از سوال‌های آزمون توصیف کند (آلن و ین، ۱۹۷۹/ترجمه دلاور، ۱۳۷٤).

2-Latent Characteristic
3-Item Discriminating Parameter
4-Internal Consistency Reliability
5-Alfa's Cronbach

۱- در آزمون‌های چندگزینه‌ای (گزیده- پاسخ)، سازنده آزمون برای هر سوال باید دو پارامتر را برآورد کند: (الف) پارامتر دشواری سوال، و (ب) پارامتر تمیز یا قدرت تشخیص سوال. در نظریه کلاسیک اندازه‌گیری، p_i به عنوان برآوردکننده درجه دشواری سوال است که مقدار آن بین صفر تا ۱+ است و مقادیر پایین آن نشان‌دهنده دشواری سوال و مقادیر بالای آن بیانگر آسان بودن سوال است. در این نظریه، قدرت تشخیص یا تمیز سوال نیز که مقدار آن بین **1±** است، به‌صورت همبستگی بین سوال و نمره کل آزمون مطرح می‌گردد و معمولاً از ضریب همبستگی دورشته‌ای (r_{bis}) و دورشته‌ای نقطه‌ای (r_{pbis}) استفاده می‌شود. پایین بودن این ضریب بیانگر این نکته است که پاسخ صحیح به سوال رابطه‌ای با نمره کل آزمون ندارد و یا رابطه کمی دارد. در این نظریه، سوالاتی که قدرت تشخیص پایین یا منفی دارند، از آزمون حذف می‌شوند (به‌منظور بهبود اندازه‌گیری ویژگی موردنظر). در CTT برآورد پارامترهای سوال از طریق مشخصه‌های نمونه به گروه آزمودنی بستگی دارد و با تغییر آزمودنی‌ها و انتخاب نمونه دیگری از جامعه مورد بررسی، پارامترهای سوالات به‌علت خطای نمونه‌گیری[1] دچار تغییر می‌شوند. شاخص دشواری سوال عبارت است از نسبت آزمودنی‌هایی که به سوال i پاسخ درست داده‌اند. در نتیجه، وقتی نمونه آزمودنی‌ها دارای توانایی بالایی باشند، دشواری سوال بیشتر از موقعی خواهد بود که همین سوال (یا سوالات) روی نمونه‌ای از آزمودنی‌ها با توانایی پایین اجرا شود. به‌عبارت دیگر، برای گروه نمونه قوی‌تر، سوال ساده و برای گروه نمونه ضعیف‌تر، سوال مشکل خواهد بود. شاخص قدرت تشخیص سوال نیز تحت‌تأثیر گروه

1-Sampling Error

نمونه است و در نمونه‌های همگون‌تر، ضرایب تشخیص پایین‌تری نسبت به نمونه‌های ناهمگون برای سؤلات به‌دست می‌آید. بنابراین، میزان فایده و کاربرد مشخصه‌های سوال و برآوردهای اعتبار آزمون در نظریه کلاسیک اندازه‌گیری تحت‌تأثیر میزان معرف بودن[1] نمونه آزمودنی‌ها برای جامعه‌ای است که آزمون برای آن ساخته می‌شود (همبلتون، ۱۹۸۹). مسئله وابسته بودن سوال به نمونه و شاخص‌های آماری آزمون، مشکلات عدیده‌ای را در تعمیم‌پذیری[2] نتایج آزمون به‌وجود آورده است و نتایج به‌دست آمده را محدود به گروه نمونه مورد آزمون می‌سازد (لرد، ۱۹۸۰). هم‌چنین عامل مهمی‌که در میزان پایایی آزمون اثر می‌گذارد، ماهیت گروهی است که مبنای اندازه‌گیری پایایی است. نخست باید توجه داشت که دامنه تفاوت‌های فردی گروه، میزان ضریب همبستگی را تحت‌تأثیر قرار می‌دهد. به‌عنوان مثال، هرگاه تمام افراد یک گروه از لحاظ توانش املاء هم‌سطح باشند، در این گروه همبستگی نمرات املاء با هر نوع توانش دیگر برابر با صفر خواهد بود. روشن است که در چنین گروهی نمی‌توان جایگاه فرد خاصی را از لحاظ توانش‌های دیگر برمبنای اطلاع از نمره املاء پیش‌بینی کرد (آناستازی، ۱۹۸۸/ ترجمه براهنی، ۱۳۷۹). ضرایب پایایی نیز مثل ضرایب همبستگی، در نظریه کلاسیک، تابع درجه تغییرپذیری گروهی که مبنای محاسبه ضرایب است، قرار می‌گیرد. ویژگی وابستگی به گروه نمونه، سودمندی مشخصه‌های سوال را در کار پرورش آزمون و دیگر کاربردها کاهش می‌دهد. آزمون‌سازان نیازمند پارامترهای ثابت و تغییرناپذیر برای سؤالات هستند تا بر اساس آن‌ها بتوان حتی نحوه پاسخگویی آزمودنی‌ها را به سؤالات، قبل از

1-Representativeness

2-Generelizability

اجرای آن، پیش‌بینی کرد. این امکان را نظریه سوال ـ پاسخ با فراهم ساختن پارامترهای تغییرناپذیر ایجاد کرده است (لرد، ۱۹۸۰).

