تدوین و هنجاریابی آزمون پیشرفت تحصیلی زیست‌شناسی

حسن مشتاقیان ابرقویی* [۱]، بهرام جوکار [۲]

تاریخ دریافت:۱۷ / ۱۴۰۱/۰۱ تاریخ پذیرش: ۱۴۰۱/۰۵/۱۵
ازصفحه ۲۹ تا ۴۶

چکیده:

هدف پژوهش حاضر، تدوین و هنجاریابی آزمون زیست‌شناسی پایه دهم دوره متوسطه به‌منظور اندازه‌گیری سطح آموخته‌ها و تعیین پراکندگی پیشرفت تحصیلی دانش‌آموزان در این ماده درسی بود. برای نیل به هدف کلی تحقیق، هر دو مدل کلاسیک آزمون و نظریه سؤال- پاسخ مورد استفاده قرار گرفت. ابزار اولیه پژوهش شامل ۱۵۰ سؤال چهارگزینه‌ای از محتوای زیست‌شناسی پایه دهم رشته تجربی بود که بر روی نمونه‌ای به حجم ۳۰۰ دانش‌آموز دختر و پسر اجرا شد. به دنبال تجزیه‌وتحلیل سؤال‌های آزمون فرم‌های نهایی آزمون تدوین شد. ابزار نهایی دو فرم موازی ۵۰ سؤالی بود که بر روی نمونه هنجاری به حجم ۹۳۸ دانش‌آموز دختر و پسر شهرستان شیراز اجرا گردید. ضریب پایایی برآورد شده برای ثبات درونی فرم‌های آزمون به ترتیب ۰/۸۹ و ۰/۸۸ به دست آمد. بر اساس تحلیل عاملی، هر دو فرم آزمون از یک عامل کلی اشباع بودند. نتایج نشان داد که بین میانگین نمرات پسران و دختران تفاوت معناداری وجود ندارد؛ بنابراین هنجارهای استاندارد و هم صدک برای همه آزمودنی‌ها محاسبه شد. یافته‌های حاصل از تجزیه‌وتحلیل نظریه سؤال-پاسخ نشان داد که بیش از ۹۲ درصد از سؤال‌ها به‌طور قابل‌توجهی با مدل لجستیک سه پارامتری برازش دارند. تابع آگاهی برآورد شده آزمون یک منحنی زنگوله‌ای شکل بود و در محدوده توانایی ۰/۵- تا ۲/۵+ بیشترین اطلاعات را ارائه می‌داد؛ همچنین حداکثر میزان آگاهی دهندگی فرم‌های آزمون در سطح توانایی ۱/۵+ به دست آمد که نشان می‌دهد آزمون تدوین‌شده تخمین دقیق‌تری از نمره واقعی آزمودنی‌هایی ارائه می‌دهد که سطح توانایی آن‌ها بالاتر از متوسط پیوستار توانایی است.

کلمات کلیدی: آزمون پیشرفت زیست‌شناسی، مدل کلاسیک اندازه‌گیری، نظریه سؤال- پاسخ، هنجاریابی.

۱. دکتری سنجش و اندازه گیری؛ دبیر زیست شناسی دبیرستان باقرالعلوم، آموزش و پرورش ناحیه ۴ شیراز * kavir311@gmail.com
۲. استاد روانشناسی تربیتی، عضو هیات علمی دانشکده روانشناسی و علوم تربیتی دانشگاه شیراز .

مقدمه و بیان مسئله

از امتحانات غیررسمی معلمان گرفته تا آزمون‌های استانداردشده که در سطح ملی اجرا می‌شود، سنجش[1] برای مدت‌های طولانی بخش جدایی‌ناپذیر از فرآیند آموزشی بوده است (پلگرینو، چودوفسکی و گلیزر[2]، ۲۰۰۱: ۱۹). بر طبق استانداردهای آزمون‌های آموزشی و روان‌شناختی[3]، سنجش را می‌توان «هرگونه روش سیستماتیک به دست آوردن اطلاعات از آزمون‌ها و دیگر منابع که به‌منظور استنباط درباره ویژگی‌های افراد، اشیا یا برنامه‌ها به کار می‌رود، تعریف کرد» (سیمون، ارسیکان و روسو[4]، ۲۰۱۳: ۱۷۲). به‌طورکلی هدف از سنجش کمک به یادگیری، اندازه‌گیری پیشرفت فردی دانش آموزان و ارزیابی برنامه‌ها است (پلگرینو و همکاران، ۲۰۰۱: ۳۷). اگرچه ظاهراً چنین تصور می‌شود که سنجش مرحله پایانی فعالیت‌های آموزشی معلم است؛ اما واقعیت کنونی این است که سنجش و اندازه‌گیری، اغلب تعیین‌کننده نوع فعالیت‌های آموزشی معلم و یادگیری دانش‌آموزان است و می‌تواند تأثیر بسزایی در فرآیند یاددهی-یادگیری داشته باشد (سیزک[5]، ۱۹۹۳). کاملاً روشن است که سنجش یک بخش جداشده از سیستم آموزشی نیست. آنچه اندازه گرفته می‌شود و اینکه اطلاعات به چه شکل استفاده شود تا حد زیادی به برنامه درسی و روش‌های آموزشی مورد استفاده وابسته است. باوجوداین، سنجش تأثیر قوی بر هر دو برنامه درسی و شیوه‌های آموزشی دارد (لین[6]، ۲۰۰۰؛ پلگرینو و همکاران، ۲۰۰۱). نتایج حاصل از چندین مطالعه عمده درباره تأثیر آزمودن[7] روی آموزش و برنامه درسی، نشان داده است اطلاعاتی که بر اساس آن آزمون‌ها ساخته می‌شوند، تأثیر معناداری نه‌تنها بر روی تدریس، بلکه بر روی برنامه آموزشی و انگیزش دانش آموزان دارد (سالمون-کاکس[8]، ۱۹۸۱؛ کروکس[9]، ۱۹۸۸؛ فلپس[10]، ۲۰۱۲). ازنظر بلک و همکاران[11] (۲۰۰۳) فعالیت‌های سنجش اگر اطلاعاتی به‌عنوان بازخورد برای معلمان و دانش آموزان فراهم کند، می‌تواند به اصلاح فرآیندهای یاددهی- یادگیری بینجامد. آن‌ها به معلمان، دانش آموزان و والدین آن‌ها در تعیین سطح و کیفیت یادگیری دانش آموزان کمک می‌کنند. آن‌ها به معلمان کمک می‌کند تا درک کنند که چگونه آموزش را بر اساس شواهدی از یادگیری دانش-آموزان منطبق سازند. همچنین به مدیران و سرپرستان کمک می‌کند تا سطح پیشرفت فردی دانش آموزان را در کلاس، مدرسه یا منطقه تعیین کنند؛ و در آخر آنکه سنجش به سیاست‌گذاران و عموم مردم کمک می‌کند تا به ارزیابی تأثیر سیستم‌های آموزشی بپردازند (پلگرینو و همکاران، ۲۰۰۱: ۱۹).

1. Assessment
2. Pellegrino, Chudowsky & Glaser
3. The Standards for Educational and Psychological Testing
4. Simon, Ercikan & Rousseau
5. Cizek
6. Linn
7. testing
8. Salmon-Cox
9. Crooks
10. Phelps
11. Black et al

علی‌رغم نقش سنجش در جهت‌بخشی به فعالیت‌های یاددهی- یادگیری، امروزه با انتقادهای بسیاری روبرو شده است. برخی مواقع این انتقادها به ابزارهای سنجش بوده است؛ اما بیشتر انتقادها به سنجش، نه ناشی از خود آزمون‌ها، بلکه استفاده‌های ثانویه‌ای است که از نتایج این آزمون‌ها صورت می‌گیرد. زمانی که نتایج آزمون‌ها و امتحانات به‌عنوان اندازه‌های پاسخگویی معیار قضاوت، پاداش و تنبیه معلمان، کارایی مدارس و برنامه‌ها به کار می‌رود، موجب پیامدهای ناخواسته بی‌شماری می‌شود. برای نمونه، معلمان ممکن است با هزینه کردن دیگر اهداف آموزشی، وقت زیادی را صرف آماده‌سازی دانش آموزان برای امتحان کنند؛ ضمن آنکه ممکن است تقلب را در بین افراد رواج دهد (کورتز، ۲۰۰۸). همچنین استفاده از این نوع سنجش، موجب کم‌سواد شدن دانش آموزان و یادگیری سطحی می‌شود (سیزک، ۱۹۹۳).

