پایایی امتحانات نهایی سال سوم متوسطه با استفاده از نظریهٔ تعمیم‌پذیری

■ نورعلی فرخی* ■ لیلا بهرامی**

چکیده:

هدف پژوهش حاضر، نشـان دادن کاربرد نظریهٔ تعمیم‌پذیری در برآورد پایایی داده‌های امتحانات نهایی است. بدین منظور، امتحانات نهایی ادبیات فارسی و زیست‌شناسی رشتهٔ علوم تجربی سال سـوم متوسـطه که در خردادماه ۱۳۹۰ برگزار شـده بود انتخاب گردید. در حوزهٔ امتحانات ناحیهٔ (۱) شهرسـتان خرم‌آباد، از اوراق امتحانی درس ادبیات فارسـی با ۶۱ سـؤال، ۵۴۰ ورقه و از درس زیست‌شناسـی با ۷۸ سـؤال، ۴۲۰ ورقه بـا ملاحظات علمی و عملی به‌طور تصادفی انتخاب شـد. تعداد مصححان دروس مذکور به‌ترتیب، ۹ و ۷ مصحح بود. در درس ادبیات فارسی، رویهٔ سؤالات با ۴۳/۸٪ و در درس زیست‌شناسی، اثر تعاملی سؤالات با دانش‌آموزان با ۶۴/۴٪ بیشترین سهم را در واریانس نمرهٔ کل داشتند. نتایج نشان داد که هم نمرات دانش‌آموزان و هم سؤالات از پایایی بالایی برخـوردار بوده‌اند. همچنین، برای هر کـدام از طرح‌های اندازه‌گیری، مطالعهٔ D صورت گرفت. این پژوهش نشان داد که با استفاده از اصل تقارن نظریهٔ تعمیم‌پذیری، پژوهشگران آموزشی می‌توانند هر کدام از اجزای یک سیسـتم آموزشی را به‌عنوان هدف اندازه‌گیری خود انتخاب کنند و پایایی آن را برآورد نمایند.

کلید واژه‌ها: **پایایی، امتحانات نهایی، نظریهٔ تعمیم‌پذیری**

▣ تاریخ دریافت مقاله: ۹۴/۹/۱ ▣ تاریخ شروع بررسی: ۹۴/۹/۲۲ ▣ تاریخ پذیرش مقاله: ۹۵/۲/۲۸

* دانشیار گروه سنجش و اندازه گیری دانشگاه علامه طباطبایی..... farrokhinoorali@yahoo.com
** کارشناس ارشد سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی..... bahrami.leala@yahoo.com

مقدمه

ارزشـیابی پیشـرفت تحصیلی فرآیند منظمی اسـت که با به‌کارگیری روش‌هـای علمی، عملکرد یادگیرندگان و میزان آموخته‌هایشـان را می‌سنجد و در مورد عملکرد آن‌ها با توجه به اهداف آموزشی موردنظر به قضاوت و داوری می‌پردازد. امتحانات نهایی سال سوم متوسطه نوعی از ارزشیابی پیشرفت تحصیلی است که در پایان دورهٔ آموزشی اجرا می‌شود. از نتایج این امتحانات جهت سنجش آموخته‌های دانش‌آموزان، تصمیم‌گیری در مورد رد یا قبول شـدن آن‌ها و نیز ارزشـیابی اثربخشی برنامهٔ آموزشی و شـیوهٔ تدریس معلم یا مدرس استفاده می‌شود. همچنین، با توجه به قانون «سنجش و پذیرش دانشجو در دانشـگاه‌ها و مراکز آموزش عالی کشور» مصوب ۱۳۹۲/٦/۱۰ مجلس شورای اسلامی، سهم سوابق تحصیلی در کنکور سراسـری (سـهم ۸۵ درصدی) تعیین‌کنندهٔ اصلی در سرنوشت داوطلبان است که هرسـاله به‌طور تدریجی بالا می‌رود (عمادی، ٤ آذر ۱۳۹۲). با در نظر گرفتن نقش امتحانات نهایی در سوابق تحصیلی، می‌توان گفت که این امتحانات پلی برای ورود به دانشگاه محسوب می‌شود.