۲- نارسایی دیگر نظریه کلاسیک اندازه‌گیری در ارتباط با نارسایی قبلی است. نه‌تنها آماره‌های سؤالات تحت‌تأثیر توزیع توانایی آزمودنی‌ها و تفاوت‌های بین نمونه‌ها (تفاوت‌های میان‌گروهی) است، بلکه اندازه‌های توصیف‌کننده افراد گروه نیز از ویژگی‌های سوالات آزمون تأثیر می‌پذیرد و نمونه‌های متفاوت سوال، منجر به برآوردهای متفاوتی از توانایی افراد مورد سنجش می‌شود. در مدل کلاسیک، چه نمره مشاهده شده و چه نمره حقیقی، هر دو وابسته به آزمون مورد استفاده است. معمولاً تعداد یا نسبت پاسخ صحیح به سوالات آزمون، ملاک محاسبه نمره مشاهده شده است و نمره مشاهده شده نیز برآوردی از نمره حقیقی محسوب می‌شود. البته این‌که نمره حقیقی و نمره مشاهده شده وابسته به آزمودنی هستند، بدیهی و آشکار است. اساساً آزمون‌ها برای برآورد همین تفاوت‌ها طراحی و ساخته می‌شوند. از آن‌جا که برحسب میزان دشواری آزمون، تعداد یا نسبت پاسخ‌های صحیح به سوالات تغییر می‌کند، نمرات مشاهده شده افراد نیز دستخوش نوسان شده و هنگامی‌که یک فرد یا گروه که به دو آزمون ساده و دشوار ـ که معطوف به سنجش یک صفت یا توانایی است ـ سنجش شوند، نمرات متفاوتی در توانایی یا صفت مورد نظر به‌دست خواهند آورد (فراهانی، ۱۳۷۵).

لرد در سال‌های ۱۹۵۲، ۱۹۵۳ و همکاران روان‌سنج قبل از او (برای مثال گالیکسن، ۱۹۵۰) به نظریه‌ها و مدل‌هایی از روانسنجی علاقه‌مند بودند که بتواند به توصیف آزمودنی‌ها مستقل از انتخاب ویژه سوالات یا ارزیابی اموری که در یک آزمون استفاده خواهند شد، منجر

شود. هم‌چنین، برخی از روان‌سنج‌ها احساس کردند که کیفیت اندازه‌گیری، در صورت ایجاد شاخص‌های سوالات و آزمون مستقل از نمونه، افزایش می‌یابد. برای متخصصان اندازه‌گیری که برای برآورد یا ثبات سوال و آزمودنی اهمیت زیادی قایل هستند، یکی از راه‌حل‌ها استفاده از مفاهیم، مدل‌ها و روش‌های مربوط به نظریه سؤال- پاسخ است. نظریه سؤال- پاسخ یک نظریه جامع آماری درباره عملکرد سؤال، آزمون و آزمودنی و چگونگی سنجش توانایی‌هایی است که به‌وسیله سؤالات در آزمون اندازه‌گیری شده است. مقیاس سوال ـ پاسخ‌ها ممکن است گسسته[1] یا پیوسته[2] باشد، و دو ارزشی[3] یا چند ارزشی[4] نمره‌گذاری شود. طبقات نمره سوال ممکن است منظم و یا نامنظم باشد. و یک توانایی یا چند توانایی در آزمون مستتر باشد. در مورد سؤال- پاسخ‌ها و توانایی یا توانایی‌هایی که می‌توان مشخص کرد، چند روش یا مدل وجود دارد (همبلتون و جونز[5]، ۱۹۹۳). در نظریه سوال- پاسخ، به‌جای تأکید بر نمرات کل آزمون، بر پاسخ آزمودنی‌ها بر هر یک از سوال‌های آزمون تأکید می‌شود. با استفاده از مدل‌های ریاضی پیچیده‌تر از آنچه در نظریه کلاسیک اندازه‌گیری به‌کار می‌رود، نظریه سؤال- پاسخ یک تابع ریاضی به‌دست می‌دهد که با آن می‌توان احتمال پاسخ درست به یک سؤال را به‌عنوان تابعی از «توانایی» آزمون‌شونده و هم‌چنین برخی ویژگی‌های سوال معرفی کرد. برخلاف روش‌های پیچیده ریاضی، در نظریه سؤال- پاسخ، اندیشه زیربنایی بسیار روشن و منطقی است. اندیشه زیربنایی این نظریه همان منطق زیربنای اندیشه بینه[6] روانشناس فرانسوی است.

1-Discrete
2-Continuous
3-Dichotomous
4-Polytomous
5-Hambleton & Jones
6-Binet

بینه معتقد بود که نسبت کودکانی‌که پاسخ یک سوال را می‌دانند با سن افزایش می‌یابد. در نظریه سوال- پاسخ به‌جای سن، توانایی به‌کار می‌رود و گفته می‌شود که احتمال پاسخ درست به سوال با افزایش توانایی افزایش می‌یابد. به دیگر سخن، در نظریه سوال- پاسخ فرض بر این است که احتمال پاسخ درست به یک سوال جبر با افزایش دانش جبر آزمودنی افزایش می‌یابد و این بالقوه سخنی منطقی است (سیف، ۱۳۸۰).

با توجه به محدودیت‌های نظریه کلاسیک اندازه‌گیری، یکی از مهم‌ترین ویژگی‌های بارز نظریه سوال- پاسخ در این است که برآورد پارامترها به‌گونه‌ای در این نظریه صورت می‌گیرد که مستقل از نمونه است و به‌عبارت دیگر، توصیف‌کننده‌های آماری یک سؤال آزمون، مانند شاخص‌های دشواری یا تمیز، از نمونه خاصی از آزمودنی‌ها که از جامعه مورد بررسی بیرون کشیده شده و آزمون برایشان اجرا شده است، مستقل هستند (همبلتون، ۱۹۸۹) و همان‌طور که همبلتون و جونز (۱۹۹۳) مطرح می‌سازند، IRT آماره‌هایی را مستقل از گروه‌هایی که آماره‌ها از آن‌ها به‌دست می‌آید، فراهم می‌سازد و این آماره‌ها از نمونه‌ای به نمونه دیگر، ثابت (یا نسبتاً ثابت) است. لرد (۱۹۸۰) بیان می‌کند که در نظریه سوال- پاسخ، احتمال پاسخ صحیح به یک سوال در آزمودنی‌هایی با سطح توانایی معین، فقط به توانایی آن‌ها بستگی دارد و به تعداد افرادی که در سطوح متفاوت توانایی قرار دارند، ارتباطی ندارد. به‌عبارتی، در IRT، احتمال پاسخ صحیح به یک سوال به توزیع توانایی گروه نمونه بستگی ندارد و بدون توجه به ویژگی‌های گروه نمونه، پارامترهای سوال ثابت و نامتغیر است. طبق نظر لرد و ناویک (۱۹۶۸)، تأثیر نپذیرفتن آماره‌های سؤال از ویژگی‌های آزمودنی‌ها، به نظریه سؤال- پاسخ امکان