به‌هرحال صرف‌نظر از جنبه مثبت یا منفی آزمون‌ها، همان‌طور که جیپس[12] (۲۰۰۳) اظهار می‌کند؛ این یک حقیقت است که آزمودن تأثیر قدرتمندی بر شیوه‌ای که معلمان وظیفه‌شان را به اجرا درمی‌آورند دارد؛ به‌طوری که امروزه تأثیر سنجش و اندازه‌گیری بر جریان آموزش معلم، به پدید آمدن مفهومی به نام «آموزش جهت داده‌شده به‌وسیله اندازه‌گیری[13]» انجامیده است. منظور از آموزش جهت داده‌شده به‌وسیله اندازه‌گیری، یک رویکرد کلی نسبت به آموزش است که در آن هدف اصلی معلم بالا بردن سطح عملکرد دانش آموزان در یک روش خاص اندازه‌گیری است (لفرانسوا، ۲۰۰۰؛ نقل از سیف؛ ۱۳۸٤). پوفام[14] (۱۹۸۷) به‌عنوان مدافع «آموزش جهت داده‌شده به‌وسیله اندازه‌گیری»، آن را به‌عنوان مقرون به‌صرفه‌ترین روش بهبود کیفیت آموزش عمومی قلمداد می‌نماید. ازنظر پوفام، آموزش جهت داده‌شده توسط اندازه‌گیری، می‌تواند یک نیروی بسیار مؤثر برای بهبود آموزشی باشد و با مراقبت‌های روشن می‌توان از اثرات منفی بالقوه آن، از جمله تحریف برنامه‌های آموزشی جلوگیری کرد (جیپس، ۲۰۰۳: ۳۱-۳۳).

از جمله ابزارهای سنجش و اندازه‌گیری آموزشی، می‌توان به آزمون‌های پیشرفت اشاره کرد. آزمون پیشرفت، یک روش سیستماتیک برای تعیین میزان آموخته‌های دانش آموزان و اندازه‌گیری برونداد یادگیری است. به کمک آزمون‌های پیشرفت می‌توان موفقیت دانش آموزان را در طول زمان پیگیری نمود، موفقیت تحصیلی گروه‌های مختلف (کلاس‌ها/ مدارس) را مقایسه نمود، نتایج آن را در تصمیم‌گیری‌های خطیر مورداستفاده قرارداد، نقاط قوت و ضعف دانش آموزان را اندازه‌گیری کرد و اثربخشی برنامه آموزشی را معلوم نمود (ایساکس، زارا، هربرت، کومبس و اسمیت[15]، ۲۰۱۳). ازنظر گرونلاند استفاده از این آزمون‌ها نه‌تنها موجب بهبود تصمیمات آموزشی می‌شوند بلکه از طریق: ۱) بهبود انگیزش دانش‌آموز، ۲) افزایش نگهداری و انتقال آموخته‌ها، ۳) افزایش خود فهمی دانش‌آموز و ٤) فراهم نمودن بازخورد در ارتباط با کارایی آموزشی، می‌تواند به یادگیری کمک کند (گرونلاند[16]، ۱۹۸۸: ۵). بااین‌وجود، حدود تأثیرگذاری و مشارکت آزمون‌های پیشرفت تحصیلی در بهبود یادگیری و آموزش، به میزان زیادی به اصول زیربنایی حاکم برساخت و استفاده از آن‌ها بستگی دارد. به‌عبارت‌دیگر، آزمون‌های پیشرفت می‌تواند توجه دانش آموزان را به اهداف

12. Gipps
13. Measurement-Driven Instruction
14. Popham
15 Smith
16. Gronlund

اساسی آموزش هدایت کند یا آنان را از آن دور نماید. موجب شود تا دانش آموزان بر جنبه‌ای محدود از محتوای دوره تمرکز کنند یا توجه آن‌ها را به حیطه‌های مهم آن معطوف کند. یادگیری سطحی را پاداش دهند یا درک عمیق را تکلیف کنند؛ و در امر تصمیم‌گیری آموزشی؛ اطلاعاتی قابل اعتماد و یا اطلاعات سودار و مخدوش فراهم کند (همان: ۷).

علی‌رغم کاربرد وسیع آزمون‌های پیشرفت و اهمیت آن‌ها در ارزیابی و راهنمایی دانش آموزان، بسیاری از معلمان در زمینه ساخت آزمون‌های مناسب، آموزش ندیده‌اند یا آموزش اندکی دارند (گرونلاند؛ ۱۹۸۸). نتایج برخی مطالعات نشان می‌دهد که اکثر آزمون‌های معلم ساخته، فاقد اعتبار[17] و پایایی[18] است و به نظر می‌رسد بسیاری از معلمان فاقد مهارت‌های ساخت آزمون هستند (آلل- ویلیامز[19]، ۲۰۰۲؛ بیکر[20]، ۲۰۰۳). در این رابطه پوفام اظهار می‌کند: آزمون‌های معلم ساخته، بیشتر بر سطوح پایین‌تر مهارت‌های شناختی تأکیددارند و عمدتاً به یادآوری مطالب قبلاً آموخته‌شده می‌پردازند (۱۹۸۷). به نظر می‌رسد که بیشتر سؤال‌های آزمون‌های معلم ساخته تنها نیاز به یادآوری اطلاعات بسیار جزئی و خاص دارد و همچنین آموخته‌های دانش آموزان بیش از آنکه مبتنی بر مهارت‌های پیچیده و بالاتر شناختی باشد شامل محفوظات و یادآوری مطالب جزئی و کم‌اهمیت است (هومن، ۱۳۷۲).

مطالعات تائید می‌کنند که سبک سنجش بر روی سبک یادگیری دانش آموزان تأثیر خواهد گذاشت. کراکس (۱۹۸۸) نشان داد که دانش‌آموزانی که رویکرد سطحی را برای یادگیری به¬کار می‌برند، سازگاری کمتری با ارزیابی‌هایی داشته‌اند که رویکرد یادگیری عمیق را بکار می‌گیرند؛ از طرف دیگر دانش‌آموزانی که می‌توانند با موفقیت رویکرد عمیق را در یادگیری به¬کار برند به‌سادگی می‌توانند خود را با رویکرد سطحی سازگار نمایند. وی نتیجه می‌گیرد که توسعه استراتژی یادگیری عمیق از طریق تأکید بیشتر بر سؤالات سطوح بالاتر حوزه شناختی در ارزیابی حاصل می‌شود. شکی نیست که می‌توان دانش آموزان را به‌جای یادگیری طوطی‌وار به‌سوی یادگیری فرآیندهای بالاتر شناختی که مستلزم درک فعال مفاهیم و اصول بنیادی و حل مسئله است هدایت کرد. یکی از راه‌های تحقق این امر استفاده از بازخوردهای مناسب حاصل از به‌کارگیری شیوه‌های مطلوب سنجش و ارزیابی است. نتایج حاصل از یک تحلیل چندجانبه که توسط فیلیدو و روسو[21] انجام‌شده نشان داده است که «استفاده مکرر از سؤال‌های سطح بالای شناختی ضمن آموزش، بر پیشرفت دانش آموزان تأثیر مثبت دارد» (گیج وبرلاینر /ترجمه خوئی نژاد، ۱۳۷٤: ۷٤۲). فردریکسون[22]، با توجه به تأثیر آزمون‌ها معتقد است: یک وظیفه مهم مربیان و روانشناسان، ساخت و توسعه ابزارهایی است که به‌خوبی منعکس کننده طیف وسیعی از هدف‌های آموزشی است و کاربرد شیوه‌هایی برای استفاده از آن ابزارها در بهبود فرآیند آموزشی است (جیپس، ۲۰۰۳: ۳۲).

یکی از انواع آزمون‌های پیشرفت تحصیلی، آزمون‌های دروس ویژه می‌باشند. این نوع آزمون‌ها برای اندازه‌گیری آموخته‌های دانش آموزان در یک درس و یا برنامه آموزشی ویژه تهیه و استاندارد می‌شوند و ابزار مفیدی برای ارزشیابی سطح آموخته‌ها و نقاط قوت و ضعف دانش آموزان کلاس در مقایسه با

17. validity
18. reliability
19. Alele-Williams
20. Baker
21. Filido & Roso
22. Frederiksen

سطح محلی؛ منطقه‌ای یا ملی به شمار می‌آیند. (گلاور و برونینگ / ترجمه خرازی، ۱۳۷۵) تحلیل پرسش‌های این گونه آزمون‌ها به معلمان کمک می‌کند تا هدف‌های تدریس خود را بهتر بشناسند و یا آن دسته هدف‌هایی را که در عمل کمتر مورد توجه آن‌ها بوده، مدنظر قرار دهند؛ فرآیند تدریس خود را اصلاح کنند و انتظارات خود را از دانش آموزان روشن نمایند؛ یادگیری دانش آموزان را تسهیل و توجه آن‌ها را به برنامه درسی متمرکز نمایند (دفتر همکاری‌های علمی بین‌المللی وزارت آموزش‌وپرورش، ۱۳۷۹). ازجمله برنامه‌های درسی ویژه دوره متوسطه در نظام آموزش‌وپرورش ایران، زیست‌شناسی است. هدف غایی از این برنامه درسی، رشد همه‌جانبه فرد در ابعاد زیستی و تعامل آن با سایر ابعاد رشد و تسلط دانش آموزان بر چارچوب مفهومی و شیوه‌های پژوهش در قلمرو علم زیست‌شناسی است (سازمان پژوهش و برنامه‌ریزی آموزشی، ۱۳۸۹: ٦).

