Journal of Research in Teaching فصلنامه علمی تدریس پژوهی

Vol 13, No 4, Winter , 2025 صفحات: ۹۳-۶۸ سال سیزدهم- شمارهٔ چهارم – زمستان ۱٤۰٤

Comparison of the Psychometric Properties of Essay and Multiple-Choice Questions in Math and Science for Sixth-Grade Students Based on Classical Test Theory and Item Response Theory

Afshin afzali*[1]، Abolghasem Yaghoobi،[2] Mohammad aref Pilehvarpour[3] ،Kazhal Azizi[4]،Darya Ghanei[5]

دریافت مقاله: ۱٤۰۳/۰٦/۲٥ پذیرش مقاله: ۱٤۰٤/۰۹/۰٦

Received Date: 2024/09/15 **Accepted Date: 2025/11/27**

Abstract

Introduction: This study analyzes the psychometric properties of essay and multiple-choice questions in math and science for sixth-grade students using Classical Test Theory (CTT) and Item Response Theory (IRT). There are two prominent theories for analyzing test questions: Classical Test Theory (CTT) and Item Response Theory (IRT). In CTT, the unit of analysis is the entire test, while in IRT, the unit of analysis is each individual item. CTT has been a foundational theory in measurement for several decades, defined as a simple linear model stating that the observed score on a test is the sum of the true score and measurement error. This model consists of three components: the observed score, the true score, and the error score. The central idea regarding the relationship between the true score, observed score, and measurement error provides CTT with the ability to explain factors affecting test scores. CTT is based on three assumptions: first, the correlation between error scores and true scores is zero; second, errors have a mean of zero; and third, measurements of parallel tests are uncorrelated. CTT has been used for decades as a model for assessing the reliability and validity of measurement tools. According to the literature, CTT involves three main concepts: (a) the test score, also known as the observed score, (b) the true score, and (c) error scores. CTT focuses on two main aspects: item difficulty and item discrimination. Item difficulty refers to the proportion of individuals who can correctly answer the question. Generally, the more difficult the question, the lower

[1] Associate Professor, Department of Psychology, Faculty of Economics and Social Sciences, Bu-Ali Sina University, Hamedan, Iran

*** (Corresponding Author): Email: Afzali.Afshin@basu.ac.ir**

[2] Professor, Department of Psychology, Faculty of Economics and Social Sciences, Bu-Ali Sina University, Hamedan, Iran

[۳] .M.A educational psychology, Bu-Ali Sina University, Hamedan, Iran.

[٤]. M.A educational psychology, Bu-Ali Sina University, Hamedan, Iran.

[٥]. M.A educational psychology, Bu-Ali Sina University, Hamedan, Iran.

the percentage of individuals answering correctly. The primary index for measuring item difficulty is the difficulty index. On the other hand, item discrimination refers to the ability of an item to differentiate between "high-performing" and "low-performing" individuals. IRT is based on the assumption that the abilities of one or more participants, denoted by θ (theta), are predictable. In Item Response Theory (IRT), important parameters for each item are defined: the discrimination parameter (a), the difficulty parameter (b), and the guessing parameter (c).

Method: The sample consisted of 388 sixth-grade students from Hamadan, selected using cluster sampling. Given the study's objective to evaluate the performance of multiple-choice and essay questions in science and math, a survey approach was utilized with methods based on CTT and IRT. The study population included all sixth-grade students in Hamadan during the 2023–2024 academic year, with a sample size of 388 determined using the Morgan table. This sample was selected randomly from six schools (three girls' schools and three boys' schools). To collect data, two teacher-made tests for science and math, containing both multiple-choice and essay questions, were used. To assess validity, each test was reviewed by four teachers (with at least six years of teaching experience) and then piloted. After incorporating the experts' feedback, the tests were finalized and used for data collection. A grading (partial credit) method was used for scoring the essay questions (Saif, 2016). To analyze the results, the e-IRT software was utilized. Parameters for multiple-choice questions were estimated based on the three-parameter model, while essay questions were analyzed using the Graded Response Model.

Results: Results indicated that essay questions performed better than multiple-choice questions in both science and math. Specifically, for essay questions, the average discrimination index in science was 0.208 and in math was 0.55, while the average difficulty index in science was 2.591 and in math was 2.342, reflecting better discrimination and difficulty for essay questions. Additionally, analysis of essay questions using IRT showed that all four questions in math had a discrimination parameter above 1.35. In science, the discrimination values for the questions were 0.087, 1.090, 0.844, 1.419, and 0.533, respectively. Furthermore, the threshold parameter showed positive changes at each step in both science and math, indicating better discrimination and threshold functioning of essay questions.

The better performance of descriptive questions compared to multiple-choice questions can be attributed to several factors. Descriptive questions allow students to explore topics in depth and demonstrate critical and analytical thinking abilities. They are particularly suitable for assessing higher-order skills such as analysis, evaluation, and synthesis of information (Anderson, 2001). Unlike multiple-choice questions that restrict students to selecting one option, descriptive questions enable them to express their ideas in detail and creatively (Biggs, 2011). They also allow assessment of complex or multifaceted topics by enabling students to tailor responses based on prior knowledge and experience (Moon, 2006).Despite these advantages, descriptive questions are used less frequently for several reasons. Scoring them is time-consuming and may involve human error, leading to scorer variability (Brown, 2013). They may also demonstrate lower reliability because responses can be influenced by writing ability, fatigue, or time constraints (Gipps, 1994). Additionally, descriptive questions usually cover fewer content areas and may not provide a comprehensive

assessment of the entire curriculum (Race, 2014).This study has limitations that should be considered when interpreting the findings. The types of questions used may not have fully reflected all aspects of students' abilities. Although CTT and IRT provided valuable psychometric information, they may not have captured all complex aspects of item characteristics. The study focused only on math and science; therefore, generalization to other subjects should be made cautiously. Moreover, factors such as testing conditions and student stress were not fully controlled and may have influenced the results.

Keywords: Psychometric properties, Classical Test Theory, Item Response Theory, Discrimination index, Threshold index, Detection index, Guessing index

مقایسه ویژگی‌های روان‌سنجی سؤالات تشریحی و چندگزینه‌ای در دروس ریاضی و علوم پایه ششم آموزش ابتدایی، بر اساس تئوری‌های کلاسیک و سؤال-پاسخ

افشین افضلی *[۱]،ابوالقاسم یعقوبی [۲]،محمد عارف پیله ورپور [۳]،کژال عزیزی [۴]،دریا قانعی [۵]

چکیده

هدف:این پژوهش با هدف تحلیل و مقایسه ویژگی‌های روان‌سنجی سؤالات تشریحی و چندگزینه‌ای در دروس ریاضی و علوم پایه ششم ابتدایی بر اساس تئوری کلاسیک آزمون و تئوری سؤال–پاسخ انجام شد.

روش: پژوهش با روش توصیفی و بر روی ۳۸۸ دانش‌آموز پایه ششم ابتدایی شهر همدان که به‌صورت تصادفی خوشه‌ای انتخاب شدند، انجام شد. داده‌ها از طریق دو آزمون معلم‌ساخته علوم و ریاضی (شامل سؤالات تشریحی و چندگزینه‌ای) گردآوری و بر اساس تئوری کلاسیک آزمون و تئوری سؤال–پاسخ با استفاده از افزونه e-IRT تحلیل شدند.

یافته ها:نتایج نشان داد سؤالات تشریحی در هر دو درس ریاضی و علوم نسبت به سؤالات چندگزینه‌ای عملکرد بهتری دارند. میانگین ضریب تمیز در سؤالات تشریحی برای درس علوم ۰/۲۰۸و برای درس ریاضی ۰/۵۵و میانگین ضریب دشواری به‌ترتیب ۲/۵۹۱ و ۲/۳۴۲ه‌دست آمد. این یافته‌ها حاکی از آن است که سؤالات تشریحی در سنجش و تمایز توانایی دانش‌آموزان اثربخش‌تر عمل می‌کنند.

کلیدواژه‌ها: ویژگی های روانسنجی،تئوری کلاسیک،تئوری سوال_پاسخ،ضریب تمیز،ضریب آستانه

[۱].دانشیار گروه روانشناسی، دانشکده علوم اقتصادی و اجتماعی،دانشگاه بوعلی سینا، همدان، ایران

*(نویسنده مسئول) : **Email:Afzali.Afshin@basu.ac.ir**

[۲].استاد گروه روانشناسی، دانشکده علوم اقتصادی و اجتماعی،دانشگاه بوعلی سینا، همدان، ایران

[۳]کارشناسی ارشد روان شناسی تربیتی، دانشگاه بوعلی سینا، همدان، ایران.

[۴]کارشناسی ارشد روان شناسی تربیتی، دانشگاه بوعلی سینا، همدان، ایران.

[۵].کارشناسی ارشد روان شناسی تربیتی، دانشگاه بوعلی سینا، همدان، ایران.

مقدمه

ارزشیابی آموزشی به عنوان فرآیندی نظام‌مند برای تعیین میزان تحقق اهداف یادگیری، یکی از ارکان اساسی نظام تعلیم و تربیت به شمار می‌رود. از طریق ارزشیابی و اندازه‌گیری، معلمان می‌توانند تصویری روشن از نقاط قوت و ضعف دانش‌آموزان به دست آورند و بر اساس آن، تصمیمات آموزشی مناسب اتخاذ کنند؛ به گونه‌ای که بدون سنجش دقیق، شناسایی نیازها و توانایی‌های یادگیرندگان امکان‌پذیر نخواهد بود (Kusumawati , 2018).

در میان ابزارهای گوناگون سنجش، آزمون‌ها پرکاربردترین شیوه ارزیابی پیشرفت تحصیلی محسوب می‌شوند. آزمون‌ها امکان سنجش میزان تسلط دانش‌آموزان بر دانش و مهارت‌های مشخص را فراهم می‌کنند و در تصمیم‌گیربا این حال، اعتبار این تصمیمات زمانی تضمین می‌شود که سؤالات آزمون از کیفیت روان‌سنجی مطلوب برخوردار باشند؛ یعنی از نظر دشواری، قدرت تمیز و کارایی گزینه‌های انحرافی در سطح مناسبی قرار گیرنداز این‌رو، تحلیل ویژگی‌های روان‌سنجی سؤالات پس از اجرای آزمون، ضرورتی انکارناپذیر در تضمین روایی و پایایی نتایج است(Alemu , 2024) .