می‌دهد که نحوه پاسخگویی افراد به سوال‌های یک آزمون را پیش‌بینی کند؛ این امر فراتر از کنترل نظریه کلاسیک اندازه‌گیری می‌باشد و به‌همین جهت برخلاف نظریه کلاسیک که نمی‌تواند احتمال پاسخ هر آزمودنی به سؤالات را پیش‌بینی کند (مگر این‌که قبلاً سؤال‌ها بر روی افراد مشابهی اجرا و مدرج[1] شده باشد)، در نظریه سؤال- پاسخ می‌توان پاسخ هر آزمودنی را به هر سؤال حدس زد (حتی اگر قبلاً آزمودنی‌های مشابه آن‌ها، به سؤالات مشابه پاسخ نداده باشند).

در مورد عدم تداخل ویژگی‌های سوالات در آزمودنی‌ها نیز باید گفت بین دو نظریه تفاوت وجود دارد. سوالی رایج در مورد نمره‌های آزمون، مربوط به جایگاه نسبی شخص در حوزه‌های مختلف است. اساساً آزمون‌ها برای تشخیص تفاوت‌های درون‌فردی[2] و بین‌گروهی[3] اشخاص ساخته می‌شوند. هدف از بررسی تفاوت‌های درون‌فردی، این است که تفاوت‌های مشاهده شده در نمرات یک فرد در متغیرهای مختلف حایز پایایی یا ثبات[4] است یا خیر؟ در مقایسه‌های بین‌گروهی، مسئله این است که آیا تفاوت‌های مشاهده شده بین نمرات افراد مختلف در یک متغیر خاص به‌قدر کافی پایا است که بتواند مثلاً مبنای انتخاب افراد برای شغل خاصی قرار گیرد یا خیر؟ (مگنوسون، ۱۹۶۲ / ترجمه براهنی، ۱۳۷۰). در مدل کلاسیک اندازه‌گیری برای پاسخ به این سوالات از تفاوت بین نمرات مشاهده شده در افراد آزمودنی استفاده می‌شود. در این نظریه، هم نمره مشاهده شده و هم نمره حقیقی، هر دو وابسته به آزمون مورد استفاده هستند. نمره‌گذاری این افراد در این نظریه بر مبنای تعداد یا نسبت

1-Colibration
2-Interindividual
3-Intraindividual
4-Stability

پاسخ‌های صحیح به سوالات آزمون است. لذا نمره افراد وابسته به سطح دشواری سؤالات مورد استفاده در آزمون است. اهمیت این ضعف زمانی آشکارتر می‌شود که ازمودنی‌ها در فرم‌های مختلف یک آزمون یا بخش‌های مختلف آن، مورد مقایسه قرار می‌گیرند. (همبلتون، ۱۹۸۹؛ ویس و پوس، ۱۹۹٤). این ضعف علاوه بر دشواری تعبیر و تفسیر نمرات حاصل از یک آزمون و تعیین جایگاه افراد در پیوستار ویژگی موردنظر، تجزیه و تحلیل و تفسیر نمرات حاصل از دو یا چند آزمون را که بر روی یک گروه ثابت اجرا شده است دشوار می‌سازد. لذا نه می‌توان به مقایسه افراد و قضاوت در مورد تفاوت‌های درون‌فردی پرداخت، و نه از تفاوت‌های میان‌گروهی سخن گفت. این امر انتظار سازندگان آزمون را در برآورد توانایی آزمودنی بدون تورش و نااریب دچار اشکال می‌سازد (همبلتون و کوک، ۱۹۷۷).

برعکس نظریه کلاسیک اندازه‌گیری، از آنجا که در نظریه سوال- پاسخ توانایی‌های افراد، مجزا از توانایی‌های سایر اعضای گروه برآورد می‌شود. توانایی و کارکرد فرد از یک آزمون به آزمون دیگر (آزمون همتا)، تفاوت چندانی نشان نمی‌دهد. به‌عنوان مثال، بنا به نظریه کلاسیک اگر دو آزمون A و B با تعداد سوال‌های برابر برای ارزشیابی توانایی واحد یک فرد خاص مورد استفاده قرار گیرد، فردی که در آزمون A با دادن پاسخ صحیح به تعداد زیادی سوال مرتبه بالایی کسب کرده باشد، ضرورتاً در آزمون B چنین وضعیتی نخواهد داشت و رتبه وی تغییر خواهد کرد. این موضوع مؤید تغییرپذیری جایگاه فرد با توجه به نمرات مشاهده شده از یک آزمون به آزمون دیگر خواهد بود. وقتی فردی به دو آزمون ساده و دشوار پاسخ می‌دهد، نمرات مشاهده شده او تحت‌تأثیر دشواری سوال قرار می‌گیرد و در نتیجه

متفاوت خواهد بود. برآورد میزان توانایی این فرد نیز تحت‌تأثیر دشواری سوال‌های آزمون، کمتر یا بیشتر خواهد شد. برعکس، در نظریه سوال- پاسخ، آزمونگر، آزمودنی‌ها را برحسب میزان توانایی به‌دست آمده مقایسه می‌کند. برای مثال اگر دو دانش‌آموز که در دو آزمون ساده و دشوار که معطوف به سنجش صفت معینی هستند، نمره یکسانی را کسب کنند، برآورد میزان توانایی آنان بر اساس نظریه IRT متفاوت از یکدیگر خواهد بود و توانایی دانش‌آموزی که به سوالات آزمون دشوار پاسخ داده است، بیشتر از میزان توانایی دانش‌آموزی است که به آزمون ساده و آسان پاسخ داده است. این ویژگی از مهم‌ترین مزیت‌های نظریه سوال- پاسخ بر نظریه کلاسیک است که باعث کارآمدی این نظریه در سنجش انطباقی[1] شده است (همبلتون و کوک[2]، ۱۹۷۷). از آنجا که در آزمون‌های تحصیلی و شغلی لازم است توانایی‌ها و استعدادهای فرد آزمون‌شونده به‌گونه‌ای برآورد شود که در آزمون‌های مشابه همواره جایگاه ثابتی داشته باشد، نظریه سوال- پاسخ راه‌حل مناسبی برای برآوردن این نیاز خواهد بود. به‌عبارت دیگر، وابستگی آزمون، دیگر یک مسئله در IRT نیست، زیرا نمرات توانایی آزمودنی‌ها به‌نوع آزمون به‌کار رفته وابسته نیست. در IRT، درک عملکرد هر یک از سوالات (منسوب به استقلال موضعی[3]: یک سؤال در یک آزمون کاملاً مستقل از دیگر سؤالات آزمون است) امکان‌پذیر است و هر سؤال می‌تواند به‌طور مستقل از دیگر سؤالات ارزیابی شود.