هرچند آموزش زیست‌شناسی می‌تواند نقش مهمی در رشد تفکر علمی و روحیه حقیقت‌جوئی در دانش آموزان ایفا کند؛ به نظر می‌رسد ضعف دانش آموزان در این ماده درسی جدی است. بر طبق یافته‌های بین‌المللی نتایج مطالعه تیمز[23] ۲۰۱۱، دانش آموزان ایرانی پایه هشتم در درس زیست‌شناسی در بین ٤۲ کشور موردمطالعه، رتبه ۲۲ را به خود اختصاص داده‌اند. متوسط نمره مقیاس زیست‌شناسی پایه هشتم این دانش آموزان ٤٦٦ بوده است که از میانگین کلی (٤٧٤) به‌طور معنی‌داری کمتر است. همچنین ازنظر نمره مقیاس محتوای آزمون، دانش آموزان ایرانی پایه هشتم بالاترین میانگین را در حوزه شناختی دانش (٤٧٩) و کمترین میانگین را در حوزه شناختی کاربرد (٤٧٠) به دست آورده‌اند که هردو به‌طور معنی‌داری متفاوت از میانگین سه حوزه (دانش، کاربرد، استدلال) (٤٧٤) است. شواهد داخلی نیز نشانگر چنین کاستی‌هایی است. برای نمونه، میانگین نمرات امتحانات هماهنگ کشوری در سال ۱۳۹٤ برای دانش‌آموزان رشته تجربی دختر ۱۲/۷۱ و پسر ۱۱/٤۸ بوده است؛ و نزدیک به ۱۵ درصد دانش آموزان رشته تجربی، شکست تحصیلی را تجربه نموده و مردود شده‌اند (مرکز سنجش وزارت آموزش‌وپرورش، ۱۳۹٤).

برای تغییر این پیامدهای خطیر، حداقل سه گزینه وجود دارد: تغییر سیستم آموزشی، تغییر آزمون‌ها و شیوه‌های سنجش و یا تغییر هردو. در اینجا، تأکید ما بر راه دوم است. البته این به معنای فقدان شایستگی گزینه‌های اول و سوم نیست؛ و باید توجه داشت که آزمون‌ها خودشان به‌تنهایی موجب بهبود آموزش و یادگیری نمی‌شوند؛ و بسیاری از عوامل دیگر ازجمله کیفیت برنامه درسی، مهارت و تجربه معلمان و حمایت دانش آموزان در خارج از کلاس، بر آموزش و یادگیری تأثیر می‌گذارند (پلگرینو و همکاران، ۲۰۰۱: ۳۱) اما همان‌طور که اسکات[24] استدلال می‌کند، آزمون‌ها را می‌توان، به‌عنوان ابزاری مترقی، منصفانه، منطقی و معقول برای اصلاحات آموزشی به کار برد (۲۰۱۱). در همین رابطه، وولفولک (۲۰۰٤) بیان می‌دارد، اگر آزمون‌ها مشخص کنند که معلمان واقعاً چه چیزهایی را آموزش می‌دهند و دانش‌آموزان چه چیزهایی را می‌آموزند - که واقعاً چنین است - پس راه بهبود آموزش، راهی مستقیم اما رو به بالا است: ارزیابی توانایی‌ها و عادات مهم و اساسی؛ بنابراین با ساخت و به¬کار بردن آزمون‌هایی که به‌جای مطالب سطحی، فرآیندهای بالاتر

23. Third International Mathematics and Science Study (TIMSS)
24. Scott

شناختی را مورد تأکید و سنجش قرار دهد و استفاده از آن‌ها جهت ارزشیابی دانش آموزان می‌توان در فرآیند یادگیری آن‌ها تغییر بنیادی و مطلوب ایجاد کرد. بی‌شک «کوشش در جهت اصلاح برنامه درسی مدارس تنها زمانی می‌تواند مؤثر باشد که امتحانات و دیگر روش‌های ارزشیابی تقویت گردد، لذا طراحی امتحانات جدید به‌عنوان بخش اساسی از فرآیند اصلاح آموزش برنامه درسی مدارس تصدیق می‌گردد» (دایره المعارف تعلیم و تربیت، جلد ۸: ۱۱۱).

با توجه به آنچه مطرح شد، ضرورت در دسترس داشتن آزمون‌های استاندارد شده پیشرفت تحصیلی احساس می‌شود؛ لذا پژوهش حاضر در پی آن بود که به‌منظور اندازه‌گیری پیشرفت و مهارت‌های اساسی دانش آموزان در برنامه درسی زیست‌شناسی، آزمونی بر اساس اهداف و محتوای آموزشی این ماده درسی برای پایه دهم دوره متوسطه نظری تدوین نماید؛ تا بدین‌وسیله ابزار مناسبی جهت اندازه‌گیری و سنجش آموخته‌های دانش آموزان فراهم نماید.

مبانی نظری تدوین و استانداردسازی آزمون‌های پیشرفت بر اصول و روش‌های روان‌سنجی نهاده شده است. امروزه دو روش، مدل کلاسیک آزمون (گالیکسون[25]، ۱۹۵۰) و نظریه سؤال-پاسخ[26] (IRT)، برای ساختن آزمون‌ها و تفسیر نمرات، به متخصصان اندازه‌گیری کمک کرده است (همبلتون[27]، ۱۹۸۹). در سال‌های اخیر به‌دلیل محدودیت‌های نظریه کلاسیک، کاربرد آن کاهش یافته و از طرفی به‌دلیل ظهور کامپیوتر و نرم‌افزارها، استفاده از IRT رواج یافته است. همبلتون، راجرز و سوامینتان[28] (۱۹۹۱) سه محدودیت عمده نظریه کلاسیک آزمون تغییرپذیری شاخص‌های سؤال، وابستگی توانایی آزمودنی به آزمون، عدم امکان پیش‌بینی عملکرد آزمودنی در یک سؤال آزمون- را شناسایی کرده‌اند. علاوه بر این، مفاهیم و تعاریف اساسی تئوری کلاسیک غیرقابل آزمون هستند و به‌سادگی، درست فرض می‌شوند و هیچ راهی برای تعیین تجربی ارتباط بین مفروضات نظریه کلاسیک آزمون با واقعیت وجود ندارد (همبلتون و کوک، ۱۹۷۷؛ همبلتون و همکاران، ۱۹۹۱؛ دی مارس، ۲۰۱۰). این در حالی است که مطلوب آن است که (الف) آماره‌های سؤال وابسته به گروه نباشند، (ب) نمرات آزمون به دشواری آزمون وابسته نباشد، (ج) مدل‌های آزمون، مبنایی برای تطبیق سؤال‌های آزمون با سطح توانایی فراهم کنند، (د) مدل‌های آزمون مبتنی بر فرضیات غیرقابل قبول نباشد. در حال حاضر شواهد قابل توجهی وجود دارد که نشان می‌دهد این ویژگی‌های مطلوب و موارد دیگر را می‌توان در چارچوب نظریه سؤال-پاسخ، به دست آورد (همبلتون و سوامینتان، ۱۹۸۹). در تئوری سؤال-پاسخ فرض می‌شود که (الف) زیربنای عملکرد آزمودنی در آزمون، یک توانایی یا ویژگی واحد است و (ب) رابطه بین عملکرد آزمون‌شونده در هر سؤال و توانایی اندازه‌گیری شده توسط آزمون را می‌توان با یک تابع افزایشی تکنوا[29] توصیف کرد. این تابع یک تابع مشخصه سؤال[30] نامیده می‌شود و احتمال پاسخگویی صحیح آزمودنی‌ها را در سطوح مختلف توانایی ارائه می‌دهد. آزمون‌شوندگانی که توانایی بیشتری دارند، نسبت به آزمون‌شوندگان با توانایی کمتر، احتمال بیشتری برای پاسخ صحیح به سؤال دارند (دی مارس[31]، ۲۰۱۰). توابع مشخصه سؤال، همان‌طور که معمولاً در مدل‌های تک‌بعدی[32] آزمون مرسوم است، معمولاً با یک، دو یا سه پارامتر توصیف می‌شوند (دی آیالا[33]، ۲۰۰۹). مدل سه

25. Gulliksen
26. Item Response Theory
27. Hambleton
28. Hambleton, Rogers & Swaminathan
29. monotonic
30. Item Characteristic Function
31. DeMars
32. unidimensional

پارامتری پیچیده‌ترین مدل IRT است. در این مدل سه پارامتر دشواری، تمیز و حدس وجود دارد که باید برآورد شود. یوری[34] (۱۹۷۷) ضمن مقایسه مدل‌های یک، دو و سه پارامتری نشان داد که مدل سه پارامتری به‌بهترین وجه قادر است آزمون‌های چندگزینه‌ای را با استفاده از داده‌های واقعی آزمون توصیف کند.

مبتنی بر اهداف این مطالعه و به‌منظور سنجش پیشرفت تحصیلی و سطح توانایی دانش‌آموزان در زیست‌شناسی، با توجه به اهداف آموزشی و برنامه‌های درسی این موضوع، آزمونی تدوین و هنجاریابی شد. ازآنجاکه جهت تولید آزمون، باید از ویژگی‌های ساختی و فنی آن اطمینان حاصل کرد؛ سؤال‌های پژوهشی زیر مدنظر قرار گرفت.