در حوزه تحلیل سؤالات، دو چارچوب نظری عمده یعنی نظریه آزمون کلاسیک [1](CTT) و نظریه سؤال–پاسخ [2](IRT) بیشترین کاربرد را دارند. نظریه کلاسیک که سابقه‌ای طولانی در سنجش دارد، بر رابطه نمره مشاهده‌شده، نمره واقعی و خطای اندازه‌گیری استوار است و شاخص‌هایی چون دشواری و تمیز را برای تحلیل سؤالات به کار می‌گیرد در مقابل، نظریه سؤال–پاسخ با بهره‌گیری از مدل‌های ریاضی، رابطه بین توانایی نهفته آزمودنی و ویژگی‌های هر سؤال را به‌صورت جداگانه مدل‌سازی می‌کند و پارامترهایی چون تمیز(a) ، دشواری (b) و حدس (c) را برآورد می‌نماید بسیاری از پژوهشگران بر مزایای IRT در ارائه اطلاعات دقیق‌تر و کاهش خطای اندازه‌گیری تأکید کرده‌اندبا این حال، همچنان درباره برتری مطلق یکی از این دو رویکرد اجماع کامل وجود ندارد.(Oladele & Adegoke, 2020)

ضرورت پرداختن به این موضوع زمانی برجسته‌تر می‌شود که نتایج مطالعات بین‌المللی مانند نشان می‌دهد عملکرد دانش‌آموزان ایرانی در ریاضی و علوم پایین‌تر از میانگین جهانی است . این وضعیت می‌تواند ناشی از عوامل گوناگون آموزشی باشد، اما بخشی از آن ممکن است به کیفیت طراحی و تحلیل سؤالات آزمون‌ها بازگردد. مرور پژوهش‌های پیشین نشان می‌دهد که اغلب مطالعات، یا صرفاً بر سؤالات چندگزینه‌ای تمرکز داشته‌اند یا یک درس خاص را بررسی کرده‌اند ؛ در نتیجه، مقایسه هم‌زمان ویژگی‌های روان‌سنجی سؤالات تشریحی و چندگزینه‌ای در دو درس بنیادین ریاضی و علوم کمتر مورد توجه قرار گرفته است(Mullis et al.,2020)

بر این اساس، مسئله اصلی پژوهش حاضر آن است که آیا سؤالات تشریحی و چندگزینه‌ای در دروس ریاضی و علوم از نظر ویژگی‌های روان‌سنجی در تئوری کلاسیک و سوال_ پاسخ (دشواری، تمیز و سایر پارامترهای مرتبط) عملکرد متفاوتی دارند و کدام‌یک در سنجش دقیق‌تر توانایی‌های دانش‌آموزان

[1] Classical Test Theory
[2] Item Response Theory

کارآمدترند. پاسخ به این پرسش می‌تواند شواهد علمی معتبری برای بهبود طراحی آزمون‌ها، ارتقای کیفیت ارزشیابی‌های مدرسه‌ای و اتخاذ تصمیمات آموزشی آگاهانه‌تر فراهم آورد.

چارچوب مفهومی

ارزشیابی جزء مهمی از برنامه درسی آموزش و یادگیری است و یکی از کاربردهای مهم آن ، نظارت مداوم بر فعالیت‌های یادگیری و ارائه بازخورد به دانش‌آموزان و معلمان است , Mehta & Mokhasi) (2014که بهتر است توسط معلمان به‌طور مداوم برای نظارت بر فرآیند پیشرفت و بهبود نتایج یادگیری دانش‌آموزان انجام شود (Maba et al , 2017) . هدف از ارزشیابی این است که به دانش‌آموزان اجازه داده شود تا آنچه را که آموخته‌اند نشان دهند و پیشرفت یادگیری را در طول زمان شناسایی کنند، دانش‌آموزان را انگیزه دهند و آن‌ها را در رتبه‌بندی‌های کلاسی طبقه‌بندی کنند Fernandez et al) (2019. ارزشیابی آموزشی شامل فرآیند تشریح کسب و تسلط بر دانش به‌منظور کمک به اتخاذ تصمیمات آگاهانه درباره مراحل پیش‌روی در فرآیند آموزشی است (Ravela et al,2009) . در این فرآیند، به مواردی چون توانایی ها ، سبک‌های یادگیری، نگرش‌ها، پیشرفت‌ها و نتایج دانش‌آموزان توجه می‌شود. تصمیماتی که پس از ارزشیابی اتخاذ می‌شود ممکن است متفاوت باشد؛ از اجرای برنامه‌های سیستمی گسترده تا بهبود تدریس و یادگیری در کلاس، تغییر در روش‌های تدریس در کلاس، یا ارزیابی پذیرش دانش‌آموزان به مقاطع بالاتر آموزشی مانند دانشگاه (Clarke,2011) . سیستم‌های ارزیابی آموزشی مؤثر به کسب اطلاعات با کیفیت برای پاسخگویی به نیازهای تصمیم‌گیری و حمایت و بهبود یادگیری دانش‌آموزان کمک می‌کنند (Butakor , 2022) .سنجش[1] و ارزشیابی جز لا ینفک تعلیم و تربیت بشمار میرود که بدون استمرار دقیق آن، رسیدن به اهداف مورد نظر بصورت مطلوب، نا ممکن خواهد بود.

یکی از پر کاربرد ترین ابزار های سنجش در آموزش ،برگزاری آزمون[2] است.آزمون ها روش هایی هستند که برای تعیین توانایی دانش آموزان در انجام وظایف خاص یا نشان دادن تسلط بر یک مهارت یا دانش محتوا استفاده می شوند (Adom et al ,2020). آزمون دادن به دانش‌آموزان برای بهبود ارتباطات، انگیزه‌بخشی به آن‌ها برای تلاش بیشتر، شناسایی نیازهای آموزشی اصلاحی، تصمیم‌گیری در مورد ارتقاء و شناسایی مشکلات مرتبط با برنامه درسی اهمیت دارد و آزمون‌ها نقش کلیدی در ارزشیابی آموزشی دارند. برای ارزیابی میزان دانش و مهارت‌های دانش‌آموزان در یک زمینه تحصیلی، باید از سؤالات آزمونی که به‌خوبی طراحی شده‌اند استفاده شود(Quansah et al , 2018) .

1. Assessment
2. Test

ویژگی‌های داخلی یک آزمون و سؤالات آن به طور فنی به عنوان خصوصیات یا ویژگی‌های روان‌سنجی آن شناخته می‌شوند. ویژگی‌های روان‌سنجی سؤالات یک آزمون به ویژگی‌های خاصی در آزمون اشاره دارد که ارزیابی شرایط بر اساس آن‌ها صورت می‌گیرد. شاخص‌های سختی یک سوال خاص در آزمون، توانایی آن در تمایز بین افرادی که سطوح مختلفی از سازه مورد اندازه‌گیری را دارند، و منطقی بودن گزینه‌های انحرافی، هر یک ویژگی روان‌سنجی سوال به شمار می‌روند. بررسی و شناخت ویژگی‌های روان‌سنجی آزمون‌ها بسیار مهم است. روان‌سنجان و دیگر متخصصانی که آزمون‌ها را طراحی می‌کنند، باید نحوه عملکرد یک آزمون را ارزیابی و توصیف کنند تا بتوانند آن را به سطح مشخصی از کیفیت برسانند. همچنین، آگاهی از ویژگی‌های روان‌سنجی یک آزمون و سؤالات آن، شواهدی فراهم می‌کند که نشان می دهد اطلاعات به‌دست‌آمده از طریق چنین آزمونی می‌تواند مبنای قوی برای تصمیم‌گیری باشد) .Chukwu Ohiri, 2023)

پس از برنامه‌ریزی، توسعه و اجرای دقیق آزمون، انجام تحلیل روان‌سنجی بر روی داده‌های حاصل از آزمون ضروری است تا شواهدی برای ارزیابی کیفیت سؤالات و قضاوت در مورد اعتبار و پایایی آن‌ها فراهم شود. برای این منظور، چارچوب‌های نظری مختلفی در حوزه سنجش و اندازه‌گیری توسعه یافته‌اند که امکان بررسی دقیق‌تر ویژگی‌های آزمون و سؤالات آن را فراهم می‌سازند .(Bichi, 2016) این چارچوب‌ها معمولاً متغیرهای مشاهده‌پذیر مانند نمرات آزمون یا نمرات هر سؤال را به متغیرهای نهفته همچون توانایی واقعی یا ویژگی‌های پنهان آزمودنی‌ها مرتبط می‌کنند و در میان رویکردهای موجود، دو نظریه اصلی بیشترین کاربرد را در تحلیل سؤالات آزمون دارند :نظریه آزمون کلاسیک (CTT) و نظریه سؤال–پاسخ .(IRT)نظریه CTT بیشتر بر تحلیل کلی آزمون تمرکز دارد، در حالی که نظریه IRT امکان بررسی ویژگی‌های هر سؤال به‌طور جداگانه و در ارتباط با سطح توانایی آزمودنی‌ها را فراهم می‌سازد. در حالی که در نظریه IRT، واحد تحلیل هر سوال (مورد) است (Hambleton et al., 1991؛ Baker, 2001؛ McAlpine, 2002). این روش‌های تحلیل عمدتاً به منظور اطمینان از کیفیت ارزیابی به کار می‌روند، به این معنی که تضمین می‌کنند سؤالات دارای سطح دشواری مناسب هستند و به درستی توانایی تمایز میان دانشجویان با سطوح مختلف عملکرد را دارند، به‌طوری که قادر به تشخیص تفاوت میان "بهترین" و "ضعیف‌ترین" دانشجویان باشند. هر دو نظریه در مورد اینکه کدامیک از دیگری برتر است، در رقابتند و این موضوع همچنان منبع بحث و جدل مداوم میان حامیان هر کدام از نظریه‌ها است (McAlpine, 2002).تئوری آزمون کلاسیک (CTT) از چندین دهه پیش، بنیانگذاری برای نظریه اندازه‌گیری بوده است. تئوری آزمون کلاسیک توسط متخصصان به این معنا تعریف شده است که یک مدل خطی ساده است که بیان می‌کند نمره مشاهده‌شده در یک آزمون مجموع نمره واقعی و خطای اندازه‌گیری است. این مدل خطی ساده از سه جزء تشکیل شده است: نمره مشاهده‌شده، نمره واقعی و امتیاز خطا. این ایده مرکزی درباره رابطه بین نمره واقعی، نمره مشاهده‌شده و خطای اندازه‌گیری، توانایی تئوری آزمون کلاسیک

را فراهم می‌کند که عواملی که بر نمرات آزمون تأثیر می‌گذارند را شرح می‌دهد. در تئوری آزمون کلاسیک (CTT)، سه فرضیه در نظر گرفته می‌شود. اول، همبستگی بین نمره خطا و نمره واقعی برابر صفر است. فرضیه دوم می‌گوید که خطاها ، میانگین صفر دارند. فرضیه سوم این است که اندازه‌گیری‌های ازمون موازی‌ها ، بی‌ارتباط هستند. نظریه کلاسیک آزمون (CTT) دهه‌ها به عنوان مدلی برای ارزیابی قابلیت اطمینان و اعتبار ابزارهای اندازه‌گیری استفاده شده است. بر اساس منابع، CTTنظریه نمره آزمون است که سه مفهوم اصلی را با خود به همراه دارد: (الف) نمره آزمون که به عنوان نمره مشاهده شده نیز شناخته می‌شود، (ب) نمره واقعی و (ج) نمره‌های خطا (oladele & adegoke , 2020) .