1-Adaptive Testing

2-Hambleton & Cook

۳- مفروضه استقلال موضعی (Local Independence) بیان می‌کند که برای یک نمره واقعی یا ارزش خصیصه مکنون معین، نمره‌های سوال مستقل از یکدیگرند و عملکرد آزمودنی در یک سوال مستقل از عملکرد او در سایر سوالات است و بنابراین هیچ سوالی نباید راهنمایی برای پاسخگویی به سایر سوالات فراهم کند. طبق این مفروضه، تنها سطح توانایی آزمودنی و ویژگی‌های هر سوال بر عملکرد آزمودنی در آن سوال مؤثر است (همبلتون و جونز، ۱۹۹۳).

با وجود آن‌که کار با نظریه سوال- پاسخ از سال‌های دهه ۱۹۴۰ آغاز شده است و مبانی نظری آن با شتاب هر چه بیشتر گسترش یافته است و نوشته‌ها و مقالات در زمینه این نظریه در غرب، چشم‌انداز بسیار روشنی را پیش‌روی سازندگان آزمون قرار داده است، اما در کشور ما این نظریه هنوز آن‌چنان که باید و شاید حتی نزد صاحب‌نظران و دانشگاهیان شناخته شده نیست و تعداد تحقیقات صورت گرفته در زمینه نظریه سؤال- پاسخ بسیار اندک می‌باشد. با این حال، موقعیت‌های عملی بسیاری پیش می‌آید که استفاده از مدل‌های نظریه سؤال- پاسخ را ضروری می‌سازد. سازمان سنجش آموزش کشور و دیگر موسسات آموزشی هرساله خیل عظیمی از داوطلبان ورود به دانشگاه‌ها را مورد آزمون قرار می‌دهند تا از میان آن‌ها، تعداد اندکی به دانشگاه‌ها و مراکز آموزش عالی راه‌یابند. اما نظریه سوال- پاسخ، هنوز در اندازه‌گیری توانایی‌های داوطلبان در پاسخ‌گویی به سؤالات به این سازمان‌ها راه نیافته است. با وجود این، الزامات جامعه ما، به‌ویژه در سال‌های اخیر، ایجاب می‌کند که هر چه دقیق‌تر و درست‌تر افراد بر مبنای توانایی‌های ذهنی خود از یکدیگر متمایز شده و برای تحصیل در دانشگاه‌ها و موسسات آموزش عالی انتخاب شوند.

با توجه به مطالب فوق، این پرسش مطرح است که دو روش اندازه‌گیری کلاسیک و سوال- پاسخ از نظر تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سؤال و ویژگی‌های سؤال بر خصوصیات آزمودنی‌ها چه تفاوتی با یکدیگر دارند؟

به‌منظور پاسخ به سوال فوق و با توجه به اهمیت و فواید نظریه سؤال- پاسخ و نیز به‌جهت عدم توجه به آن در ایران به‌طور مناسب و شایسته، این پژوهش در پی آن بوده است

که مزیت‌های نظریه سؤال- پاسخ را بر نظریه کلاسیک از جنبه تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سؤال و ویژگی‌های سؤال بر خصوصیات آزمودنی‌ها بررسی کند. لذا با استفاده از داده‌های حاصل از اجرای این پژوهش، سؤالات زیر مورد بررسی قرار گرفت:

۱ـ نظریه کلاسیک اندازه‌گیری و نظریه سؤال- پاسخ، از نظر تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سؤال، چه تفاوتی دارند؟

۲ـ نظریه کلاسیک اندازه‌گیری و نظریه سؤال- پاسخ، از نظر تأثیر ویژگی‌های سؤال بر خصوصیات آزمودنی‌ها، چه تفاوتی دارند؟

روش

جامعه

به‌منظور پاسخ به پرسش‌های پژوهشی موردنظر، از روش کاربردی- توصیفی استفاده شد و دو جامعه مورد بررسی قرار گرفت. این دو جامعه عبارتند از کلیه داوطلبان ورود به دانشگاه‌های کشور در رشته ریاضی ـ فیزیک و هم‌چنین کلیه داوطلبان ورود به دانشگاه‌های کشور در رشته علوم‌تجربی در سال ۱۳۷۸. جدول ۱، تعداد داوطلبان شرکت کننده در کنکور سراسری ۱۳۷۸ را در رشته ریاضی- فیزیک و علوم‌تجربی نشان می‌دهد.

جدول ۱. تعداد داوطلبان شرکت‌کننده در کنکور سراسری ۱۳۷۸ در رشته ریاضی ـ فیزیک و علوم تجربی

رشته	تعداد داوطلبان زن	تعداد داوطلبان مرد	کل
ریاضی ـ فیزیک	۹۸٤۲۰	۲۰۲۱٦۷	۳۰۰۵۸۷
علوم تجربی	۲۸۸۱۷۷	۱٦٦٥٦٦	٤٥٤۷٤۳

نمونه

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- 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، نمره کسب کرده بودند)، از بین افراد باقیمانده چند نمونه‌گیری صورت گرفت.