۱- آیا آزمون زیست‌شناسی تدوین‌شده از پایایی و اعتبار کافی برخوردار است تا بتوان از آن به‌عنوان ابزاری پایا و معتبر استفاده کرد؟

۲- آیا محتوای آزمون زیست‌شناسی تدوین‌شده مبتنی بر تحلیل عاملی از یک عامل کلی اشباع است؟

۳- آیا تفاوتی بین عملکرد دانش‌آموز پسر و دختر در آزمون زیست‌شناسی تدوین‌شده وجود دارد؟

٤- آزمون زیست‌شناسی تدوین‌شده تا چه اندازه با مدل لجستیک سه پارامتری در تئوری سؤال-پاسخ برازش دارد؟

۵- اندازه‌های برآورد پارامترهای سؤال (دشواری، تشخیص، حدس) در آزمون زیست‌شناسی تدوین‌شده چقدر است؟

٦- میزان آگاهی‌های برآورد شده سؤالات آزمون تدوین‌شده در چه سطحی قرار دارند؟

۷- هنجارهای آزمون زیست‌شناسی تدوین‌شده برای دانش‌آموزان مدارس دولتی چگونه است؟

روش پژوهش

طرح پژوهشی حاضر از نوع «تحقیق و توسعه[35]» بوده است که در زمره پژوهش‌های کاربردی قرار می‌گیرد. برخی از مؤلفه‌های اصلی طرح پژوهش عبارت بود از ۱- تعریف حیطه مورداندازه‌گیری، ۲- بیان اهداف و موارد استفاده ابزار اندازه‌گیری، ۳- تعریف رفتاری دانش و مهارت‌های مورداندازه‌گیری، ٤- تعیین الگوی سؤال‌های آزمون، ۵- اجرای مقدماتی و تجربی آزمون به‌منظور اصلاح و یا حذف سؤالات نامناسب، ٦- هنجاریابی آزمون و ارزیابی پایایی و اعتبار آزمون، ۷- تهیه دستورالعمل‌ها و راهنمای اجرا، نمره‌گذاری و تفسیر آزمون.

شرکت‌کنندگان

جامعه آماری این پژوهش را کلیه دانش‌آموزان پایه دهم دبیرستان‌های دولتی شهر شیراز در سال تحصیلی ۱۳۹٤-۱۳۹۳ تشکیل می‌داد. حجم کل جامعه آماری ۳۹۰۳۹ دانش‌آموز بود. از این جامعه ۳۰۰ دانش‌آموز دختر و پسر به‌صورت تصادفی برای نمونه تجربی و ۹۳۸ دانش‌آموز (٤۸۲

33. de Ayala
34. Urry
35. research & development

دختر و ٤٥٦ پسر) برای مرحله نهایی هنجاریابی به روش نمونه‌گیری چندمرحله‌ای[۳۶] از دانش آموزان مدارس دولتی چهار ناحیه آموزشی شهر شیراز انتخاب شدند. معیار تعیین حجم نمونه در مرحله تجربی، پارامترهای دشواری و تمیز سؤال بود (ثرندایک، ۱۹۸۲ / ترجمه هومن، ١٣٦٦) و برای نمونه هنجاریابی معیارهای مهم تعیین حجم نمونه عبارت بودند از:

- حجم نمونه برای تحلیل عاملی کافی باشد.
- حجم نمونه برای تجزیه‌وتحلیل داده‌ها با مدل لجستیک سه پارامتری در نظریه سؤال- پاسخ کافی باشد.

ابزار اندازه‌گیری

برای تدوین سؤالات آزمون ضمن تعیین محتوای کتاب درسی زیست‌شناسی پایه دهم (چاپ ۱۳۹۳) و تعریف حیطه مورداندازه‌گیری (منطبق با حیطه شناختی بلوم)، به کمک محقق و تیمی از دبیران زیست‌شناسی ۱۵۰ سؤال چهارگزینه‌ای بر اساس جدول مشخصات آزمون و منطبق بر اصول روان‌سنجی طراحی گردید. سؤالات تولیدشده پس از اصلاح و ویراستاری بر روی یک نمونه در دسترس اجرا شد تا ازنظر قابل‌فهم بودن محتوا و مدت‌زمان لازم برای پاسخگویی، اطمینان حاصل شود. سپس سؤال‌ها در سه فرم موازی - ازنظر محتوا و هدف- توزیع شد. جهت اجرای تجربی، آزمون اولیه تدوین‌شده بر روی ۳۰۰ دانش‌آموز (۱۰۰ دانش‌آموز برای هر فرم) دبیرستان‌های متوسطه دولتی شیراز که به‌صورت تصادفی ساده انتخاب‌شده بودند، اجرا شد. پس از جمع‌آوری داده‌ها و ضمن بررسی شواهد اعتبار یابی، سؤالات آزمون بر اساس معیارهای فنی زیر مورد تجزیه‌وتحلیل قرار گرفت:

۱- دشواری هر سؤال حداقل ۰/۲۵ و حداکثر ۰/۷۵ باشد.
۲- شاخص‌های تمیز هر سؤال کمتر از ۰/۳۰ نباشد.
۳- ضریب همبستگی دورشته‌ای هر سؤال حداقل ۰/۲۵ باشد.
٤- گزینه‌های انحرافی هر سؤال دارای قدرت جذب کافی باشند.

بر اساس معیارهای ذکرشده، ۱۰۰ سؤال برای فرم‌های نهایی آزمون انتخاب شدند. سؤال‌های نهایی در دو شکل موازی الف و ب هریک با ۵۰ سؤال چهارگزینه‌ای توزیع و از ساده به دشوار مرتب شدند. ضمن تعیین شواهد اعتبار یابی، فرم‌های نهایی آزمون تدوین‌شده بر روی نمونه هنجاری به شکلی اجرا شد که فرم‌های آزمون در هر کلاس به‌صورت تصادفی به آزمودنی‌ها اختصاص یابد. زمان پاسخگویی به مجموعه سؤالات آزمون «٦٠» دقیقه برای هر فرم بود. به هر پاسخ صحیح نمره «۱» و به هر پاسخ اشتباه نمره صفر تعلق گرفت. نمرات خام کل برای هر فرم آزمون از مجموع تعداد پاسخ‌های صحیح به دست آمد. در هر فرم آزمون «۸» درصد از کل سؤال‌ها شامل مفاهیم روش علمی، «۱۷» درصد مفاهیم زیست‌شناسی سلولی و مولکولی، «۱۵» درصد مفاهیم تغذیه و گوارش، «۱۹» درصد مفاهیم تنفس و دفع، «۲۹» درصد مفاهیم گردش مواد و ۱۲ درصد مفاهیم

36. multistage sampling

زیست گیاهی را در برمی‌گرفت. سؤال‌های آزمون سطح دانش و فرآیندهای شناختی بالاتر (فهم، کاربرد و تحلیل) را مورد ارزیابی قرار می‌داد (جدول ۱).

جدول۱ - ترکیب و توزیع سؤال‌های آزمون زیست‌شناسی بر حسب محتوا و عملکرد

عملکرد / محتوا	دانش			فهم	کاربرد	تحلیل	جمع
	اصلاحات و واقعیت ها	روش‌ها و ملاک‌ها	مفاهیم و اصول				
نگرش علمی		۳		۵			۸
ساختار سلولی	۸				۲	۱	۱۱
ساختارشیمیایی	۲	۱	۱	۲			۶
تغذیه و گوارش	۵	۳	۱	۴	۲		۱۵
دفع مواد	۱			۴	۱	۲	۸
زیست گیاهی	۷		۱	۱	۳		۱۲
گردش مواد	۸	۲	۱	۵	۶	۷	۲۹
تنفس	۲	۳	۱	۴	۱		۱۱
جمع	۲۳	۱۲	۵	۲۷	۱۴	۹	۱۰۰

یافته‌های پژوهش

مشخصه‌های توصیفی سؤالات آزمون:

جدول ۲ خلاصه مشخصه‌های دشواری و تمیز سؤالات آزمون مبتنی بر مدل کلاسیک اندازه‌گیری را نشان می‌دهد. بر مبنای این جدول دامنه دشواری سؤالات آزمون بین ۰/۲ تا ۰/۸ گسترده شده بود و همه سؤالات از قدرت تمیز بالاتر از ۰/۳ برخوردار بودند. متوسط دشواری سؤالات آزمون ۰/٤٤ و متوسط قدرت تشخیص آن‌ها ۰/۵۵ به‌دست‌آمد؛ بنابراین مبتنی بر معیارهای روان‌سنجی (کاپلان و ساکوزو[37]، ۲۰۱۸) سؤالات آزمون کمی دشوار اما از قدرت تشخیص خوبی برخوردار بوده است

جدول۲ - خلاصه دشواری و قدرت تشخیص سؤال‌های آزمون زیست‌شناسی

فرم آزمون	دشواری سؤال			قدرت تشخیص سؤال		
آزمون	حداقل	حداکثر	متوسط	حداقل	حداکثر	متوسط
الف	۰/۱۹	۰/۸۳	۰/٤۳	۰/۳۱	۰/٦۲	۰/۵۳
ب	۰/۲۳	۰/۷۹	۰/٤۵	۰/۲۹	۰/٦۷	۰/۵۷

37 .Kaplan and Saccuzzo

پایایی آزمون:

ازآنجایی‌که در این مطالعه هر سؤال به‌صورت دو ارزشی (با مقادیر ۱ یا ۰) امتیازدهی شده بود، فرمول ۲۰ کودر-ریچاردسون[38] (۱۹۳۷) برای برآورد پایایی محاسبه شد. ضرایب پایایی برای فرم الف و ب به ترتیب «۰/۸۹» (خطای استاندارد اندازه‌گیری ٤/۳۵) و «۰/۸۸» (خطای استاندارد اندازه‌گیری ٤/۳۳) برآورد شد (جدول ۳). با توجه به معیار هلمشتاتر[39] (۱۹٦٤) برای ضرایب پایایی آزمون‌های پیشرفت، ضرایب برآورد شده برای هر دو فرم آزمون قابل‌قبول بود ضمن آنکه نشان می‌داد که سؤال‌های آزمون زیست‌شناسی از ثبات داخلی رضایت بخشی برخوردار بوده است.