کارکرد اخیر، بیش‌ازپیش بر حساسـیت این امتحانات افزوده است. به‌طوری‌که از امتحانات نهایی دورهٔ متوسـطه می‌توان به‌عنوان یکی از مهم‌ترین ابزارهای سـنجش در حوزهٔ آموزش‌وپرورش نام برد. زمانی این امتحانات، به‌عنوان ابزار سنجش، کارا و سودمند تلقی می‌شوند و می‌توان به نتایج آن‌ها اعتماد کرد که کیفیت آزمون‌ها در سطح مطلوبی باشد. کیفیت یک ابزار اندازه‌گیری توسط سه حوزهٔ پایایی[۱]، روایی و عملی بودن توصیف شده است (شولتز، تروی و پولمن[۲]، ۲۰۱۱).

اگرچه پایایی می‌تواند به‌طورکلی در چارچوب همسانی و یا تعمیم‌پذیری تعریف شود، شاخص‌های آماری ویژهٔ پایایی، مبتنی بر الگوی آماری و منابع خطا تغییر می‌کند. الگوی آماری ممکن است بر پایهٔ نظریه کلاسـیک آزمون (CTT)[۳]، نظریهٔ تعمیم‌پذیری (GT)[٤]، یا نظریهٔ پرسـش- پاسخ[٥] باشد (میلر[٦]، ۲۰۱۰). ازاین‌رو، در انجام هر مطالعه‌ای، و در هنگام ارائهٔ نتایج، باید پایایی به‌دست‌آمدهٔ اندازه‌ها و روش برآورد پایایی گزارش شود.

نگاهی به مجلات علمی کشور نشان می‌دهد که در بسیاری از مطالعات در زمینهٔ آموزش‌وپرورش و در رشـته‌های گوناگون، با وجود منابع چندگانهٔ خطا[۷] باز پژوهشـگران ضرایب پایایی کلاسـیک را گزارش می‌دهند. این در حالی اسـت که در وضعیت‌هـای اندازه‌گیری پیچیده که منابع چندگانه‌ای از خطای اندازه‌گیری وجود دارد، CTT نمی‌تواند به بررسی منابع چندگانه بپردازد. زیرا، شیوه‌های سنتی پایایی تنها برای یک رویه[۸] طراحی شده‌اند (فن و سان[۹]، ۲۰۱۳). سوئن و لی[۱۰] (۲۰۰۷) نیز اذعان دارند که این‌گونه نیست که CTT وجود منابع چندگانه خطاهای اندازه‌گیری را انکار کند، بلکه حقیقت این است که این نظریه نمی‌تواند از لحاظ مفهومی و آماری آن را در خود جای دهد. درحالی‌که GT نه‌تنها می‌تواند از نظر مفهومی تصور داشتن انواع مختلفی از ضریب پایایی را در خود لحاظ کند، بلکه می‌تواند یک مکانیسم عملی برای انجام آن نیز داشته باشد.

برنان[11] (۲۰۱۰) معتقد است که GT چارچوب مفهومی گسترده و مجموعهٔ محکمی از روش‌های آماری را برای پرداختن به مسائل اندازه‌گیری متعدد فراهم کرده است، تا حدی که این نظریه را می‌توان، به دلیل به‌کاربستن روش‌های تحلیل واریانس (ANOVA)[12] به‌عنوان بسطی از CTT تلقی کرد. تمایزی که GT میان رویه‌های اندازه‌گیری ثابت و تصادفی[13] قائل می‌شود، و همچنین قابلیت این نظریه در پرداختن به طرح‌های مختلف مطالعهٔ تصمیم[14] (مطالعهٔ D) دو ویژگی مهم آن هستند که منجر به برطرف کردن برخی تناقضات آشکار CTT از پایایی شده است (برنان، ۲۰۱۱). همچنین، GT نسبت به CTT، برآورد مناسبی از پایایی را برای آزمون‌های ملاک‌مرجع[15] فراهم می‌آورد. ازآنجاکه CTT نمی‌تواند خطای اندازه‌گیری منظم را در خود جای دهد، تنها برای سنجش هنجارمرجع[16] مناسب است (سوئن و لی، ۲۰۰۷؛ کومازاوا[17]، ۲۰۰۹).