برطبق نظر طرفداران نظریه کلاسیک، یکی از مزایای این نظریه آن است که حجم نمونه در آن نسبت به نظریه سؤال- پاسخ کمتر است. مطالعات اندکی وجود دارد که تأثیر حجم نمونه را بر برآوردهایی که برای پایایی پارامترها صورت می‌گیرد، به‌طور منظم بررسی کرده باشند. حداقل حجم نمونه پیشنهادی برای به‌کارگیری مؤثر CTT در دامنه‌ای از ۳۰۰ تا ۵۰۰ متغیر است. البته ترجیح داده می‌شود که حجم نمونه ۱۰۰۰ باشد (تروسکوسکی[1]، ۱۹۹۹ به‌نقل از نانالی[2]، ۱۹٦۷). مطابق پژوهش‌های موجود در زمینه آزمون‌سازی، کاربرد موفق مدل‌های IRT مستلزم استفاده از سؤالات و آزمودنی‌هایی با حجم

1-Truskosky

2-Nanaly

بزرگ است تا بتوان به‌طور همزمان صفت مکنون و پارامترهای سؤال را برآورد کرد (لرد، ۱۹۶۸؛ همبلتون و سوآمیناتان، ۱۹۸۵). بر اساس مدل انتخاب شده در نظریه سؤال- پاسخ به محقق توصیه می‌شود که حجم نمونه را برای مدل‌های تک‌پارامتری حداقل ۲۰۰ (رایت و استون، ۱۹۷۹)، برای مدل‌های دو پارامتری تا ۵۰۰ (گدمن و راجو، ۱۹۸۶)، و برای مدل‌های سه پارامتری بالای ۱۰۰۰ (تیسن و وایتر، ۱۹۸۲) درنظر بگیرد. بطور کلی، همبلتون (۱۹۸۹) با درنظر گرفتن نتایج تحقیقات مطرح می‌سازد که برای مدل سه پارامتری، حداقل حجم نمونه‌ای برابر با ۱۰۰۰ لازم است. 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، از بین گروه‌های نمونه مختلف، در نهایت دو نمونه ۱۰۰۰ نفری که در آن‌ها تفاوت توانایی آزمودنی‌ها بیشتر بود، انتخاب شد. از بین داوطلبان گروه آزمایشی تجربی نیز یک گروه ۱۰۰۰ نفری به روش نمونه‌گیری منظم انتخاب شدند.

ابزار و روش اجرا و تحلیل داده‌ها

در آزمون ورودی دانشگاه‌ها در ایران از آزمون‌های پیشرفت تحصیلی استفاده می‌شود این آزمون‌ها در سازمان سنجش توسط اساتید و طراحان زبده و کارکشته کشور طراحی و تهیه می‌شود. ابزار مورد استفاده برای جمع‌آوری داده‌های پژوهش، پاسخنامه‌های داوطلبان در آزمون اختصاصی درس ریاضی در رشته ریاضی- فیزیک با ۵۵ سؤال و آزمون اختصاصی درس زیست‌شناسی در رشته علوم‌تجربی با ۵۰ سؤال بود.

در مورد تعداد سؤالات در زمینه مقایسه مدل‌ها، صاحب‌نظران حداقل حجم نمونه سوال را ۱۵ ذکر کرده‌اند. لیکن مانند آزمودنی‌ها، تعداد سؤالات خرده‌آزمون‌ها نیز از تعداد پارامترهای مدل تأثیر می‌پذیرد. در نتیجه به‌عقیده برخی، مدل یک پارامتری با ۲۰ سؤال، دو پارامتری با ۳۰ سؤال، و سه پارامتری با ۶۰ سؤال دارای کارآیی است. البته با افزایش حجم نمونه آزمودنی، تعداد کمتری از سؤالات- حتی در حد ۲۰ سوال- نیز می‌تواند برآوردهای قابل قبولی از پارامترها به‌دست دهد (همبلتون، ۱۹۸۹).

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

به منظور تحلیل داده‌ها در نظریه سوال- پاسخ، از روش‌های آماری مانند آزمون t وابسته، ضریب همبستگی پیرسون، جذر میانگین خطاها[1] (RMSE)، و آزمون‌های معنی‌داری آنها استفاده شد و سپس مقایسه‌های موردنظر انجام گرفت.

1-Root Mean Square Error

نتایج

با استفاده از آزمون اختصاصی ریاضی رشته ریاضی- فیزیک، سوال اول تحقیق، یعنی تأثیر خصوصیات آزمودنی‌ها بر ویژگی‌های سؤال، بررسی شد. در تحلیل سؤالات آزمون ریاضی به روش کلاسیک، ۷ سوال که فاقد برازندگی بود، حذف شد و ٤٨ سوال باقی ماند. بعد با توجه به نمونه‌ای که در اختیار بود، چندبار نمونه‌گیری صورت گرفت و هر بار نمره‌کل آزمودنی‌ها محاسبه شد و سرانجام دو نمونه‌ای که از نظر میانگین نمره‌کل افراد، بیشتر با یکدیگر تفاوت داشتند، انتخاب و شاخص‌های کلاسیک آن‌ها با هم مقایسه شد. بار دیگر با روش نظریه سوال- پاسخ این سؤالات تحلیل شد. پارامترهای سوال با توجه به مدل سه‌پارامتری این نظریه برآورد شد و مجذور خی سؤالات محاسبه شد و به‌دنبال آن، آزمون برازندگی برای هر سؤال و کل آزمون با استفاده از فرمول نیکویی برازندگی «ین» محاسبه شد. این مقدار نیز با مقدار بحرانی جدول χ^2 با درجه آزادی $n(m-k)$ مقایسه شد. در بررسی برازندگی هر سؤال و کل آزمون با استفاده از آزمون مجذور خی، کلیه محاسبات توسط برنامه نرم‌افزاری BILOG انجام شد و در نهایت ۱۰ سوال در دو مرحله به‌علت عدم برازندگی حذف شد و ٤٥ سوال جهت تجزیه و تحلیل باقی ماند. این‌بار نیز تعداد زیادی نمونه با ٤٥ سوال گرفته شد و توانایی (θ) آن‌ها محاسبه و سرانجام دو نمونه‌ای که بین میانگین θهای آن‌ها تفاوت زیادی بود، انتخاب شد. با این دو نمونه، پارامترهای سوال برآورد شده و سپس مقیاس‌هایش یکسان‌سازی شد (از آن‌جا که سوال‌ها یکی بود، θها استاندارد شد). سرانجام بر اساس شاخص‌های CTT و پارامترهای IRT هر یک بر روی دو گروه نمونه، به روش‌های