جدول ۳- ضرایب پایایی و خطاهای استاندارد آزمون زیست‌شناسی

فرم آزمون	حجم گروه نمونه			تعداد سؤال	میانگین	انحراف استاندارد	ضریب پایایی	خطای استاندارد
	پسر	دختر	کل					
الف	۲۴۶	۲۲۲	۴۶۸	۵۰	۲۲/۸۵	۹/۴۰	۰/۸۸۶	۴/۳۵
ب	۲۴۴	۲۲۶	۴۷۰	۵۰	۲۲/۷۰	۹/۲۴	۰/۸۸۴	۴/۳۲

اعتبار آزمون:

اگرچه چندین روش خاص برای اعتبار یابی آزمون‌های آموزشی شرح داده‌شده است، اما اعتبار محتوا[40] در سنجش پیشرفت تحصیلی حائز اهمیت بسیار است؛ بنابراین در این پژوهش علاوه بر اعتبار سازه (از طریق تحلیل عاملی)، از اعتبار محتوایی نیز استفاده‌شد. اعتبار محتوایی آزمون از ابتدای تدوین سؤالات، ضمن تعریف دقیق حیطه محتوایی و تهیه جدول مشخصات آزمون (جدول ۱) و تدوین سؤالات آزمون بر اساس اهداف و محتوای آموزشی در آن تنیده شد. همچنین به‌منظور تعیین تجربی اعتبار محتوایی آزمون و برآورد میزان تطابق محتوای سؤالات با حیطه موردسنجش از ضریب توافق (۰/٦۲) بین نظرات کارشناسان گروه‌های آموزشی استفاده شد که نشانگر متناسب بودن سؤالات با محتوای تدریس است. نتایج تحلیل عاملی بر اساس روش مؤلفه اصلی نیز نشان داد که «۱۵/۷» درصد از کل واریانس در فرم الف و ۱۵/٦ درصد از کل واریانس در فرم ب توسط یک عامل کلی که کاملاً متمایز از بقیه عوامل است توضیح داده می‌شود. همچنین ٦۸/۷ درصد از واریانس مشترک بین سؤالات در فرم الف و ٦۹/۳ درصد از واریانس مشترک بین سؤالات در فرم ب با این عامل کلی تبیین می‌شود (جدول ٤).

38. Kuder – Richardson
39. Helmstadter
40. content-related validity evidence

جدول ٤- ارزش ویژه، درصد واریانس و درصد واریانس تراکمی سؤال‌های آزمون زیست‌شناسی

عامل	ارزش ویژه		درصد واریانس		درصد واریانس تراکمی	
	فرم الف	فرم ب	فرم الف	فرم ب	فرم الف	فرم ب
۱	۷/۸۴۰	۷/۷۸۷	۱۵/۶۸	۱۵/۵۷	۱۵/۶۸	۱۵/۵۷
۲	۱/۸۳۶	۱/۸۹۰	۳/۶۷	۳/۷۸	۱۹/۳۵	۱۹/۳۵
۳	۱/۷۳۸	۱/۵۴۶	۳/۴۸	۳/۰۹	۲۲/۸۳	۲۲/۴۵

تفاوت بین گروه‌ها

برای بررسی اینکه آیا عامل جنسیت در نمره کل آزمون زیست‌شناسی دخیل است یا خیر، از آزمون t برای گروه‌های مستقل استفاده شد. بر اساس یافته‌ها احتمال اینکه اختلاف میانگین بین دو گروه‌ها ناشی از شانس باشد بسیار زیاد بود و شواهد کافی برای رد فرضیه صفر در سطح معنی‌داری آلفای ۰/۰۵ وجود نداشت؛ بنابراین بین عملکرد پسران و دختران در آزمون زیست‌شناسی تفاوت معنی‌داری وجود نداشت (جدول ۵). نبود تفاوت معنی‌دار بین میانگین نمرات پسران و دختران مؤید این است که سازه اندازه‌گیری شده توسط آزمون زیست‌شناسی، مستقل از جنسیت آزمودنی‌ها است.

جدول ٥- آزمون تفاوت بین عملکرد دختران و پسران در فرم‌های آزمون زیست‌شناسی

فرم آزمون	پسران		دختران			
	M	SD	M	SD	df	t
الف	۲۲/۷۳	۱۰/۶۳	۲۲/۹۷	۸/۰۷	۴۲۵/۳	٭۰/۷۸
ب	۲۲/۳۴	۱۰/۳۹	۲۳/۰۳	۸/۰۲	۴۲۴/۷	٭۰/۴۳

P <۰/۰۵ ٭

هنجار آزمون

در پژوهش حاضر از یک شاخص نسبی (نمرات مشتق شده) جهت ارائه مجموعه هنجارها استفاده شد. برای این منظور، نمرات خام آزمون با استفاده از توزیع فراوانی به نمرات صدک و نمرات استاندارد نرمال شده (نمرات Z و نمرات T با میانگین «۵۰» و انحراف استاندارد «۱۰») تبدیل شدند. به‌عنوان نمونه، جدول ۶ خلاصه هنجار آزمون زیست‌شناسی برای فرم الف را نشان داده است. بر اساس این جدول، دانش‌آموزی با نمره خام ۴۶ بالاتر از ۹۶ درصد افراد جامعه قرار می‌گیرد و نمره استاندارد این فرد ۱/۷۴ واحد انحراف استاندارد بالاتر از میانگین جامعه قرار می‌گیرد. همچنین این فرد دارای نمره ۶۷/۴ در مقیاس T است.

جدول ۶- خلاصه هنجار آزمون پیشرفت علوم زیستی (فرم الف آزمون)

نمره خام	درصد فراوانی	رتبه درصدی	Z نمره	T نمره
۷	۰/۹	۰/۵۳	-۲/۶۰۳	۲۳/۹۷
۸	۰/۴	۱/۱۸	-۲/۲۸۹	۲۷/۱۱
۹	۱/۱	۱/۹۲	-۲/۰۸۶	۲۹/۱۴
۱۰	۱/۱	۲/۹۹	-۱/۸۹۳	۳۱/۰۷
۱۱	۱/۷	۴/۳۸	-۱/۷۱۶	۳۲/۸۴
:	:	:	:	:
۴۶	۰/۹	۹۶/۰۵	۱/۷۴۰	۶۷/۴۰
۴۷	۱/۵	۹۷/۲۲	۱/۸۹۳	۶۸/۹۳
۴۸	۰/۲	۹۸/۰۸	۲/۰۴۰	۷۰/۴۱
۴۹	۰/۹	۹۸/۶۱	۲/۱۶۲	۷۱/۶۲
۵۰	۱/۱	۹۹,۵۷	۲/۵۳۱	۷۵/۳۱

همترازسازی[41]

پس از اطمینان از اینکه هر دو فرم آزمون دارای شاخص‌های آماری (تعداد سؤال‌ها، میانگین، انحراف معیار، ضریب پایایی و خطای استاندارد) یکسان هستند، برای اینکه نمره آزمون‌شونده تحت تأثیر فرمی که در آن شرکت کرده قرار نگیرد از روش‌های همترازسازی خطی و غیرخطی (انتقال بر اساس سطح) پیشنهادشده توسط ثرندایک (۱۹۸۲) استفاده شد. برای همترازسازی خطی، پس از محاسبه میانگین و انحراف معیار هر یک از گروه‌های شرکت‌کننده در فرم‌های آزمون، فرم الف (A) به‌عنوان آزمون لنگر انتخاب شد و با استفاده از معادله «(۱)» نمره معادل ب (B) محاسبه شد (جدول ۷).

معادله ۱ $x_B = \bar{x}_B + s_B / s_A (x_A - \bar{x}_B)$

جایی که «XB» و «XA» میانگین هستند و «SB» و «SA» انحرافات استاندارد هستند.

بر اساس معادله ۱ جدول همترازسازی دو فرم آزمون تهیه شد که خلاصه‌ای از آن در جدول ۷ منعکس‌شده است. مطابق این جدول: نمره خام ۹ در فرم ب معادل با نمره خام ۸/۹۲ در فرم الف است که نشانگر ۲/۲۰ واحد انحراف استاندارد پایین‌تر از میانگین جامعه (نمره Z) و برابر با نمره ۲۸/۰۸ در مقیاس نمره T (مقیاسی با میانگین ۵۰ و انحراف معیار ۱۰) است.

41. equating

جدول ۷- خلاصه همترازسازی خطی آزمون پیشرفت علوم زیستی (تبدیل نمره خام فرم ب به فرم الف)

نمره خام ب	نمره خام الف	Z نمره	T نمره
۵	۴/۸۵	-۲/۹۸۵	۲۰/۱۵
۶	۵/۸۷	-۲/۶۹۳	۲۳/۰۷
۷	۶/۸۹	-۲/۵۳۲	۲۴/۶۸
۸	۷/۹۰	-۲/۳۷۱	۲۶/۲۹
۹	۸/۹۲	-۲/۱۹۲	۲۸/۰۸
:	:	:	:
۴۶	۴۶/۵۶	۱/۷۶۷	۶۷/۶۷
۴۷	۴۷/۵۷	۱/۸۹۵	۶۸/۹۵
۴۸	۴۸/۵۹	۲/۰۴۲	۷۰/۴۲
۴۹	۴۹/۶۱	۲/۱۶۴	۷۱/۶۴
.	.	.	.