اگرچه تمرکز اصلی نظریه آزمون کلاسیک بر اطلاعات مربوط به آزمون است، اما آماره‌های سوال که نشان‌دهنده سختی سوال[1] (معمولاً با نماد p نشان داده می‌شود) و تمایز سوال[2] (معمولاً با نماد D یا r نشان داده می‌شود)هستند، نیز توسعه داده شده و بخش‌های مهمی از این مدل هستند. این دو آماره اصلی سوال در تحلیل سوال و انتخاب سوال در توسعه آزمون‌ها استفاده می‌شوند (Adegoke, 2013) . CTT بر دو جنبه اصلی متمرکز است: دشواری و تمایز سؤالات. دشواری یک سوال به تعداد افرادی که قادرند به درستی به آن پاسخ دهند بستگی دارد. به طور ساده، هرچه سوال دشوارتر باشد، درصد افرادی که به درستی به آن پاسخ می‌دهند، کمتر خواهد بود. شاخص اصلی برای اندازه‌گیری دشواری سوال، شاخص دشواری است. از سوی دیگر، تمایز سوال به قابلیت آن در تفکیک "بهترین" از "ضعیف‌ترین" افراد مربوط می‌شود. به عبارت دیگر، هرچه تمایز سوال بیشتر باشد، تعداد افرادی که در گروه "بهترین" قرار دارند و به درستی به آن پاسخ می‌دهند، بیشتر خواهد بود و تعداد افرادی که در گروه "ضعیف‌ترین" هستند و به درستی به آن پاسخ می‌دهند، کمتر خواهد بود. شاخص اصلی برای اندازه‌گیری تمایز سوال، شاخص تمایز است(Azevedo et al , 2019) . تئوری آزمون کلاسیک و مدل‌های مرتبط بیش از ۶۰ سال است که مورد تحقیق و به کارگیری قرار گرفته‌اند و بسیاری از برنامه‌های آزمون امروزی همچنان بر اساس مدل‌ها و روش‌های اندازه‌گیری کلاسیک پایدار هستند (Oladele & Adegoke, 2020) .تا اوایل دهه ۱۹۷۰، نظریه آزمون کلاسیک (CTT) به عنوان چارچوب اصلی برای تحلیل و توسعه آزمون‌های استاندارد شناخته می‌شد. اما از آن زمان به بعد، نظریه سوال پاسخ به آیتم به طور کامل جایگزین نقش CTT شده و اکنون به عنوان چارچوب نظری اصلی در این زمینه علمی شناخته می‌شود . (ERGUVEN , 2013) تئوری سوال پاسخ (IRT) یک مدل آماری است که عملکرد آزمون و سوال های آزمون را توصیف می‌کند و توضیح می‌دهد که چگونه نتایج آزمون با توانایی‌های منعکس شده در سوال های آزمون مرتبط هستند (Embretson & Reise, 2013). نظریه سوال پاسخ (IRT) از توابع ریاضی استفاده می‌کند، برخلاف نظریه کلاسیک آزمون (CTT) که از مدل X=T+E بهره می‌برد.

[1] threshold parameters

[2] discrimination parameters

بر اساس مطالعات (Hambleton & Swaminathan, 1985)، IRT با یک رابطه دقیق بین پاسخ‌ها و ویژگی‌ها مشخص می‌شوند. علاوه بر این، IRT بر این فرض استوار است که توانایی‌های یک یا چند شرکت‌کننده از تتا (θ) که یکی از پارامترها را تشکیل می‌دهد، قابل پیش‌بینی است. (Ayanwale , et al , 2022). در نظریه پاسخ به آیتم(IRT) ، پارامترهای مهمی برای هر آیتم تعریف می‌شوند. این پارامترها عبارتند از:پارامتر تمایز :(a) این پارامتر نشان می‌دهد که چقدر یک آیتم قادر است بین دانش‌آموزان ماهر و کم‌تر ماهر تمایز بیانجامد. هرچه مقدار این پارامتر بیشتر باشد، آیتم توانایی بیشتری در تمایز دادن بین دانش‌آموزان دارد.پارامتر سختی :(b) این پارامتر نشان می‌دهد که یک آیتم چقدر سخت یا آسان است. اگر بی پارامتر برای یک آیتم بزرگ باشد، آن آیتم برای دانش‌آموزان با توانایی پایین سخت خواهد بود و برعکس.پارامتر حدس زدن :(c) این پارامتر نشان می‌دهد که چقدر احتمال دارد که یک دانش‌آموز با حدس زدن به درستی به یک آیتم پاسخ دهد. این پارامتر معمولاً برای آیتم‌های دوگانه (بله/خیر) مورد استفاده قرار می‌گیرد(ERGUVEN , 2013) ..در نظریه پاسخ به سؤال (IRT) سه مدل لجستیک وجود دارد: مدل لجستیک یک‌پارامتره[1] (PL۱)، مدل لجستیک دوپارامتره (PL۲) و مدل لجستیک سه‌پارامتره (PL۳). این مدل‌ها بر اساس تعداد پارامترهایی که برای توصیف ویژگی‌های هر سوال استفاده می‌شوند، از یکدیگر متمایز می‌شوند. پارامترهای سؤالات شامل شاخص سختی سوال (b)، شاخص تمایز سوال (a) و پارامتر حدس کاذب (c) هستند. این سه عنصر به‌گونه‌ای با یکدیگر مرتبط هستند که باعث شکل‌گیری تابع یا منحنی پاسخ می‌شود که به آن منحنی ویژگی‌های سؤالات (ICC) گفته می‌شود (Kusumawati , 2018)..بر اساس(Hambleton & et al ,1991)، مدل‌های (IRT) بر دو اصل پایه‌ای استوارند :۱ . عملکرد یک آزمون‌دهنده در یک آزمون می‌تواند پیش‌بینی یا توضیح داده شود از طریق مجموعه‌ای از ویژگی‌ها که به آنها ویژگی‌های پنهان گفته می‌شود؛ و ۲ .رابطه بین عملکرد آزمون‌دهنده و مجموعه‌ای از ویژگی‌های یک سؤال، که مبنای عملکرد آن است، می‌تواند از طریق یک تابع افزایشی و یکنواخت توصیف شود که به آن تابع ویژگی سؤال یا منحنی ویژگی سؤال (ICC) گفته می‌شود.پارامترهای آیتم در نظریه پاسخ به آیتم (IRT) می‌توانند به شرطی برآورد شوند که مدل آماری استفاده‌شده فرضیات را برآورده کند. این فرضیات شامل یک نواختی بعدی، استقلال محلی و تغییر ناپذیری پارامترها هستند. برای انجام تحلیل IRT، باید این فرضیات برآورده شوند. در صورتی که فرضیات برآورده نشوند، تحلیل انجام‌شده همان نظریه کلاسیک آزمون‌ها (CTT) خواهد بود. برآورده شدن فرضیات بر اساس کیفیت ابزار آزمون است؛ بنابراین، توسعه‌دهنده آزمون باید دانش کافی داشته باشد تا سؤالات تولیدشده از نظر کیفیت محتوا نیز مناسب باشند(Hambleton & Swaminathan, 1985; Santoso et al., 2022)).

[1] .One-Parameter Logistic Model

تحقیقات متعدد تأیید کرده‌اند که چارچوب IRT مزایای زیادی را ارائه می‌دهد که توجه مؤسسات ارزیابی آموزشی، توسعه‌دهندگان آزمون و سیاست‌گذاران در صنعت ارزیابی را جلب کرده است و آن‌ها از این نظریه برای اتخاذ تصمیمات معتبر و قابل اعتماد استفاده کرده‌اند.(Ayanwale et al., 2019)

در مقایسه مدل‌های CTT و IRT در توسعه آزمون، IRTمجموعه‌ای غنی‌تر از ابزارها را برای توسعه آزمون فراهم می‌کند. IRT یک پارامتر سوم (شبه حدس‌زدن) را فراهم می‌کند که در CTT معادل ندارد. IRT همچنین ابزاری برای ارزیابی درجه معادل بودن اندازه‌گیری در نقاط مختلف مقیاس نمره، بر اساس مجموعه‌های مختلف سوال ها ، فراهم می‌کند. بنابراین، نتایج تحلیل سوال ها ارائه شده توسط هر دو CTTو IRT تقریباً قابل مقایسه هستند، اما IRT آمارهای سوال اضافی و مکانیزم پیچیده‌تری برای کاهش خطای اندازه‌گیری فراهم می‌کند.در مدل‌هایIRT ، پارامترهای سوال ها برای کل جمعیت ثابت فرض می‌شوند و نهایتاً مستقل از نمونه‌های خاص افراد هستند. این به این معنی است که می‌توان با استفاده از مجموعه‌های مختلفی از سوال ها ، به افراد مختلف امتیازات تقریباً مشابهی اختصاص داد.از طرفی، درCTT ، پارامترهای سوال و ویژگی‌های آزمون به‌طور وابسته به نمونه خاص مشخص می‌شوند و قابلیت جدا کردن دقیق این دو مفهوم وجود ندارد که در IRT امکان‌پذیر است. به عبارت دیگر، مدل‌های IRTبه خوبی توانایی آزمون‌دهنده‌ها و مراحل دشواری سوال ها را به‌طور مستقل شناسایی می‌کنند، در حالی که CTT این امکان را ندارد و سوال هایی که در نمونه‌های با توانایی متفاوت استفاده می‌شوند، به گونه‌ای تفسیر می‌شوند که در یک جامعه با توانایی بالا به‌طور آسان و در افراد با توانایی پایین به‌طور دشوار معنی می‌دهند.تخمین توانایی آزمون‌دهنده در مدل‌های نظریه سوال پاسخ (IRT) بستگی به پاسخ‌های او دارد که معمولاً اطلاعات کاملی را ارائه می‌دهند. توانایی به عنوان یک متغیر پیوسته در نظر گرفته می‌شود و IRT تخمین‌های پیوسته را ارائه می‌دهد. در مقابل، نظریه کلاسیک (CTT)، به ویژه در آزمون‌های دودویی، تخمین‌های گسسته‌ای ارائه می‌دهد که ممکن است منجر به اختلافات در ارزیابی پیشرفت دانش‌آموزان با امتیازهای کلی نهایی شود. به این معنی که تخمین‌های توانایی دانش‌آموزان به وسیله IRT ممکن است از نتایج کلی امتیازهای خام CTT متمایز باشند (bichi & talib, 2018).