ضریب‌همبستگی، آزمون t، و جذر میانگین مربع خطاها (RMSE)، مقایسه‌های مورد نظر انجام گرفت. جدول ۲ مقایسه تفاوت مقادیر برآورد شده پارامترهای سوال در دو گروه آزمودنی در CTT و IRT را نشان می‌دهد. البته این مقایسه برای پارامتر حدس‌پذیری، به‌دلیل ثابت بودن میزان آن در نظریه کلاسیک میسر نگردید و مقایسه‌ها برای شاخص یا پارامتر دشواری و قدرت تشخیص سؤالات انجام پذیرفت.

جدول ۲. مقایسه تفاوت مقادیر برآورد شده پارامترهای سوال در دو گروه نمونه

نظریه مورد استفاده	تعداد سوال	تفاوت شاخص یا پارامتر سطح دشواری				تفاوت شاخص یا پارامتر قدرت تشخیص			
		میانگین تفاوت	انحراف معیار تفاوت	حداقل تفاوت	حداکثر تفاوت	میانگین تفاوت	انحراف معیار تفاوت	حداقل تفاوت	حداکثر تفاوت
CTT	٤٨	۰/۱٤	۰/۰۹	۰/۰۲۱	۰/۲٦	۰/۱	۰/۳	-۰/۳۲	۰/۷٦
IRT	٤٥	۰/۰۱	۰/۰۸	-۰/۲۰	۰/۲۷	۰/۰	۰/۸	-۰/۲٦	۰/۱۷

از بررسی جدول ۲ می‌توان دریافت که در نظریه IRT، تفاوت‌ها در حد کمتر از خطای استاندارد ($\frac{s}{\sqrt{n}}$ = خطای استاندارد برآورد) است و در نتیجه مقادیر برآورد شده در دو گروه آزمودنی با یکدیگر تفاوت معنی‌دار ندارند، برعکس، در نظریه کلاسیک، تفاوت چند برابر خطای استاندارد برآورد است و در نتیجه تفاوت بین دو مقدار برآوردشده در دو گروه از آزمودنی‌ها معنی‌دار می‌باشد.

همچنین همان‌طور که جدول ۳ نشان می‌دهد، ضرایب همبستگی برآورد شاخص‌های دشواری، قدرت تشخیص و حدس‌پذیری در نظریه کلاسیک مقادیر بسیار پایینی را نشان

می‌دهد، در حالی‌که در IRT 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 IRT چنین نیست.

جدول ۳. مقایسه ضرایب همبستگی شاخص‌ها یا پارامترهای سوال در CTT و IRT

نظریه مورد مقایسه	تعداد سوالات	ضریب همبستگی برآورد شاخص	سطح معنی‌داری	پارامتر قدرت تشخیص	سطح معنی‌داری	ضریب همبستگی برآورد شاخص	سطح معنی‌داری
CTT	٤٨	۰/۳۵۰	۰/۰۱۵	۰/۳۰۳	۰/۰۳٦	-	-
IRT	٤٥	۰/۹۹۷	۰/۰۰۰	۰/۹۵۹	۰/۰۰۰	۰/۹۹٤	۰/۰۰۰

مقایسه دو نظریه با آزمون t نیز بیانگر همین نکته می‌باشد. جدول ٤ مقادیر برآورد شده شاخص‌ها یا پارامترهای سوال را در دو نظریه به‌وسیله آزمون t با یکدیگر مقایسه می‌کند.

جدول ٤. مقایسه پارامتر دشواری و قدرت تشخیص سوال در CTT و IRT با آزمون t

نظریه مورد مقایسه	تعداد سوالات	آزمون t برای سطح دشواری	سطح معنی‌داری	آزمون t برای قدرت تشخیص	سطح معنی‌داری
CTT	٤٨	٦/٤٥ (d.f =٤٧)	۰/۰۰۰	۹/۳۷۹ (d.f =٤٧)	۰/۰۰۰
IRT	٤٥	۰/۸٦۲ (d.f =٤٤)	۰/۳۹٤	-۰/۱۳۲ (d.f =٤٤)	۰/۸۹٥

همان‌طور که از جدول ٤ ملاحظه می‌شود، آزمون t نیز نشان داد که دو نظریه در برآورد پارامترهای سوالات با یکدیگر تفاوت دارند. مقدار t در نظریه کلاسیک تفاوت معنی‌داری را بین مقادیر برآورد شده پارامترها در دو گروه نشان داد، در حالی‌که در نظریه IRT این تفاوت‌ها در هیچ‌مورد معنی‌دار نیست. این امر نیز تأییدکننده ثابت و نامتغیر بودن برآورد پارامترها در IRT است و این خود مزیتی اساسی برای سازنده آزمون است که برآورد پارامترها فقط تحت‌تأثیر خود سوال قرار دارد و خصوصیات آزمودنی‌ها بر آن‌ها هیچ تأثیری ندارد.

مقایسه RMSE های به‌دست آمده در مورد پارامترهای سوالات این آزمون به دو روش کلاسیک و سوال ـ پاسخ نیز تفاوت شدید دو نظریه را در برآورد پارامترها نشان داد. این تفاوت در جدول ٥ منعکس شده است.

جدول ٥. مقایسه RMSE پارامتر دشواری و قدرت تشخیص سوال در CTT و IRT

نظریه مورد مقایسه	تعداد سوالات	RMSE دشواری	RMSE قدرت تشخیص
CTT	٤٨	١/١٢٨	١/١٦٨
IRT	٤٥	٠/٠٧٠	٠/٢٨٥

همان‌طور که در جدول ٥ مشاهده می‌شود، در تمام پارامترهای سوال در IRT، مقادیر RMSE بیانگر عدم تفاوت معنی‌دار بین برآورد پارامترها در دو گروه نمونه است. در CTT این مقادیر بیانگر وجود تفاوت معنی‌دار در هر دو مورد در دو گروه نمونه است. این امر مؤید این نکته است که برآورد شاخص‌ها در نظریه کلاسیک تحت‌تأثیر گروه آزمودنی‌ها قرار می‌گیرد و غیرثابت و یکسان است.