برای همترازسازی غیرخطی نمره خام مربوط به صدک های انتخابی (۲، ۵، ۱۰، ۲۵، ۵۰، ۷۵، ۹۰، ۹۵ و ۹۸) آزمودنی‌ها در هر دو فرم آزمون برابر در نظر گرفته شد (جدول ۸). مطابق این جدول فردی که در هر دو فرم آزمون نمره ۲۱ کسب نماید در صدک ۵۰ام قرار می گیرد به این معنا که نمره وی از نمره نیمی از افراد جامعه بالاتر است.

جدول ۸- همترازسازی غیرخطی فرم‌های آزمون زیست‌شناسی

صدک برگزیده	۲ام	۵ام	۱۰ام	۲۵ام	۵۰ام	۷۵ام	۹۰ام	۹۵ام	۹۸ام
فرم الف	۹	۱۱	۱۳	۱۶	۲۱	۲۷	۳۷	۴۵	۴۸/۶۲
فرم ب	۱۰	۱۱/۵۵	۱۳	۱۶	۲۱	۲۸	۳۵	۴۴/۴۵	۴۸/۵۸

ارزیابی خوبی برازش

یکی از مهم‌ترین موضوعات در نظریه سؤال- پاسخ، مناسب بودن و برازش[42] یک مدل تابع ریاضی با داده‌های جمع‌آوری‌شده از یک اجرای آزمایشی است. در این تحقیق برای ارزیابی برازش مدل

42. fitness

لجستیک سه پارامتری با داده‌های آزمون زیست‌شناسی از نرم‌افزار ASCAL [۴۳] استفاده شد. ASCAL از یک رویکرد بیشینه درستنمایی مشترک[۴۴] با توزیع پیشین[۴۵] و آزمون کای اسکور جهت برازش سؤال‌ها با مدل استفاده می‌کند. بر اساس نتایج حاصل: از ۱۰۰ سؤال هر دو فرم آزمون، تنها ۸ سؤال با مدل برازش نداشت و ۹۲٪ سؤال‌ها با مدل ریاضی برازش خوبی نشان داد. عدم برازش برخی سؤالات ممکن است به دلیل وجود تعدادی سؤال ساده (اثر کف) و تعدادی سؤال دشوار (اثر سقف) بوده باشد که طبیعتاً نسبت به دیگر سؤالات از قدرت تشخیص کمتری برخوردارند (امبرتسون و رایس، ۲۰۰۰/ ترجمه شریفی و همکاران، ۱۳۸۸).

برآورد پارامترهای سؤال

مبتنی بر مدل ۳ پارامتری نظریه سؤال- پاسخ، برای هر سؤال آزمون سه پارامتر دشواری سؤال، شیب یا تمیز سؤال و حدس کاذب برآورد می‌شود که در ادامه برآورد پارامترهای مذکور توصیف‌شده است.

۱- پارامتر دشواری سؤال: (bi) نشان‌دهنده نقطه‌ای از مقیاس توانایی است که در آن آزمون‌شونده احتمال ۲/(ci+۱) برای پاسخ صحیح به یک ماده آزمون دارد (همبلتون، ۱۹۸۹). محدوده دشواری سؤال‌های آزمون برای فرم الف، بین ۰/۹۸۹- (سؤال ۱) تا ۱/۹۵٤+ (سؤال ٤٥) و برای فرم ب بین ۰/۹٦۲- (سؤال ۲) تا ۲/۱٦۲ (سؤال ٤۹) برآورد شد. متوسط پارامتر دشواری سؤال‌ها برای فرم الف ۰/۷٤+ و برای فرم ب ۰/۷۳+ به دست آمد؛ که نشان می‌دهد آزمون نسبتاً دشوار بوده است.

۲- پارامتر تمیز سؤال: (ai) نشان‌دهنده قدرت تشخیص سؤال جهت تفکیک افراد قوي و ضعیف است. دامنه پارامتر تمیز (شیب) سؤال برای فرم الف بین ۰/٤۰٤ (سؤال ۲) تا ۱/۹۹۳ (سؤال ۳۵) و برای فرم ب بین ۰/٤۰۸ (سؤال ٤) تا ۲/۱۸۲ (سؤال ٤۸) برآورد شد. متوسط تمیز سؤال برای فرم الف، ۱/۱۳ و برای فرم ب، ۱/۱٦ به دست آمد که نشان می‌دهد سؤالات آزمون از قدرت تمیز کافی برخوردار بوده است.

۳- پارامتر حدس کاذب (ci): احتمال دادن پاسخ صحیح از روی حدس به سؤال توسط آزمون‌شوندگاني با کمترین توانايي را نشان می‌دهد (فالکنر-باند و ولز[۴۶]، ۲۰۱٦). دامنه پارامتر حدس از حداقل ۰/۱۰ (سؤال ۲ فرم ب تا حداکثر ۰/۲۷ (سؤال ۲۷ فرم الف) برآورد شد. متوسط پارامتر حدس برای فرم الف، ۰/۲۳ و برای فرم ب، ۰/۲۱ به دست آمد.

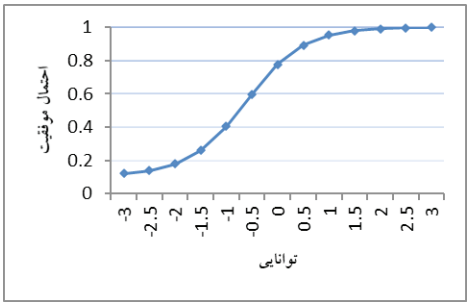
منحنی مشخصه سؤال و آزمون

منحنی مشخصه سؤال (ICC) یک تابع ریاضی است که احتمال موفقیت در یک سؤال را به توانایی اندازه‌گیری شده توسط مجموعه سؤال‌ها یا آزمونی که شامل سؤال‌ها است مرتبط می‌کند (همبلتون و سوامینتان، ۱۹۸۵) در این مطالعه ICC ها به شکل توابع توزیع لجستیک سه پارامتری بر اساس معادله ۳ برآورد شدند.

43. A Microcomputer Program for Estimating Logistic IRT Item Parameters
44. joint maximum likelihood
45. prior distributions
46. Faulkner-Bond and Wells

$$p_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} (i = 1,2,...,n)$$ معادله ۳

به‌طوری که pi (θ) احتمال این است که یک آزمودنی با توانایی θ به سؤال i به‌درستی پاسخ دهد، bi پارامتر دشواری سؤال، ai پارامتر تمیز سؤال و ci پارامتر حدس کاذب سؤال است و D مقدار ثابت برابر با ۱/۷ است. برای نمونه، شکل۱ منحنی ویژه سؤال ٤ فرم الف آزمون را نشان داده است. بر اساس این نمودار، احتمال اینکه یک آزمودنی با توانایی متوسط به این سؤال پاسخ صحیح دهد برابر ۰/۸ است. این سؤال دارای قدرت تشخیص مناسب (ai = ۱/۰۵) است، احتمال دادن پاسخ صحیح به سؤال، صرفا از روی حدس اندک است (ci = ۰/۱۱) و دشواری آن کمی پایین‌تر از متوسط (b = -۰/٦۱) برآورد می‌شود.



شکل ۱- منحنی مشخصه سؤال آزمون زیست‌شناسی (سؤال ٤ فرم الف)

منحنی مشخصه آزمون، نمره خام مورد انتظار در یک آزمون را به خصیصه مکنون (توانایی) مورداندازه‌گیری مرتبط می‌سازد به‌عبارت‌دیگر منحنی ویژه آزمون منعکس‌کننده رابطه تکنوا بین نمره واقعی و نمرات توانایی برای مجموعه خاصی از سؤال‌های آزمون است (همبلتون، ۱۹۸۹). برای محاسبه تابع منحنی ویژه آزمون، مجموع احتمال موفقیت برای سطوح مشخصی از توانایی برای سؤال‌های مختلف محاسبه شد؛ و سپس منحنی ویژه آزمون ترسیم گردید که بر اساس آن می‌توان نمره واقعی فرد در آزمون را برآورد کرد.