پیشینه پژوهش

(Oghenerume& Egberha , 2024) در پژوهشی به تحلیل مقایسه‌ای ویژگی‌های آماری سؤالات آزمون‌های چندگزینه‌ای WASSCE و NECO SSCE سال ۲۰۲۳ با استفاده از مدل سه‌پارامتری نظریه سؤال–پاسخ (3PLM) پرداخته شد. نتایج نشان داد که تفاوت معنی‌داری در پارامترهای دشواری و تمایز بین سؤالات دو آزمون وجود ندارد، اما در پارامتر حدس تفاوت معناداری مشاهده شد و در نتیجه پیشنهاد گردید که استفاده از مدل سه‌پارامتری IRT می‌تواند به بهبود کیفیت سؤالات چندگزینه‌ای کمک کند .(Hamidah & Istiyono, 2022) همچنین در پژوهشی دیگر، کیفیت سؤالات چندگزینه‌ای آزمون

ملی علوم طبیعی در مدارس ابتدایی با استفاده از مدل دوپارامتری نظریه سؤال–پاسخ بررسی شد و نتایج نشان داد که سؤالات از نظر سطح دشواری و قدرت تمایز در سطح مناسبی قرار دارند (Cobbinah & Ntumi, 2022). در مطالعه‌ای دیگر نیز بر روی آیتم‌های آزمون‌های ریاضی شورای امتحانات غرب آفریقا در غنا، با استفاده از نظریه سؤال–پاسخ (IRT) نشان داده شد که سؤالات آزمون‌های سال ۲۰۲۰ از نظر دشواری در سطح قابل قبولی قرار دارند و توانسته‌اند به خوبی بین داوطلبان تمایز قائل شوند.مرور این مطالعات نشان می‌دهد که بیشتر پژوهش‌های پیشین عمدتاً بر آزمون‌های چندگزینه‌ای و آن هم در چارچوب یک درس یا یک مقطع آموزشی متمرکز بوده‌اند. همچنین تمرکز اصلی این تحقیقات بیشتر بر تحلیل آماری ویژگی‌های سؤالات (مانند دشواری، تمیز و حدس) بوده و کمتر به مقایسه‌ی انواع مختلف سؤالات (تشریحی و چندگزینه‌ای) در دروس متفاوت پرداخته‌اند. از این رو، تاکنون تحلیل روان‌سنجی تلفیقی سؤالات تشریحی و چندگزینه‌ای در دروسی مانند ریاضی و علوم کمتر مورد توجه قرار گرفته است. پژوهش حاضر با تمرکز بر این خلأ، می‌کوشد تصویری جامع‌تر از ویژگی‌های روان‌سنجی سؤالات ارائه دهد تا هم برای طراحان آزمون و هم برای سیاست‌گذاران آموزشی کاربردی باشد.

پژوهش حاضر با هدف بررسی عملکرد ویژگی‌های روان‌سنجی سؤالات تشریحی و چندگزینه‌ای در دروس علوم و ریاضیات انجام گرفت. مرور پیشینه‌های داخلی و خارجی نشان می‌دهد که هرچند پژوهش‌های متعددی به تحلیل ویژگی‌های روان‌سنجی سؤالات پرداخته‌اند، اغلب آن‌ها محدود به سؤالات چندگزینه‌ای یا یک درس خاص بوده‌اند و کمتر مطالعه‌ای به مقایسه‌ی هم‌زمان سؤالات تشریحی و چندگزینه‌ای در دو درس علوم و ریاضی توجه کرده است. بنابراین، گره اصلی این موضوع در آن است که تاکنون تصویر جامعی از تفاوت‌ها و شباهت‌های ویژگی‌های روان‌سنجی این دو نوع سؤال در این دروس ارائه نشده است. پژوهش حاضر با هدف پر کردن این خلأ و فراهم‌سازی شواهدی دقیق‌تر برای طراحان آزمون و سیاست‌گذاران آموزشی انجام شده است.

روش شناسی پژوهش

پژوهش حاضر از نوع توصیفی است. جامعه آماری شامل کلیه دانش‌آموزان پایه ششم ابتدایی شهر همدان در سال تحصیلی ۱۴۰۲-۱۴۰۳ بود. بر اساس جدول مورگان، ۳۸۸ نفر به‌عنوان نمونه انتخاب شدند که به شیوه تصادفی خوشه‌ای از میان شش مدرسه (سه مدرسه دخترانه و سه مدرسه پسرانه) گزینش گردیدند. ابزار گردآوری داده‌ها شامل دو آزمون معلم‌ساخته در درس‌های علوم و ریاضی بود که هرکدام دارای بخش‌های چندگزینه‌ای و تشریحی بودند؛ به‌گونه‌ای که آزمون علوم شامل ۷ سؤال چندگزینه‌ای و ۵ سؤال تشریحی و آزمون ریاضی شامل ۸ سؤال چندگزینه‌ای و ۴ سؤال تشریحی بود. در نمره‌گذاری بخش چندگزینه‌ای به پاسخ صحیح نمره ۱ و به پاسخ غلط نمره ۰ اختصاص داده شد. در بخش تشریحی از یک روبریک مدرج استفاده گردید که شامل چهار سطح بود: پاسخ کامل (۳ نمره)، پاسخ متوسط (۲

نمره)، پاسخ ضعیف (۱ نمره) و پاسخ کاملاً نادرست یا بدون پاسخ (۰ نمره). این روبریک پیش از اجرای اصلی طراحی و در اختیار چهار معلم با حداقل شش سال سابقه تدریس در پایه ششم قرار گرفت تا فرآیند نمره‌دهی یکسان‌سازی شود. برای اطمینان از روایی محتوایی نیز سؤالات آزمون‌ها در اختیار همین معلمان قرار گرفت و پس از اعمال اصلاحات پیشنهادی ایشان، ابزار نهایی مورد تأیید قرار گرفت و در مرحله اصلی پژوهش اجرا شد. داده‌ها پس از گردآوری در نرم‌افزار Excel وارد و با استفاده از افزونه eIRT تحلیل شدند. این افزونه امکان برازش مدل‌های مختلف نظریه سؤال‌پاسخ (IRT) را فراهم می‌سازد. در این پژوهش، سؤالات چندگزینه‌ای با استفاده از مدل سه‌پارامتری لاجیت (3PLM) و سؤالات تشریحی با استفاده از مدل پاسخ مدرج[1] (GRM) تحلیل گردیدند. به این ترتیب، شاخص‌های دشواری، تمیز و حدس‌زنی در سؤالات چندگزینه‌ای و عملکرد تدریجی پاسخ‌ها در سؤالات تشریحی بررسی شدند (Saif, 2016).

یافته‌ها

[1] graded response model

جدول ۱. مقایسه سؤالات تشریحی بر اساس تئوری سوال_پاسخ

		پارامترتمیز	آستانه گزینه های پاسخ					شاخص های برازش		
	سؤالات	A	B_0	B_1	B_2	B_3	B_4	X^2	df	P
ریاضی	۱	۱/۸۸	-۰/۸۸	-۰/۵۸	۰/۱۱	۰/۶۷	۰/۸۵	۳/۶۴	۴۰	۱
	۲	۱/۴۶	-۲/۵۷	-۲/۰۷	۱/۰۶	۰/۳۸	-۰/۲۱	۵/۰۰	۴۰	۱
	۳	۱/۴۷	-۱/۶۴	-۱/۲۹	۰/۴۱	۰/۳۵	۰/۵۸	۶/۱۹	۴۰	۱
	۴	۲/۰۹	-۱/۲۴	-۰/۸۱	۰/۰۱	۰/۷۳	۱/۱۱	۷/۴۸	۴۰	۱
علوم	۱	۰/۰۸	-۳۸/۶۱	-۲۵/۶۹	-۹/۵۵	۲/۸۴	۱۲/۰۰	۰/۶۸	۴۰	۱
	۲	۱/۰۹	-۲/۴۳	-۲/۰۲	-۱/۰۹	-۰/۲۰	۰/۱۷	۱/۱۳	۴۰	۱
	۳	۰/۸۴	-۲/۵۸	-۲/۲۲	-۱/۳۰	-۰/۴۸	-۰/۲۳	۲/۵۳	۴۰	۱
	۴	۱/۴۱	-۱/۲۹	-۰/۸۷	۰/۳۳	۱/۲۲	۱/۳۳	۰/۴۹	۴۰	۱
	۵	۰/۵۳	-۳/۱۵	-۲/۸۴	-۱/۷۸	-۰/۹۲	-۰/۸۱	۳/۱۵	۴۰	۱

Baker, Rounds, & Zevon (2000) مقدار پارامتر های تمیز کمتر از ۰/۶۵را پایین،۰/۶۵تا ۱/۳۴ به عنوان شاخص تمیز متوسط و ۱/۳۵بالاتر را تمیز بالا معرفی نموده اند. نتایج جدول فوق نشان می دهد که در درس ریاضی هر چهار سوال دارای پارامتر تمیز بالا هستند اما در درس علوم تنها سوال ۴ در بازه ی تمیز بالا قرار دارد و سؤالات۲ و۳ در بازه تمیز متوسط و سؤالات ۱و ۵ دارای تمیز ضعیف هستند.در پارامتر آستانه نیز در درس ریاضی در سؤالات ۱،۳و۴ تغییرات در هرگام به گام دیگر مثبت است که نشانگر عملکرد مناسب این سؤالات است در سوال ۲تمیز تا گام سوم روند تغییر مثبت است .در درس علوم در تمام سؤالات تقریبا روند تغییر مثبت در هرگام وجود دارد.لازم به ذکر است که در سؤالات ریاضی از گام سوم به بعد سطوح توانایی وارد محدوده مثبت می شود که نشانگر دشواری بالای این سؤالات است اما در درس علوم این موضوع فقط در سؤالات ۱و۴ وجود دارد. بررسی شاخص های برازش هم نشان می دهد که آستانه ی پاسخ برای همه سؤالات از مقدار برازش مناسبی برخوردار است (÷>۰/۰۵)

جدول ۲ . مقایسه سؤالات تشریحی بر اساس تئوری کلاسیک

علوم				ریاضی			
ضریب تمیز		انحراف استاندارد	ضریب تشخیص	ضریب تمیز		انحراف استاندارد	ضریب تشخیص
۱	۰/۰۴	۱/۱۷	۲/۶۱	۱	۰/۵۷	۱/۵۳	۱/۹۳
۲	۰/۲۶	۱/۳۶	۲/۷۹	۲	۰/۴۹	۱/۲۸	۲/۹۹
۳	۰/۲۳	۱/۴۵	۲/۸۴	۳	۰/۵۲	۱/۴۳	۲/۴۱
۴	۰/۳۲	۱/۳۴	۱/۸۵	۴	۰/۶۱	۱/۳۸	۲/۰۳
۵	۰/۱۸	۱/۵۶	۲/۸۴	۵			
میانگین	۰/۲۰		۲/۵۹	میانگین	۰/۵		۲/۳۴

با توجه به نتایج جدول فوق در سؤالات تشریحی درس علوم کمترین ضریب تمیز مربوط به سوال ۱ با تمیز۰۴۷ /۰ است. ضریب تمیز متوسط سؤالات تشریحی درس علوم ۰/۲۰۸ است. در درس ریاضی کمترین ضریب تمیز مربوط به سوال ۲ با تمیز۰/۴۹ است. متوسط ضریب تمیز سؤالات تشریحی در درس ریاضی ۰/۵۵ است.