در بررسی سوال اول تحقیق ملاحظه شد که ویژگی‌های دو گروه نمونه که از نظر توانایی با یکدیگر تفاوت داشتند، در برآورد شاخص‌های سوال در نظریه کلاسیک تأثیر می‌گذارند؛ ولی در برآورد پارامترهای سوال در نظریه سوال ـ پاسخ بی‌تأثیر می‌باشند و این امر، مزیت عمده و اساسی نظریه IRT برCTT محسوب می‌شود. این مسئله در نظریه کلاسیک، اختلال زیادی در کار آزمون‌سازی ایجاد می‌کند.

برای بررسی سوال دوم تحقیق، یعنی تأثیر ویژگی‌های سوال بر برآورد توانایی آزمودنی‌ها، از سوالات اختصاصی درس زیست‌شناسی رشته علوم تجربی سال ۱۳۷۸ استفاده شد. آزمون زیست‌شناسی که شامل ۵۰ سوال بود به دو نیمه آزمون تقسیم شد و دو مجموعه سوال ساخته شد که سوالات فرد، یک آزمون و سوالات زوج، آزمون دوم را تشکیل می‌دادند و سپس نمرات گروه نمونه (یک گروه ۱۰۰۰ نفری) در این دو آزمون بررسی شد. جدول ٦ میزان همبستگی بین θهای برآوردشده در دو آزمون را بر یک گروه آزمودنی نشان می‌دهد.

جدول ٦. ضریب همبستگی θهای برآوردشده دو آزمون بریک گروه آزمودنی

روش مورد استفاده	تعداد آزمودنی‌ها	میانگین تتای برآوردشده		همبستگی برآوردشده	سطح معنی‌داری
		آزمون A	آزمون B		
IRT	۱۰۰۰	۰/۰۳۲	۰/۰۳۲	۰/۸۷٦	۰/۰۰۱

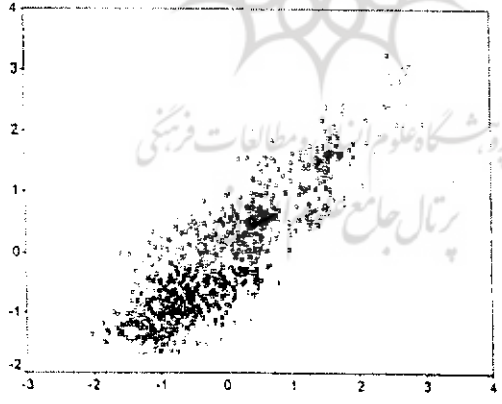
آزمون t به دلیل همبسته بودن هر دو آزمون صفر است و در نتیجه از این جنبه بررسی نشد.

با استفاده از RMSE نیز برآورد پارامترها در گروه آزمودنی با اجرای دو آزمون بررسی شد و نتایج آن در جدول ۷ آمده است.

جدول ۷. مقایسه نتایج اجرای دو آزمون بر یک گروه آزمودنیدر برآورد پارامتر توانایی با IRT

نظریه مورد استفاده	تعداد آزمودنی‌ها	میانگین برآورد پارامتر توانایی	انحراف‌معیار برآورد پارامتر توانایی	RMSE
IRT	۱۰۰۰	۰.۰۰۰۰۲۵	۰ ۰۰۰۳۸	۰/۴۹۸

همان‌طورکه ملاحظه می‌شود، استفاده از دو آزمون نشان داد که تفاوت توانایی‌های برآوردشده در یک گروه از آزمودنی‌ها و در دو آزمون معنی‌دار نیست و تفاوت‌ها در حد خطای اندازه‌گیری است (نظر به این‌که میزان RMSE برابر ۰/۴۹۸ بوده و از ۰/۸۳ کمتر است، تفاوت توانایی‌های برآوردشده معنی‌دار نیست). نمودار شماره ۱ مربوط به بررسی سطوح توانایی گروه مورد آزمون در دو آزمون زوج و فرد اجرا شده نیز بیانگر ثبات مقادیر برآوردشده پارامتر توانایی در دو آزمون است.



نمودار ۱. نمودار توزیع مقادیر برآوردشده پارامتر توانایی در دو آزمون

بحث و نتیجه‌گیری

در نظریه کلاسیک اندازه‌گیری، شاخص‌های سوال که از طریق مشخصه‌های نمونه برآورد می‌شود، به گروه نمونه بستگی دارد و با تغییر گروه نمونه و انتخاب یک گروه متفاوت از جامعه مورد بررسی، مقدار برآورد شده شاخص‌های سوالات آزمون به‌علت خطای نمونه‌گیری تغییر می‌کند. اگر یک مجموعه سوال بر روی یک گروه نمونه قوی اجرا شود، درجه دشواری سوالات، متفاوت از موقعی خواهد بود که این سوالات بر روی یک گروه آزمودنی ضعیف اجرا شود. به‌عبارتی، هنگامی که متوسط توانایی افراد کاهش یا افزایش یابد، دشواری سوال نیز تغییر می‌کند. حال آن‌که سازنده آزمون به‌دنبال مشخصه‌هایی است که به‌خوبی از عهده توصیف سوال برآیند. نسبت پاسخ‌های صحیح به سوال علاوه بر سوالات آزمون، ویژگی‌های گروه مورد آزمون را نیز توصیف می‌کند. هم‌چنین شاخص قدرت تشخیص سوالات در گروه‌های ناهمگن بالاتر از گروه‌های همگن خواهد بود. این امر ناشی از تأثیر ناهمگنی گروه نمونه بر ضرایب همبستگی است. درمقابل این ضعف نظریه کلاسیک، ثبات و عدم تغییر پارامترهای سوال در بین گروه‌های مختلف و تأثیر نپذیرفتن این پارامترها از ویژگی‌های نمونه، یکی از مهم‌ترین ویژگی‌های IRT است. در این نظریه، اگر یک مجموعه آزمون بر روی دو گروه نمونه متفاوت اجرا شود و سپس قدرت تشخیص سوال برای هر گروه محاسبه گردد، برآورد مزبور در هر دو گروه، برآوردی بدون اریب و نامتغیر است و تحت‌تأثیر ناهمگنی گروه آزمودنی‌ها قرار ندارد. در نظریه سوال- پاسخ، احتمال پاسخ صحیح به سوال برای آزمودنی‌هایی که در سطح معینی از توانایی قرار دارند، فقط به توانایی آنان بستگی دارد؛