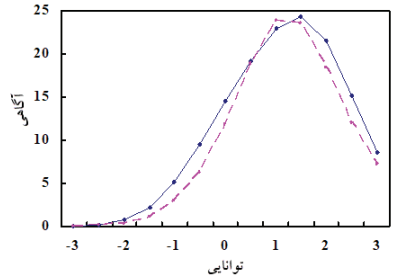
تابع آگاهی سؤال و آزمون

توابع آگاهی[47] نقش برجسته‌ای در IRT دارند و درباره وضعیت نسبی فرد در خصیصه مکنون اطلاعاتی به دست می‌دهند. در این پژوهش با استفاده از رابطه برین بام (۱۹٦۸) توابع آگاهی سؤال برای سطوح معینی از توانایی(θ) به دست آمد (معادله ٤).

$$I_i(\theta) = \frac{2.89a_i^2(1-c_i)}{\left[c_i + e^{1.7a_i(\theta-b_i)}\right]\left[1+e^{1.7a_i(\theta-b_i)}\right]^2} (i = 1,2,3,...,n)$$ معادله ٤

47. Item Informations

جایی که Ii (θ) آگاهی ارائه‌شده توسط سؤال i در سطح θ است و ai،bi و ci قبلاً تعریف‌شده بودند. هرگاه مفروضه استقلال موضعی[48] صادق باشد، تابع آگاهی آزمون برابر با مجموع توابع آگاهی سؤال‌های تشکیل‌دهنده یک آزمون است (ثرندایک، ۱۹۸۲ / ترجمه هومن، ۱۳۶۹). شکل ۲ نمودار تابع آگاهی هر دو فرم آزمون زیست‌شناسی را نشان داده است. بر اساس نمودار شکل ۲ هردو آزمون در محدوده گسترده‌ای از توانایی ۰/۵- تا ۲/۵+ اطلاعاتی پیرامون آزمودنی‌ها ارائه می‌کند اما بیشترین اطلاعات فراهم‌شده توسط هر دو آزمون در محدوده توانایی ۱/۵+ به دست می‌آید؛ بنابراین می‌توان نتیجه گرفت آزمون برای دامنه متوسط به بالای توانایی از پایایی بیشتری برخوردار است.



شکل ۲- تابع آگاهی آزمون زیست‌شناسی (فرم الف- خط ممتد- و فرم ب- خط ناپیوسته)

بحث و نتیجه‌گیری

هدف از این پژوهش ارائه ابزاری برای سنجش آموخته‌ها و آگاهی از پیشرفت تحصیلی زیست‌شناسی دانش‌آموزان پایه اول متوسطه دوم در رشته تجربی بود. بر این اساس دو فرم موازی از آزمون زیست‌شناسی بر اساس مدل‌های کلاسیک و IRT تدوین و هنجاریابی شد. برآورد ضریب پایایی نشان داد که بیش از ۸۸ درصد واریانس نمرات آزمون مربوط به واریانس واقعی صفت مورد اندازه‌گیری و ۱۲ درصد مربوط به عوامل خطا است. نتیجه حاصله در مقایسه با نتایج گزارش‌شده برای آزمون‌های مشابه نیز رضایت‌بخش است. برای نمونه ضرایب پایایی آزمون استاندارد پیشرفت استنفورد [علوم] بین ۰/۷۱ تا ۰/۹۶ گزارش شد (مهرنز و لهمان[49]، ۱۹۸۶، ۴۹۶)؛ بنابراین مجموعه سؤالات آزمون زیست‌شناسی از ثبات درونی بالا و رضایت بخشی برخوردار است. نتایج حاصل از برآورد ضریب دشواری و تمیز سؤال‌ها مبتنی بر مدل کلاسیک نشان داد سؤال‌ها از قدرت دشواری (۰/۲ الی ۰/۸) و قدرت تمیز مناسبی (بیشتر از ۰/۳) برخوردار است؛ لذا می‌تواند اطلاعات لازم در مورد تفاوت‌های فردی دانش آموزان را در اختیار آزمایش کننده قرار دهد (کاپلان و ساکوزو، ۲۰۱۸). اگرچه روش‌های متعددی در اعتبار یابی آزمون‌ها شرح داده‌شده است، اما اعتبار محتوایی آزمون‌های پیشرفت تحصیلی از اهمیت بیشتری برخوردار است[50] (AERA, APA, & NCME, ۲۰۱۴).

48. Local independence
49. Mehrens and Lehman
50. American Educational Research Association, American Psychological Association, & National Council of Measurement in Education

در ادبیات روان‌سنجی رایج است که جهت تعیین شواهد مرتبط با اعتبار محتوایی آزمون، عموماً نظرات متخصصان حیطه موردسنجش مورد مشورت قرار گیرد (پیترسون و ترگست[51]، ۱۹۸۹؛ آبراهام، ویلیامسون و وستبروک[52]، ۱۹۹٤). در این پژوهش با توجه به تدوین سؤالات امتحانی مبتنی بر تعریف دقیق حوزه محتوایی آزمون و اهداف آموزشی توسط معلمان مجرب و متخصصین آموزش‌وپرورش و برآورد ضریب تطابق محتوای سؤالات با حیطه موردسنجش (۰/٦۲) مبتنی بر نظرات کارشناسان گروه‌های آموزشی، می‌توان نتیجه گرفت که آزمون زیست‌شناسی از اعتبار محتوای کافی برای اندازه‌گیری دانش و مهارت‌های زیستی برخوردار است. نتایج تحلیل عاملی نیز نشان داد که آزمون پیشرفت زیست‌شناسی از یک عامل کلی اشباع‌شده است، بنابراین می‌توان نتیجه گرفت که عملکرد دانش آموزان در این آزمون تا حد زیادی تحت تأثیر یک عامل یا توانایی کلی است که آزمون، اساساً برای اندازه‌گیری آن ساخته‌شده است.

نتایج آزمون t نشان داد که بین عملکرد پسران و دختران تفاوت معناداری در آزمون مشاهده نمی‌شود. این بدان معناست که آزمون نسبت به گروه‌های جنسی بی‌طرفانه عمل می‌کند و عاری از سوگیری است و سازه اندازه‌گیری شده توسط آزمون، مستقل از جنسیت است. نتیجه حاصل با سایر نتایج آزمون‌های مشابه نیز هماهنگ است. برای نمونه در سومین مطالعه بین‌المللی ریاضی و علوم، تفاوت معنی‌داری بین عملکرد دانش آموزان دختر و پسر در مجموعه سؤالات زیست‌شناسی به دست نیامد (کیامنش، ۱۳۷٦).

یافته‌های حاصل از تجزیه‌وتحلیل سؤالات بر اساس مدل IRT نشان داد که برازش بین داده‌های تجربی و مدل نظری رضایت‌بخش است؛ بنابراین، می‌توان نتیجه گرفت که استفاده از مدل لجستیک سه پارامتری قابل توجیه بوده و نتایج حاصل از آن معتبر است (همبلتون و کوک، ۱۹۷۹؛ همبلتون، ۱۹۸۹). متوسط پارامتر تشخیص برای فرم‌های آزمون ۱/۱۳ (فرم الف) و ۱/۱٦ (فرم ب) به دست آمد که نشان می‌دهد سؤال‌های آزمون از قوه تمیز کافی برای تفکیک دانش آموزان قوی و ضعیف به‌ویژه در سطوح میانی و بالای توانایی برخوردار است. همچنین بالا بودن قوه تمیز اکثر سؤال‌ها نشان می‌دهد که پاسخگویی به اغلب سؤال‌های آزمون فقط به خصیصه مکنون توانایی و خیلی کم به عوامل دیگر بستگی دارد (ثرندایک، ۱۹۸۲ / ترجمه هومن، ۱۳٦۹).

بر اساس یافته‌ها، متوسط پارامتر دشواری برای فرم‌های آزمون ۰/۷۳+ بود که با توجه به دامنه مقیاس توانایی (۲±) نشان می‌دهد آزمون تدوین‌شده کمی دشوار بوده و برای دانش آموزان بالاتر از متوسط کارایی بیشتری دارد. میانگین پارامتر حدس کاذب برای فرم‌های آزمون ۰/۲۱ به دست آمد که کوچک‌تر از مقدار «۰/۲۵»در سؤال‌های چهارگزینه‌ای است. این نشان می‌دهد که دانش آموزان از حدس زدن کورکورانه صرف‌نظر کرده و یا گزینه‌های انحرافی به‌خوبی عمل کرده‌اند و توانسته است آزمودنی‌های با توانایی اندک را به خود جذب کند (همبلتون، ۱۹۸۹).

بر اساس منحنی‌های آگاهی آزمون، دامنه‌ای از توانایی که در آن، آزمون اطلاعات مناسبی را برای برآورد نمره واقعی دانش آموزان فراهم می‌کند در محدوده (۰/۵- تا ۲/۵+) گسترده شده

51. Peterson & Treagust
52. Abraham, Williamson & Westbrook

است و بیشینه آگاهی در سطح توانایی ۱/۵+ به دست می‌آید. ازآنجایی که تابع آگاهی به‌عنوان عکس مجذور فاصله اطمینان حول برآورد توانایی تعریف می‌شود (برین بام[53]، ۱۹۶۸) و با خطای استاندارد اندازه‌گیری رابطه معکوس دارد (همبلتون و کوک[54]، ۱۹۷۷) می‌توان نتیجه گرفت که آزمون پیشرفت زیست‌شناسی تخمین دقیق‌تری از نمره واقعی آزمودنی‌هایی ارائه می‌دهد که سطح توانایی آن‌ها بالاتر از متوسط پیوستار توانایی است. به‌عبارت‌دیگر خطای استاندارد اندازه‌گیری برای این سطح از توانایی کوچک‌تر و پایایی نمرات حاصل بیشتر است. همچنین ازآنجایی که تابع آگاهی هردو فرم یکسان است می‌توان نتیجه گرفت که دو فرم آزمون معادل (موازی) یکدیگرند (همبلتون، ۱۹۸۹).