جدول ۳ .مقایسه سؤالات چند گزینه ای از منظر تئوری سوال_پاسخ

	سؤالات	شیب تشخیص	آستانه، دشواری	حدس
ریاضی	۱	۱/۲۳	-۱/۰۹	۰/۱۵۷
	۲	۱/۲۴	-۲/۱۲	۰/۱۶۵
	۳	۲/۲۴	۰/۳۱	۰/۲۲۸
	۴	۱/۲۰	-۱/۸۲	۰/۱۷۸
	۵	۱/۹۶	-۱/۳۴	۰/۱۵۶
	۶	۱/۸۹	-۰/۳۴	۰/۱۳۹
	۷	۱/۲۹	-۲/۱۲	۰/۱۷۴
	۸	۱/۲۰	۰/۴۱	۰/۱۳۷
میانگین		۱/۵۵	-۰/۸۴	۰/۱۶۶
	۱	۲/۳۵	-۰/۹۱	۰/۱۵۶
	۲	۰/۴۷	-۵/۲۰	۰/۱۶۷
	۳	۰/۸۲	-۰/۲۹	۰/۱۵۸

علوم	۴	۰/۳۹	-۱/۲۰	۰/۱۷۲
	۵	۰/۵۱	-۴/۱۶	۰/۱۶۸
	۶	۰/۵۲	-۱/۳۰	۰/۱۷۰
	۷	۱/۷۷	-۱/۱۸	۰/۱۶۷
میانگین		۰/۹۷	-۲/۰۳	۰/۱۶۵

نتایج جدول فوق نشان می‌دهد که در درس علوم ضریب تشخیص سؤالات ۱ و ۷ قدرت تشخیص بالا و سوال ۳ قدرت تشخیص متوسط دارد ، تشخیص سایر سؤالات ضعیف است. در درس ریاضی سؤالات ۳ ،۴، ۵و ۶ دارای تشخیص بالا و سایر سؤالات دارای تشخیص متوسط هستند.

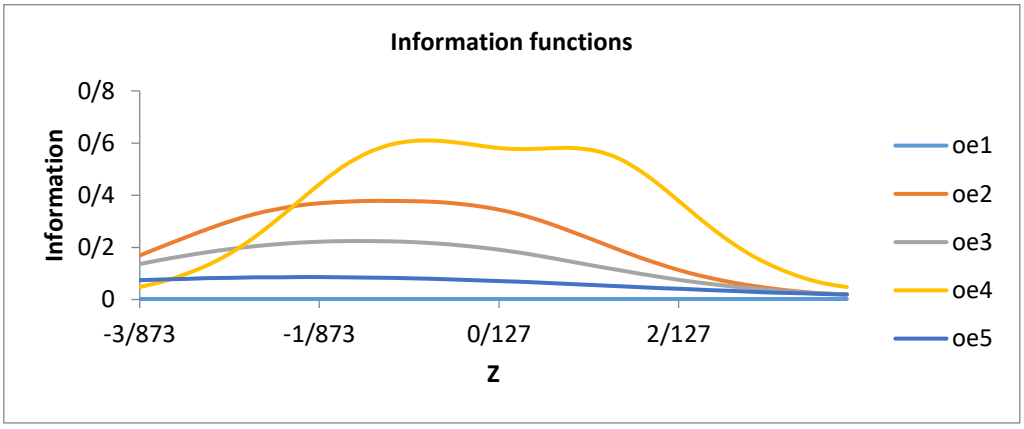
جدول۴. .مقایسه سؤالات چند گزینه ای از منظر تئوری کلاسیک

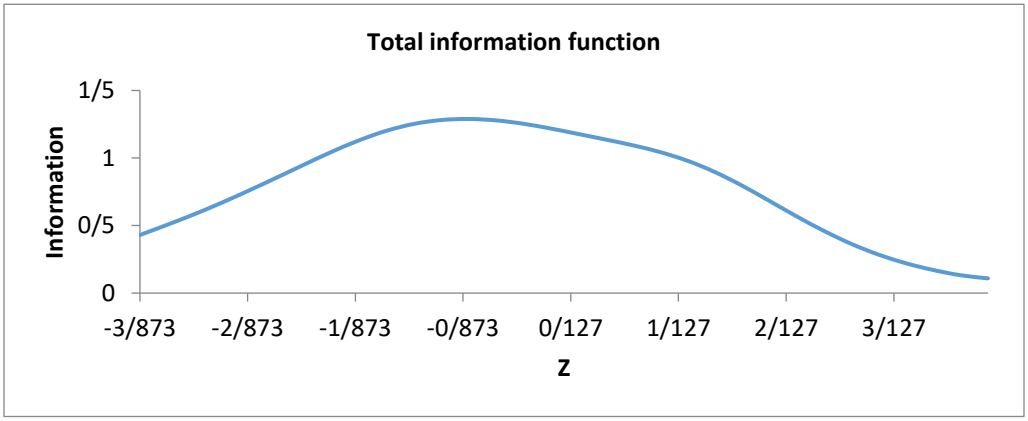
	سؤالات	ضریب دشواری	انحراف استاندارد	ضریب تشخیص
ریاضی	۱	۰/۸۷	۰/۴۱	۰/۳۱
	۲	۰/۹۱	۰/۲۸	۰/۲۵
	۳	۰/۵۴	۰/۴۹	۰/۳۲
	۴	۰/۸۹	۰/۳۰	۰/۲۸
	۵	۰/۸۶	۰/۳۴	۰/۳۶
	۶	۰/۶۵	۰/۴۷	۰/۴۰
	۷	۰/۹۱	۰/۲۷	۰/۲۲
	۸	۰/۴۸	۰/۵۰	۰/۳۰
میانگین		۰/۷۶		۰/۲۱
علوم	۱	۰/۸۰	۰/۳۹	۰/۲۴
	۲	۰/۹۲	۰/۲۵	۰/۰۵
	۳	۰/۶۱	۰/۴۸	۰/۲۱
	۴	۰/۶۸	۰/۴۶	۰/۱۹
	۵	۰/۹۰	۰/۲۹	۰/۰۸
	۶	۰/۷۱	۰/۴۵	۰/۲۲
	۷	۰/۸۳	۰/۳۷	۰/۱۹
میانگین		۰/۷۸		۰/۱۷

با توجه به نتایج جدول در درس علوم سؤالات ۶،۳،۱ دارای تشخیص بالاتر از۰/۲ هستند و سوال ۲ با ضریب تشخیص ۰/۰۵۱ دارای کمترین مقدار تشخیص است.به صورت متوسط ضریب تشخیص سؤالات

در درس علوم۰/۱۷ است. در درس ریاضی ضریب تشخیص همه سؤالات بالاتر از ۰/۲ است و به صورت متوسط ضریب تشخیص سؤالات ۲/۲۱۶ است.

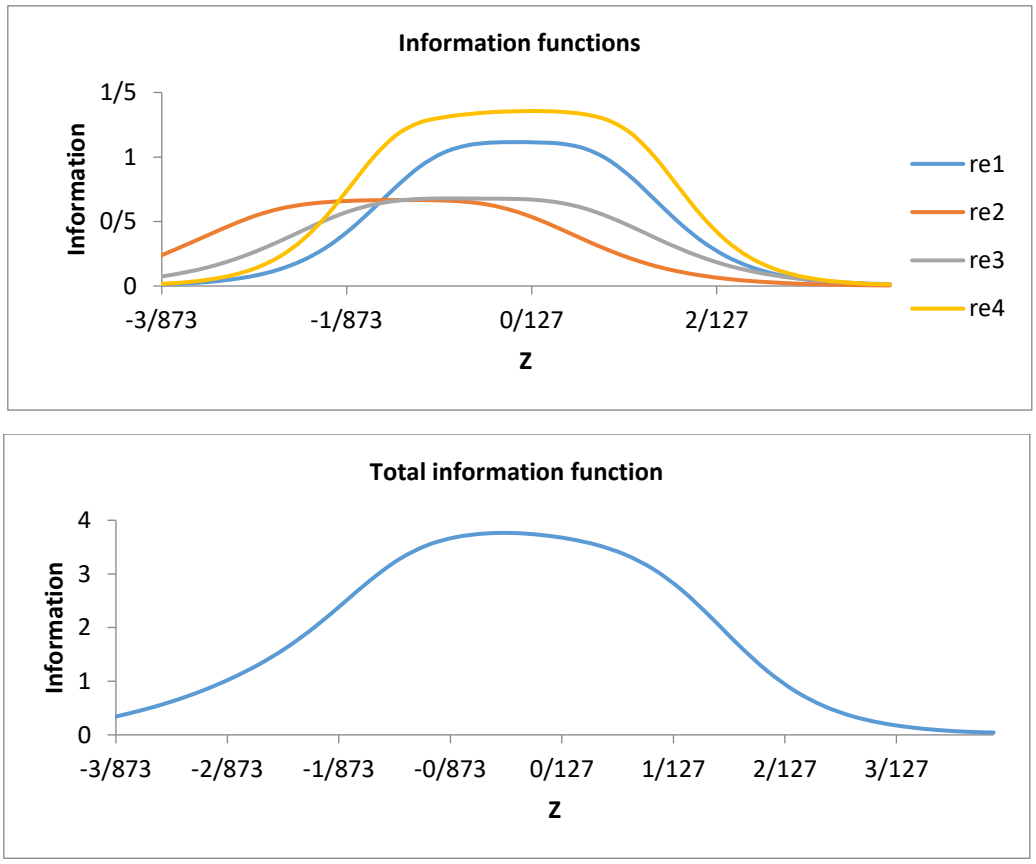
یکی از مفاهیم مورد بررسی در نظریه‌های نوین اندازه‌گیری، تابع آگاهی است که در نظریه سوال-پاسخ نقش مهمی دارد(Embretson & Reise ,2013) . از نظر آماری ،آگاهی به معنای مفهوم مقابل میزان دقت در برآورد یک پارامتر می باشد. در این نظریه، به جای اعتبار، از تابع آگاهی سوال و آزمون بهره گرفته می‌شود(Sharifi et al ,2013) .این ویژگی را می‌توان برای ارزیابی دقت اندازه‌گیری به کار برد. در حقیقت، سؤالاتی که دارای قدرت تشخیص بالاتری هستند، نسبت به سؤالاتی با قدرت تشخیص پایین‌تر، آگاهی بیشتری ارائه می‌دهند. سؤالات با قدرت تشخیص بالا، نقش مهم‌تری در دقت اندازه‌گیری ایفا می‌کنند(Habibi et al ,2013) . در شکل های زیر تابع آگاهی سؤالات تشریحی و چهارگزینه ای در دروس علوم و ریاضی به صورت زیر آمده است.