و در نتیجه به تعداد افرادی که در سطح معینی از توانایی قرار دارند و به تعداد افرادی که در سطوح متفاوت توانایی قرار دارند، بستگی ندارد. بنابراین، این احتمال به توزیع توانایی گروه آزمودنی بستگی ندارد و بدون توجه به ویژگی‌های گروه یا گروه‌های آزمودنی، پارامترهای سوال ثابت است. این امر با مطالعات همبلتون و کوک (۱۹۷۷)، همبلتون (۱۹۸۹)، ویس و یوس (۱۹۹٤) مطابقت دارد. در بررسی این سوال تحقیقی بر روی داده‌های داوطلبان کنکور سراسری ۱۳۷۸ مشخص شد که در نظریه IRT، تفاوت مقادیر برآوردشده پارامترهای سوال، در حد کمتر از خطای استاندارد اندازه‌گیری است و در نتیجه مقادیر برآوردشده در دو گروه تفاوت معنی‌داری با یکدیگر ندارند، اما در نظریه CTT، تفاوت بین مقادیر برآوردشده شاخص‌های سوال زیاد است و تفاوت بین دو گروه معنی‌دار است.

۲- از آن‌جا که نمره مشاهده شده در نظریه کلاسیک اندازه‌گیری وابسته به آزمون است و نمره‌گذاری برمبنای تعداد پاسخ‌های صحیح صورت می‌گیرد، نمره افراد وابسته به سطح دشواری یا آسانی سوالات آزمون است و در نتیجه در مقایسه آزمودنی‌ها در فرم‌های مختلف آزمون یا بخش‌های مختلف یک آزمون اختلال به‌وجود می‌آید. برعکس در نظریه سوال- پاسخ، توانایی فرد مجزا از توانایی سایر اعضای گروه برآورد می‌شود و لذا توانایی او از یک آزمون به آزمون دیگر چندان تغییر نمی‌کند و برآورد توانایی در IRT تقریباً ثابت و بدون تغییر است. این امر منطبق با مطالعات همبلتون و کوک (۱۹۷۷)، همبلتون (۱۹۸۹)، همبلتون و جونز (۱۹۹۳) است. به‌طور کلی می‌توان نتیجه گرفت که در نظریه کلاسیک، سوالات مختلف، مقادیر متفاوتی را از توانایی به‌دست می‌دهد. به‌عبارت دیگر، وقتی سوالات آسان است،

آزمودنی‌های بیشتری به آن پاسخ می‌دهند و در نتیجه توانایی آن‌ها بیشینه برآورد می‌شود. هنگامی‌که سوالات دشوار است، تعداد کمتری با همان سطح توانایی به آن پاسخ می‌دهند و در نتیجه توانایی آن‌ها کمینه برآورد می‌شود. به‌همین دلیل، در نظریه کلاسیک، برآورد پارامتر توانایی تحت‌تأثیر ویژگی‌های سوالات قرار دارد و متفاوت است. اما در نظریه سوال- پاسخ، همان‌طور که بررسی‌ها نشان داد، ویژگی‌های سوالات در برآورد پارامتر توانایی آزمودنی‌ها تأثیر ندارد و پارامتر توانایی نیز در این نظریه، همچون پارامترهای سوال یک پارامتر نامتغیر و بدون اریب است.

منابع

الف:فارسی

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

آناستازی، آ.(۱۳۷۹). روان‌آزمایی، ترجمه محمدنقی براهنی. تهران: دانشگاه تهران.(۱۹۸۸).

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

ثرندایک، رابرت.(۱۳۷۵). روان‌سنجی کاربردی. ترجمه حیدرعلی هومن. تهران: انتشارات دانشگاه تهران.(۱۹۸۲).

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

مگنوسون، داوید (۱۳۷۰). مبانی نظری آزمون‌های روانی. ترجمه محمدنقی براهنی. تهران: انتشارات دانشگاه تهران.(۱۹٦۲).

فراهانی، مهدی.(۱۳۷۵). مقایسه مدل‌های اندازه‌گیری (کلاسیک و سوال- پاسخ) از لحاظ برآورد پارامترهای سوال و توانایی. پایان‌نامه منتشرنشده کارشناسی ارشد، دانشکده روانشناسی و علوم‌تربیتی دانشگاه علامه‌طباطبایی.

ب:انگلیسی

Goldman,S.H.&Raju,N.S.(1986).Recoveryofoneandtwoparameter LogisticItemParameters:Anempricalstudy,EducationalandPsychologic al Measurement.46, 11-21.

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 (3rd end).Newyork:Memillan. 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.;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.&Swaminathan,H.(1985).ItemResponseTheory: Principles and Applications,Boston:Kluwer.

Hambleton,R.K;Swaminathan,H.&Rojers.H.J.(1991).Fundamenta ls of Item Response Theory, Newbury park, CA: sage

Hambleton,R.K.&Vander Linden, Wim.J.(1982).Advance in Item ResponseTheoryandApplications:AnIntroduction,AppliedPsychplogic al 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

Thissen,D.&Wainer,H.(1982).Some Standard Errors in Item Response Theory, Psychometrika. 47,397-412.

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

Weiss,D.J.&Yoes,M.E.(1994).ItemResponseTheory.InR.K.Hambleton&Zaal,J.N.(Eds),AdvancesinEducationalandPsychological Testing.

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