به‌طورکلی بر اساس یافته‌های حاصل از این مطالعه می‌توان نتیجه گرفت که آزمون استاندارد تدوین‌شده زیست‌شناسی، به‌طور نسبی ابزاری معتبر و قابل‌اعتماد است که می‌تواند در زمینه آموزش و ارزیابی مورداستفاده قرارگیرد. این آزمون پیشرفت که جهت اندازه‌گیری سطح آموخته‌ها و توانایی دانش آموزان توسعه‌یافته است، ساختاری دارد که می‌تواند الگویی برای آزمون‌های پایا و معتبری باشد که در این زمینه تهیه خواهد شد.

منابع

امبرتسون، سوزان ای. و استیون، پی. رایس (۲۰۰۰). نظریه های جدید روانسنجی برای روان شناسان. ترجمه حسن پاشا شریفی، ولی الله فرزاد و همکاران (۱۳۸۸). تهران: رشد.

ثرندایک، رابرت، ال. (۱۹۸۲). روان‌سنجی کاربردی. ترجمه حیدر علی هومن (۱۳۶۹). تهران: موسسه انتشارات و چاپ دانشگاه تهران.

سازمان پژوهش و برنامه‌ریزی آموزشی، دفتر برنامه‌ریزی و تألیف کتب درسی. (۱۳۸۹). راهنمای

53. Birnbaum
54. Hambleton & Cook

برنامه‌درس زیست‌شناسی، تهران: مؤلف
سیف، علی‌اکبر. (۱۳۸٤). سنجش فرایند و فراورده یادگیری. تهران: دوران.
دفتر همکاری‌های علمی بین‌المللی وزارت آموزش‌وپرورش. (۱۳۷۹). مجموعه گفتارهای ارزشیابی در آموزش. تهران: مؤلف.
کیامنش، علیرضا. (۱۳۷٦). گزارش سنجش عملکرد در سومین مطالعه بین‌المللی ریاضی و علوم سال چهارم ابتدایی و سوم راهنمایی. تهران: وزارت آموزش و پرورش.
گلاور، جان ای و برونینگ، راجر اچ. (۱۳۷۵). روان‌شناسی تربیتی، اصول و کاربرد آن. ترجمه علینقی خرازی (۱۳۷۵). تهران: مرکز نشر دانشگاهی
گیج، نیت، ل؛ و برلانیر، دیوید، سي. (۱۳۷٤). روان‌شناسی تربیتي. ترجمه غلامرضا خوي نژاد. مشهد: حکیم‌فردوسي.
هومن، حیدر علی. (۱۳۷۲). اندازه‌گیری‌های روانی و تربیتی. تهران دیبا.

Abraham, M. R., Williamson, V. M. & Westbrook, S. L. (۱۹۹۴). A cross-age study of the understanding five concepts, Journal of Research in Science Teaching, ۱۶۵-۱۴۷ :(۲) ۳۱.
Alele-Williams, G. (۲۰۰۲). Measurement and evaluation in mathematics: The way forward. Journal of basic sciences, ۷-۱:(۱)۱
American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (۲۰۱۴). Standards for educational and psychological testing. Washington, DC: Authors
Baker, J.O. (۲۰۰۳). Testing in modern classroom. London: George. Allen and Unwin
Birnbaum, A. (۱۹۶۸). Some latent trait models and their use in inferring an examinee's bility. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. ۴۷۲-۳۹۷). Reading, MA: Addison-Wesley.
Black, Paul; Harrison, Christine; Lee, Clara; Marshall, Bethan and William, Dylan (۲۰۰۳). Assessment for Learning- putting it into practice. Maidenhead, U.K.: Open university Press.
Cizek, Gregory J. (۱۹۹۳). Testing for Learning: A Remonstrance. Educational Measurement: Issues and Practice, Volume ۴۲-۴۰ :(۴)۱۲.
Crooks, T.J. (۱۹۸۸) 'The impact of classroom evaluation practices on students', Review of Educational Research, ۴ :۵۸.
de Ayala, R. J. (۲۰۰۹). The Theory and Practice of Item Response Theory, New York: Guilford Publications, Inc.
DeMars, Christine (۲۰۱۰). Item Response Theory. New York: Oxford

University Press, Inc.
Faulkner-Bond Molly and Wells Craig S. (۲۰۱۶). A Brief History of and Introduction to Item Response Theory. In Ronald K. Hambleton and Stephen G. Sireci (Ed.). Educational measurement. New York: The Guilford Press.
Frederiksen, N. (۱۹۸۴) 'The real test bias: Influences of testing on teaching and learning', American Psychologist, ۲۰۲–۱۹۳ :(۳)۳۹.
Gipps, C.V. (۲۰۰۳). Beyond Testing: Towards a Theory of Educational Assessment, London: Washington, D.C. Taylor & Francis e-Library.
Gronlund, N. E. (۱۹۸۸). How to Construct Achievement Tests (۴th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.
Gulliksen, H. (۱۹۵۰). Theory of mental tests. John Wiley & Sons Inc.
Hambleton, R. K., Rogers, H. J., & Swaminathan, H. (۱۹۹۱) Fundamentals of item response theory, Newbury Park, Cliff: Sage Publications
Hambleton, R.K. (۱۹۸۹). Principles and selected applications of item-response theory. In R. Linn (Ed.) Educational measurement, (۳th ed.). New York: American Council on Education
Hambleton, R.K.& Cook, L.L. (۱۹۷۷). Latent trait models and their use in the analysis of educational test data, Journal of Educational Measurement, ۹۶-۷۵ :۱۴.
Helmstadter, G. C. (۱۹۶۴). Principles of Psychological Measurement. New York: Appleton Century Crofts.
InternationaL Student Achievement in the TIMSS (۲۰۱۱). Science Content and Cognitive Domains. chapter ۳, TIMSS & PIRLS International Study Center. Lynch school of education, Boston College.
Isaacs, T., Zara, C., Herbert, G., Coombs, S. and Smith, C. (۲۰۱۳) Key Concepts in Educational Assessment. SAGE Publications Ltd.
Kaplan, Robert M. and Saccuzzo, Dennis P. (۲۰۱۸). Psychological testing: Principles, applications, and issues, (۹th ed.). Boston: Cengage Learning.
Linn, R. (۲۰۰۰). Assessment and accountability. Educational Researcher, ۱۶-۴ :(۲)۲۹.
Mehrens, W. A. and Lehman, H. J. (۱۹۸۶). Using Standardized Test in Education. (۴th ed). New York : Longman.
Pellegrino, J. W., Chudowsky, N. and Glaser, R. (Eds.). (۲۰۰۱) Knowing what Students Know :The Science and Design of Educational Assessment . Washington, DC. National Academy Press.
Peterson, R. F. & Treagust, D. F. (۱۹۸۹). Grade۱۲- students'

misconceptions of covalent bonding and structure, Journal of Chemical Education, ۴۶۰-۴۵۹ :(۶) ۶۶.
Phelps, Richard P. (۲۰۱۲). The Effect of Testing on Student Achievement, ۲۰۱۰–۱۹۱۰. International Journal of Testing, ۴۳–۲۱ :۱۲.
Popham, J. (۱۹۸۷) 'The merits of measurement-driven instruction', Phi Delta Kappa, ۸۲–۶۷۹ :۵.
Salmon-Cox, L. (۱۹۸۱). Teachers and standardized achievement tests: what›s really happening? Phi Delta Kappan, ۷۳۶-۷۳۰ :(۱۰)۶۲.
Simon, M., Ercikan, K., Rousseau, M., (۲۰۱۳). Improving large-scale assessment in education :theory, issues and practice . New York & London: Routledge Flamer.
Urry, V. W. (۱۹۷۷) Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, ۱۹۶-۱۸۱ :(۲)۱۴
Vale, C. D. and Gialluca K. (۱۹۸۵) ASCAL: A Microcomputer Program for Estimating Logistic IRT Item Parameters. Computer Science

Development and Standardization of Biology Achievement Test

Hassan Moshtaghian Abarghoie[1]*, Bahram Jowkar[2]

Abstract

The general purpose of this study was to develop and standardize an Achievement test to measure student learning in the biology program at the secondary school. Both classical and IRT models were used to address the research objectives of the study. The preliminary instrument consisted of 150 multiple-choice items that on a sample size of 300 male and female students were performed. The final instrument was two parallel forms of 50 items that were performed on a normative sample of 938 male and female students in Shiraz. The Estimated reliability coefficient for internal consistency with test forms was 0.89, and 0.88, respectively. Based on factor analysis; both forms of the test were an overall factor saturated. Results showed there is no significant difference between the mean scores of boys and girls. So standardized and percentile Norms for all subjects were calculated. Findings from the IRT analysis showed that more than 92% of the items are significantly fitted to Three-Parameter Logistic Model. Test information function was a bell-shaped curve and over a wide ability range from -0.5 to +2.5 provides more information. Also, the maximum information was provided at +1.5 from the ability continuum. Based on this, it can be concluded that the Biology Achievement Test provides a more accurate estimate of the True score of subjects whose ability level is higher than the average ability continuum.

Keyword: Achievement test, Classical model, Item-Response Theory, Standardization

1 -* PhD in Measurement and Assessment; teacher of Baqer Al-Uloom High School, Education District 4 Shiraz -, IRAN kavir311@gmail.com.
2 - Professor in Educational Psychology; Faculty member of Psychology and Educational Sciences Department, Shiraz University, IRAN