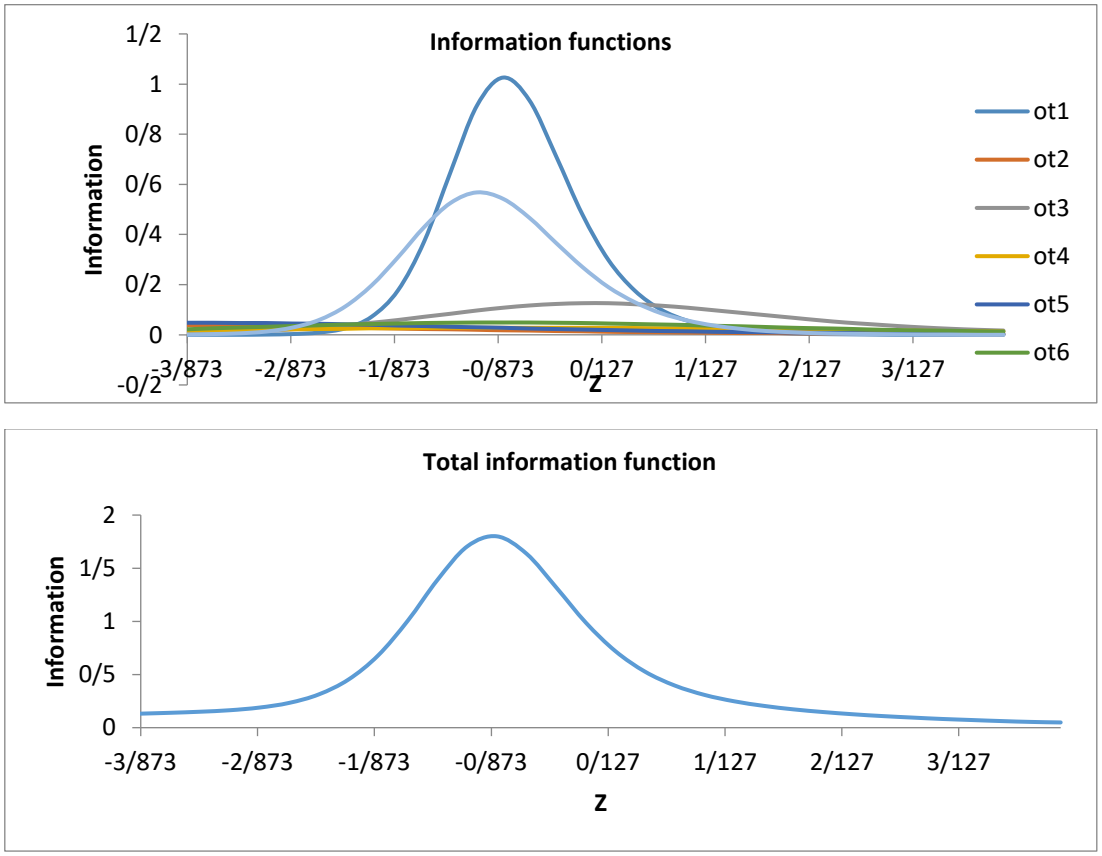
شکل۱.تابع آگاهی سؤالات تشریحی علوم

نتایج حاصل از نمودار تابع آگاهی سؤالات تشریحی علوم نشان می‌دهد که هر یک از سؤالات تشریحی علوم میزان متفاوتی از اطلاعات را در سطوح گوناگون توانایی دانش‌آموزان فراهم می‌کنند. در این میان، سؤال ۴ بیشترین اطلاعات را ارائه داده و با اوج حدود ۰,۶، بالاترین قدرت تمیز را در بین سؤالات دارد. سؤال ۲ نیز پس از آن در رتبه دوم قرار گرفته و سهم قابل توجهی در تمایز دانش‌آموزان ایفا کرده است. در مقابل، سؤالات ۱ و ۵ کمترین میزان اطلاعات را تولید کرده‌اند و قدرت تمیز پایینی دارند، در حالی که سؤال ۳ در سطح متوسط قرار می‌گیرد .بررسی دامنه توانایی (Zیاθ) نشان می‌دهد که اوج اطلاعات بیشتر سؤالات در بازه‌ی بین ۲- و ۰ واقع شده است. این یافته بیانگر آن است که مجموعه سؤالات تشریحی علوم بیشترین کارایی را در تشخیص دانش‌آموزان ضعیف تا نزدیک به متوسط داشته‌اند. با این حال، در سطوح بالاتر توانایی (بالاتر از +۱) میزان اطلاعات همه سؤالات به‌طور محسوسی کاهش یافته است.



شکل۲.تابع آگاهی سؤالات تشریحی ریاضی

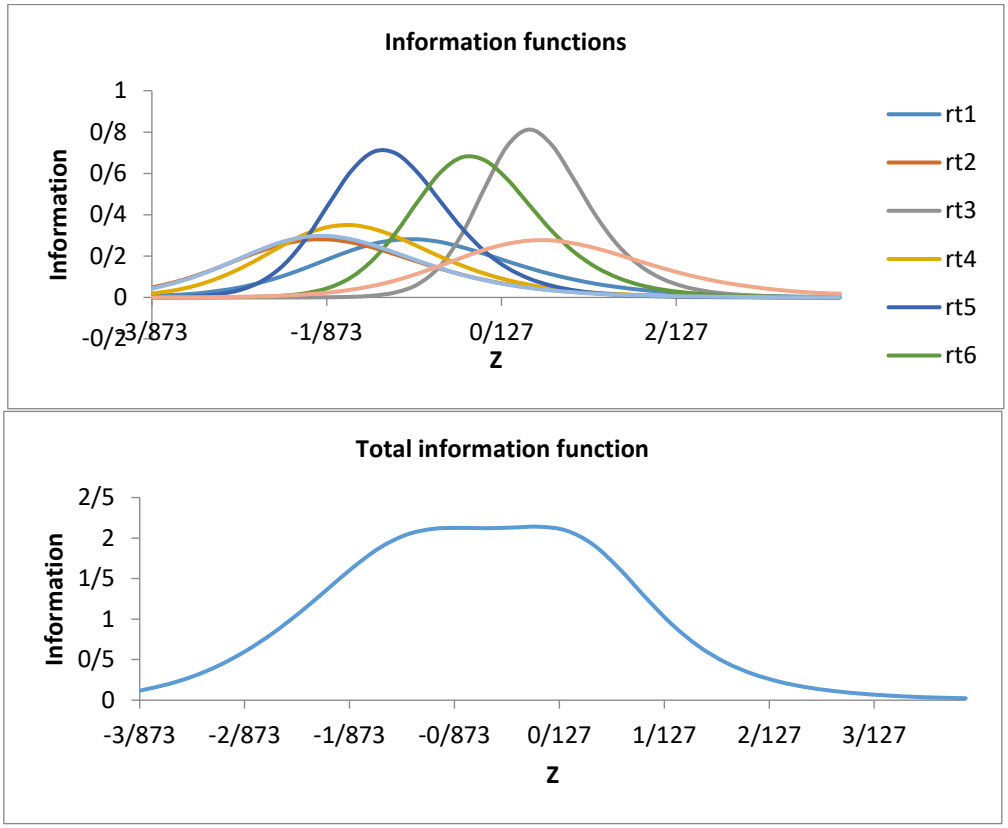
در نمودار شکل۲، تابع آگاهی هر سؤال به‌صورت جداگانه نشان داده شده است و در نمودار بالا، تابع آگاهی کل آزمون (ترکیب همه سؤالات) نمایش داده شده است .سؤالات مختلف در بازه‌های متفاوتی از توانای θیا Z بیشترین اطلاعات را ارائه می‌دهند. به عنوان مثال، سؤال‌های ۱ و ۴ در حدود توانایی متوسط (نزدیک به صفر) بیشترین آگاهی را دارند، در حالی‌که سؤال‌های دیگر مانند ۲ و ۳ در سطوح پایین‌تر توانایی مفیدتر هستند .در نمودار پایین، تابع آگاهی کل آزمون نشان می‌دهد که سؤالات تشریحی ریاضی بیشترین دقت را در حدود توانایی بین -۱ تا +۱ دارند. این یعنی آزمون تشریحی ریاضی بیشترین تمیز را بین دانش‌آموزانی با توانایی متوسط فراهم می‌کند و در سطوح خیلی بالا یا خیلی پایین توانایی، دقت کمتری دارد.



شکل۳.تابع آگاهی سؤالات چندگزینه ای علوم

نتایج حاصل از نمودار شکل ۳تابع آگاهی سؤالات چندگزینه ای علوم نشان می‌دهد ،در نموداربالا، تابع آگاهی هر سؤال جداگانه رسم شده است. مشاهده می‌شود که برخی سؤالات مانند ۱ و ۵ بیشترین

اطلاعات را در حدود توانایی نزدیک به ۱- فراهم می‌کنند، در حالی که سایر سؤالات مانند سوال۳ دامنه گسترده‌تری از توانایی را پوشش می‌دهند .نمودار پایین که آگاهی کل آزمون را نشان می‌دهد، بیانگر این است که آزمون علوم بیشترین قدرت تمیز را در سطح توانایی حدود پایین ۱- دارد. این بدین معناست که آزمون چندگزینه‌ای علوم برای تمیز دانش‌آموزانی با سطح توانایی پایین‌تر (زیر میانگین) مناسب‌تر طراحی شده و در توانایی‌های بالاتر کارایی کمتری دارد.



شکل۴.تابع آگاهی سؤالات چند گزینه ای ریاضی

نتایج حاصل از نمودار شکل۴ در نمودار بالا، نشان می دهد که هر سؤال در یک محدوده‌ی خاص بیشترین اطلاعات را ارائه می‌دهد. به طور مثال، سؤال ۳ در حدود توانایی مثبت (بالاتر از صفر) بیشترین آگاهی را دارد، در حالی که سؤال ۱ و ۵ در حدود توانایی منفی بیشترین دقت را نشان می‌دهند. این تنوع نشان می‌دهد که سؤالات چندگزینه‌ای ریاضی طیف گسترده‌تری از توانایی را پوشش داده‌اند .نمودار پایین (آگاهی کل آزمون) نشان می‌دهد که آزمون چندگزینه‌ای ریاضی بیشترین دقت و قدرت تمیز را در محدوده توانایی بین ۱- و ۱+ دارد و نسبت به دو نوع دیگر آزمون، بازه وسیع‌تری از توانایی‌ها را پوشش داده است.

این ویژگی موجب می‌شود که آزمون چندگزینه‌ای ریاضی هم برای دانش‌آموزان ضعیف‌تر و هم برای دانش‌آموزان متوسط و قوی، اطلاعات قابل اعتمادی فراهم کند.

بحث و نتیجه گیری

این پژوهش به تحلیل ویژگی‌های روان‌سنجی سؤالات تشریحی و چندگزینه‌ای در دروس ریاضی و علوم از منظر تئوری کلاسیک و سوال_پاسخ پرداخته است .

تئوری کلاسیک بر اساس مفاهیمی مانند نمره واقعی و خطای اندازه‌گیری، به ارزیابی کلی عملکرد یک آزمون و سؤالات آن می‌پردازد و معمولاً برای تعیین ضریب تشخیص و دشواری سؤالات به کار می‌روداین تئوری ساده‌تر و در کاربردهای اولیه آزمون‌سازی مفید است. (Crocker, 2006)،در مقابل، تئوری سوال_پاسخ (IRT) به عنوان رویکردی مدرن‌تر، به تحلیل دقیق‌تر سؤالات می‌پردازد و امکان بررسی شاخص‌های مختلفی همچون پارامتر تمیز، دشواری و حدس سؤالات را فراهم می‌کند (Embretson, 2006).نتایج این پژوهش به تفکیک برای تئوری کلاسیک و تئوری سوال_پاسخ ارائه شده است.

در تحلیل سؤالات چندگزینه‌ای از منظر تئوری کلاسیک، نتایج نشان می‌دهد که در درس ریاضی، تمامی سؤالات دارای ضریب تشخیص بالاتر از ۰,۲۰ بودند و به‌طور متوسط ضریب تشخیص سؤالات به میزان ۰,۲۱۶ بود. در مقابل، در درس علوم، تنها سؤالات ۱ و ۷ دارای ضریب تشخیص بالاتر از ۰,۲۰ بودند و سایر سؤالات ضریب تشخیص پایین‌تری داشتند (به‌طور متوسط ۰,۱۷). در تحلیل سؤالات تشریحی براساس تئوری کلاسیک، نتایج نشان می‌دهد که دردرس علوم کمترین ضریب تمیز مربوط به سوال ۱ با تمیز۰۴۷ /۰ است.و درس ریاضی کمترین ضریب تمیز مربوط به سوال ۲ با تمیز۰/۴۹ است. متوسط ضریب تمیز سؤالات تشریحی در درس علوم ۰/۲۰۸است و متوسط ضریب تمیز سؤالات تشریحی در درس ریاضی ۰/۵۵ است . بررسی نتایج بر اساس تئوری سوال_پاسخ نشان می‌دهد که در درس ریاضی، تمامی چهار سوال مورد بررسی ،دارای پارامتر تمیز بالاتر از ۱,۳۵ بودند. در مقایسه، در درس علوم، تنها یکی از سؤالات (سوال ۴) دارای پارامتر تمیز بالاتر از ۱,۳۵ بود و سایر سؤالات در بازه‌های تمیز متوسط (۰,۶۵ تا ۱,۳۴) و ضعیف (زیر ۰,۶۵) قرار داشتند. در بررسی پارامتر آستانه سؤالات تشریحی ریاضی تغییرات مثبت در هر گام از آستانه در سؤالات ۱، ۳ و ۴ مشاهده شد. در مقابل، در درس علوم، روند تغییرات مثبت در سؤالات ۱ و ۴ به وضوح قابل مشاهده بود. در سؤالات چندگزینه‌ای در درس علوم، سؤالات ۱ و ۷ دارای قدرت تشخیص بالایی هستند، در حالی که سوال ۳ قدرت تشخیص متوسطی دارد و سایر سؤالات نیز قدرت تشخیص نسبتا ضعیفی را نشان داده‌اندر مقابل، در درس ریاضی، سؤالات ۳، ۴، ۵ و ۶ دارای قدرت تشخیص بالایی هستند و سایر سؤالات نیز به‌طور متوسط قدرت تشخیص مناسبی ارائه کرده‌اند.

بررسی نتایج این پژوهش نشان داد که سؤالات تشریحی در هر دو درس علوم و ریاضی عملکرد بهتری نسبت به سؤالات چندگزینه‌ای داشته است . به طوری که نتایج عملکرد دانش آموزان در سؤالات تشریحی بر اساس تئوری کلاسیک نشان میدهد که متوسط ضریب تمیز در درس علوم ۰/۲۰۸و درس ریاضی ۰/۵۵ است و متوسط ضریب دشواری در درس علوم ۲/۵۹۱ و در درس ریاضی ۲/۳۴۲ می باشد که این عملکرد نشان دهنده ی قدرت تشخیص و تمیز بهتر سؤالات تشریحی است. همچنین بررسی سؤالات تشریحی از منظر تئوری سوال_پاسخ نشان داد که هر چهار سوال در درس ریاضی دارای تمیز بالاتر از ۱/۳۵می باشند و در دس علوم نیز سوال۱ دارای تمیز ۰/۰۸۷،سوال ۲ دارای تمیز ۱/۰۹۰،سوال ۳ دارای تمیز ۰/۸۴۴،سوال۴ دارای تمیز۱/۴۱۹و سوال ۵ دارای تمیز ۰/۵۳۳ می باشند و در پارامتر آستانه نیز در هردو درس علوم و ریاضی تغییرات در هرگام به گام دیگر مثبت گزارش شده است که این عملکرد نیز نشان دهنده قدرت تمیز و آستانه بهتر سؤالات تشریحی می باشد . این عملکرد بهتر سؤالات تشریحی نسبت به سؤالات چهارگزینه ای ناشی از چندین عامل می باشد که در ادامه به بررسی آنها پرداخته می شود .سؤالات تشریحی به دانش‌آموزان و فراگیران این امکان را می‌دهد که موضوعات را به‌صورت عمقی بررسی کنند و توانایی‌های تفکر انتقادی و تحلیلی خود را نشان دهند. این نوع سؤالات به‌ویژه برای ارزیابی مهارت‌هایی که نیاز به تحلیل، ارزیابی، و ترکیب اطلاعات دارند مناسب هستند.(Anderson, 2001).برخلاف سؤالات چندگزینه‌ای که دانش‌آموز را به انتخاب یک گزینه محدود می‌کنند، سؤالات تشریحی به آنها این امکان را می‌دهند که ایده‌های خود را با جزئیات و به‌صورت خلاقانه بیان کنند.(Biggs, 2011).سؤالات تشریحی این امکان را می‌دهند که سطوح بالاتری از یادگیری مانند تحلیل، ترکیب، و ارزیابی مورد ارزیابی قرار گیرند، که در دیگر روش‌های ارزیابی کمتر ممکن است و دانش‌آموزان می‌توانند پاسخ‌های خود را بر اساس تجربیات، دانش قبلی و نظرات شخصی خود تنظیم کنند.

این ویژگی به ویژه در ارزیابی موضوعات پیچیده یا چندوجهی مفید است. (Moon, 2006)با وجود مزایای سؤالات تشریحی اما با این حال این سؤالات به دلایلی کمتر مورد استفاده قرار می گیرند.تصحیح سؤالات تشریحی نیازمند زمان زیادی است و احتمال وجود خطای انسانی در ارزیابی وجود دارد. این مسئله می‌تواند منجر به تفاوت در نمره‌دهی بین ارزیابان مختلف شود.(Brown, 2013).سؤالات تشریحی ممکن است پایایی کمتری نسبت به سؤالات چندگزینه‌ای داشته باشند، زیرا ممکن است پاسخ‌دهی به آنها تحت تأثیر عوامل غیرعلمی مانند مهارت‌های نگارشی، خستگی دانش‌آموز یا زمان محدود قرار گیرد.(Gipps, 1994) .سؤالات تشریحی معمولاً تنها یک یا دو موضوع را پوشش می‌دهند و نمی‌توانند تمامی محتوای آموزشی را به طور کامل ارزیابی کنند. در نتیجه، ممکن است تصویری ناقص از دانش و توانایی‌های دانش‌آموز ارائه دهند.(Race, 2014)

با وجود نتایج ارزشمندی که این پژوهش به دست آورده است، برخی محدودیت‌ها باید در تفسیر یافته‌ها و کاربرد آن‌ها مورد توجه قرار گیرد. نخستین محدودیت به نوع سؤالات مورد استفاده بازمی‌گردد.

در آزمون تشریحی ریاضی، هرچند امکان سنجش عمیق‌تر توانایی‌های تحلیلی و فرایند حل مسئله وجود داشت، اما این نوع سؤال‌ها بیشتر بر جنبه‌های استدلالی و تشریحی تأکید دارند و ممکن است سایر ابعاد توانایی‌های ریاضی مانند سرعت محاسبات یا درک شهودی مفاهیم را کمتر پوشش داده باشند. در آزمون چندگزینه‌ای علوم نیز با وجود دقت مناسب در تفکیک سطوح پایین‌تر توانایی، به دلیل ماهیت بسته‌ی این سؤالات، احتمال آن وجود دارد که توانایی‌های خلاقیت، استدلال باز و تحلیل عمیق علمی کمتر منعکس شده باشند. در آزمون چندگزینه‌ای ریاضی نیز هرچند پوشش دامنه‌ی وسیع‌تری از توانایی‌ها مشاهده شد، اما همچنان محدودیت‌هایی در بازتاب کامل تمامی ابعاد توانایی‌های شناختی دانش‌آموزان وجود دارد.محدودیت دیگر مربوط به روش‌های آماری مورد استفاده است. هرچند استفاده همزمان از تئوری کلاسیک و نظریه سؤال–پاسخ (IRT) توانست تصویری دقیق‌تر از ویژگی‌های روان‌سنجی سؤالات ارائه دهد، اما این روش‌ها ممکن است تمامی جنبه‌های پیچیده‌تر مانند چندبعدی بودن توانایی‌ها یا اثر عوامل انگیزشی را پوشش نداده باشند. همچنین، این پژوهش صرفاً به دو درس ریاضی و علوم در یک مقطع خاص پرداخته است؛ بنابراین، نتایج آن به‌طور مستقیم به سایر دروس و سطوح تحصیلی قابل تعمیم نیست. از سوی دیگر، عواملی همچون شرایط برگزاری آزمون، اضطراب دانش‌آموزان و تفاوت‌های فردی در شیوه پاسخ‌دهی، متغیرهایی هستند که احتمالاً بر عملکرد دانش‌آموزان اثر گذاشته‌اند و در این پژوهش به‌طور کامل کنترل نشده‌اند.

با توجه به این محدودیت‌ها، پیشنهاد می‌شود در پژوهش‌های آینده آزمون‌ها به گونه‌ای طراحی شوند که ترکیبی از سؤالات تشریحی و چندگزینه‌ای را دربرگیرند تا نقاط قوت هر نوع سؤال مکمل یکدیگر باشد. سؤالات تشریحی می‌توانند برای سنجش مهارت‌های استدلالی و فرایند حل مسئله به کار روند، در حالی‌که سؤالات چندگزینه‌ای دقت و سرعت سنجش را افزایش دهند. همچنین، معلمان و طراحان آزمون بهتر است به این نکته توجه داشته باشند که سؤالات چندگزینه‌ای علوم بیشتر برای تفکیک دانش‌آموزان در سطوح پایین‌تر توانایی مفید هستند و در مقابل، سؤالات تشریحی ریاضی قدرت بیشتری در تمایز دانش‌آموزان با سطح توانایی متوسط دارند؛ بنابراین، استفاده‌ی متوازن از این دو نوع سؤال می‌تواند تصویری جامع‌تر از توانایی‌های دانش‌آموزان فراهم کند. سیاست‌گذاران آموزشی نیز می‌توانند با بهره‌گیری از نتایج این پژوهش، دستورالعمل‌هایی برای طراحی آزمون‌ها تدوین کنند تا علاوه بر سنجش دانش نظری، به مهارت‌های استدلالی، خلاقیت و تحلیل نیز توجه شود. از سوی دیگر، برای افزایش تعمیم‌پذیری، پژوهش‌های آتی بهتر است در سایر دروس و مقاطع تحصیلی انجام شوند تا امکان مقایسه‌ی بین‌دروسی فراهم گردد. در نهایت، بهره‌گیری از مدل‌های پیشرفته‌تر روان‌سنجی مانند مدل‌های چندبعدی IRT یا مدل‌های رشد می‌تواند ابعاد دقیق‌تری از توانایی‌ها را آشکار کرده و به تحلیل عمیق‌تر ویژگی‌های آزمون‌ها کمک نماید.

ملاحظات اخلاقی

در پژوهش حاضر فرم‌های رضایت‌نامه آگاهانه توسط تمامی شرکت‌کننده‌ها تکمیل شد.

حامی مالی

هزینه‌های مطالب حاضر توسط نویسنده مقاله تأمین شد.

تعارض منافع

بنابراظهار نویسنده مقاله حاضر فاقد هرگونه تعارض منافع بوده است.

References

Adegoke, B.A. (2013). Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks. Journal of Education and Practice, 4, 87-96.

Ado Abdu Bichi , Rohaya Talib .2018. Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. International Journal of Evaluation and Research in Education (IJERE)،Vol.7, No.2, pp. 142~151

Adu-Mensah, J., & Adom, D. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. International Journal of Evaluation and Research in Education.

Ahmadi, H., Shirbagi, N., & Shirbagi, S. (2023). Teachers' Understanding and Use Of Authentic Assessment In the Teaching-Learning Process. Journal of Research in Teaching, 11(4), 170-197. [In Persian]

Alimirzaei.M, Moghadamzadeh. A , Minaei. A, Eizanlou. B , & Salehi. K. (2019). Sources of the Differential Item Functioning and its Application in Education. Journal of Research in Teaching, 7(1), 133-153. [In Persian]

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.

Ayanwale, Musa Adekunle, Adeleke, J. O., & Mamadelo, T. I. (2019). Invariance Person Estimate of Basic Education Certificate Examination: Classical Test Theory and Item Response Theory Scoring Perspective. Journal of the International Society for Teacher Education, 23(1), 18–26.

Ayenew Takele Alemu , Hiwot Tesfa, Addisu Mulugeta,Enyew Tale Fenta, Mahider Awoke Belay.(2024). Quality of multiple choice question items: item analysis, International Journal of Scientific Reports 10(6):195-199

Babatunde K Oladele, Benson A. Adegoke2020 Using Test Theories Models to Assess Senior Secondary Students Ability in Constructed-Response Mathematics Tests .Journal of Education and Practice . Vol.11, No.7,

Baker, F. (2001), The Basics of Item Response Theory, ERIC: Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Bichi, A. A. (2016). Classical test theory: An introduction to linear modelling approach to test and item analysis. International Journal for Social Studies, 2(9), 27-33

Biggs, J., & Tang, C. (2011). Teaching for quality learning at university.

Brown, G., Bull, J., & Pendlebury, M. (2013). Assessing student learning in higher education. Routledge.

Butakor P.K. (2022). Using Classical Test and Item Response Theories to Evaluate Psychometric Quality of Teacher-Made Test in Ghana. European Scientific Journal, ESJ, 18, (1), 139

C. Alonso-Fernandez, I. Martinez-Ortiz, R. Caballero, M. Freire, and B. Fernandez-Manjon,. 2020 .Predicting students' knowledge after playing a serious game based on learning analytics data: A case study, Journal of Computer Assisted Learning, vol. 36, no. 3, pp. 350–358

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. In Annual Review of Statistics and Its Application (Vol. 3, pp. 297–321). Annual Reviews Inc.

Chukwu Ohiri .S .(2023) . Psychometric Analysis at Item Level of the Waec May/June Mathematics Multiple Choice Questions Using the Classical Test Theory . International Journal of Research Publication and Reviews , 4(11) , 132-138

Clarke, M. (2011). Framework for building an effective student assessment system: READ/SABER Working Paper. World Bank.

Cobbinah, A. & Ntumi, S. (2022). Difficulty, discrimination and pseudo-guessing indices of the West African Examinations Council core mathematics multiple choice items: Practical implications of using item response theory. Journal Research in Education Sciences, 13(5), 51-60

Crocker, L., & Algina, J. (2006). Introduction to Classical and Modern Test Theory. Cengage Learning.

Ebrahimi Manesh, M. R., Daneshpoor, A., Hasani Panah, T., & Haji Ramazani, E. (2024). Examination of student evaluation methods throughout the academic year. Journal of Psychology and Educational Sciences, 5 (53), 627-636. [In Persian]

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc., Mahwah. 1–371

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc., Mahwah. 1–371.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Ganglmair, A., & Lawson, R. (2010). Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. In EEuropean Advances in Consumer Research, 6, 162–167.

Gipps, C. V. (1994). Beyond testing: Towards a theory of educational assessment. Brown, G., Bull, J., & Pendlebury, M. (2013). Assessing student learning in higher education.

Habibi, M., Khodaei, E., & Ezzanlou, B. (2012). *Old and new measurement theories in behavioral and medical sciences: A review of methodology, advantages, and challenges*. Quarterly Journal of Behavioral Sciences Research, 10(4), 302-315. [In Persian]

Hambleton, R.K. and Swaminathan, H. (1985). Item response theory: principles and applications. p.332.

Hambleton, R.K. and Swaminathan, H. (1985). Item response theory: principles and applications. p.332.

Hambleton, R.K., Swaminathan, H. and Rogers, J.H. (1991), Fundamentals of Item Response Theory, Sage Publications, Newbury Park, CA

Jose Manuel Azevedo , Ema P. Oliveira, Patrícia Damas Beites ,(2019), Using Learning Analytics to evaluate the quality of multiple-choice questions. A perspective with Classical Test Theory and Item Response Theory, The International Journal of Information and Learning Technology,36(4) , 322-341

Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. Research and Evaluation in Education.

Lang, J. W. B., & Tay, L. (2021). The Science and Practice of Item Response Theory in Organizations. In Annual Review of Organizational Psychology and Organizational Behavior, 8, 311–338.

Maba, W., Perdata, I. B. K., & Astawa, I. N. (2017). Constructing assessment instrument models for teacher's performance, welfare and education quality. International Journal of Social Sciences and Humanities, 1(3), 88–96.

McAlpine, M. (2002), "Design requirements of a databank", The CAA Centre TLTP Project, Leicestershire

Mehta G, Mokhasi V. 2014 .Item analysis of multiple choice questions-an assessment of the assessment tool. Int J Health Sci Res,;4(7):197-202.

Mehtap Erguven , Two approaches to psychometric process: Classical test theory and item response theory , (2013), Journal of Education , Vol. 2 No. 2

Moon, J. (2007). Critical Thinking: An Exploration of Theory and Practice (1st ed.). Routledge.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2020). TIMSS 2019 International Results in Mathematics and Science. TIMSS & PIRLS International Study Center, Boston College.

Musa Adekunle Ayanwale , Julia Chere-Masopha and Malebohang C. Morena , (2022), The Classical Test or Item Response Measurement Theory: The Status of the Framework at the Examination Council of Lesotho , International Journal of Learning, Teaching and Educational Research , Vol. 21, No. 8, pp. 384-406

Quansah, F., Amoako, I., & Ankomah, F., 2019."Teachers' test construction skills in Senior High Schools in Ghana: Document Analysis," International Journal of Assessment Tools in Education, vol. 6, no. 1, pp. 1-8

Race, P. (2014). The lecturer's toolkit: A practical guide to assessment, learning, and teaching

Ravela, P., Arregui, P., Valverde, G., Wolfe, R., Ferrer, G., Rizo, F. M., Aylwin, M., & Wolff, L. (2009). The Educational Assessments that Latin America Needs. Washington, DC: PREAL.

Rioborue Alexander Oghenerume , & Friday Egberha,2024, Comparative Analysis of Item Statistics of WASSCE and NECO SSCE 2023 Data Processing Multiple Choice Tests Using Item Response Theory, International Journal of Educational Researchers, 15(1): 58-67

Samadieh, H., Tanhayi Roshwanlou, F., Saeedi Rezvani, T., & Talebzadeh Shushtari, L. (2019). *Psychometric properties of the Unidimensional Relationship Closeness Scale based on classical test theory and item response theory*. Educational Measurement Quarterly. [In Persian]

Santoso, A., Pardede, T., Djidu, H., Apino, E., Rafi, I., Rosyada, M. N., & Abd Hamid, H. S. (2022). The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory. Research and Evaluation in Education, 8(2), 140–151.

Seif, A. A. (2016). *Educational Measurement, Assessment, and Evaluation* (7th ed., 8th print). Doran Publishing.. [In Persian]

Sharifi, H. P., & Sharifi, N. (2013). *Principles of psychometrics and psychological testing*. Tehran: Roshd Publications. [In Persian]