با توجه به ملاک‌مرجع بودن امتحانات نهایی، علی‌رغم اهمیتی که این امتحانات دارند و تصمیمات سرنوشت‌سازی که بر مبنای نتایج آن‌ها گرفته می‌شود، پژوهش‌های معدودی در زمینهٔ پایایی این امتحانات انجام گرفته است که این معدود پژوهش‌ها نیز مبتنی بر CTT می‌باشد. ازاین‌رو، این پژوهش در پی آن است که ضمن ارائهٔ خلاصه‌ای از GT، پایایی امتحانات نهایی سال سوم متوسطه رشتهٔ تجربی در دو درس ادبیات فارسی و زیست‌شناسی را با استفاده از طرح‌های اندازه‌گیری[18] این نظریه بررسی نماید. این پژوهش با پاسخ‌گویی به سؤالات زیر، هدف فوق را تحقق خواهد بخشید.

- آیا نمرات امتحانات نهایی دانش‌آموزان از پایایی لازم برخوردار است؟
- دقت سؤالات در برآورد توانایی دانش‌آموزان تا چه حد است؟
- با تغییر تعداد سطوح رویه‌ها، چه تغییری در اندازهٔ ضرایب تعمیم‌پذیری به وجود می‌آید؟

روش پژوهش

دروس: درس‌های زیست‌شناسی و ادبیات فارسی رشتهٔ علوم تجربی سال سوم متوسطه، که آزمون آن‌ها در خردادماه ۹۰ به‌طور هم‌زمان در سراسر کشور برگزار شده بود، به دلیل دارا بودن ضرایب بالا از میان دیگر دروس اختصاصی و عمومی این رشته انتخاب شدند.

دانش‌آموزان: تعداد دانش‌آموزان سال سوم متوسطه شاخهٔ نظری در سال تحصیلی ۹۰-۸۹ در ناحیهٔ(۱) شهرستان خرم‌آباد ۳۳۰۶ نفر بود که از بین آن‌ها ۱۴۸۸ نفر، معادل ۴۵ درصد، در رشتهٔ علوم تجربی مشغول به تحصیل بودند. این تعداد دانش‌آموز ۵۹۴ نفر، معادل ۳۹/۹۲ درصد، پسر و ۸۹۴ نفر، معادل ۶۰/۰۸ درصد، دختر بودند. از بسته‌های اوراق امتحانی دروس ادبیات فارسی و زیست‌شناسی موجود در انبار امتحانات که متعلق به تمام دبیرستان‌های ناحیهٔ (۱) شهرستان خرم‌آباد بود، نمونه‌ای با حجم ۶۰۰ ورقهٔ امتحانی در دو درس مذکور (۳۰۰ ورقه برای هر جنسیت) به‌طور تصادفی انتخاب شد.

در نهایت با ملاحظات علمی و عملی، نمونهٔ دانش‌آموزان، محدود به ٥٤٠ نفر برای درس ادبیات فارسی و ٤٢٠ نفر برای درس زیست‌شناسی گردید.

سؤالات: آزمون‌های ادبیات فارسی و زیست‌شناسی خردادماه ۹۰، به‌ترتیب حاوی ۵۳ و ۳۰ سؤال بود که با احتساب ریز سؤالات، این تعداد به ٦١ و ٧٨ سؤال می‌رسید. اطلاعات کلیه سؤالات جمع‌آوری گردید.

مصححان: در کل، درس ادبیات فارسی توسط ۱۳ و درس زیست‌شناسی توسط ۷ مصحح، تصحیح شده بود. در درس ادبیات فارسی، اوراق تصحیح‌شدهٔ ٤ مصحح، به دلیل کمی تعداد، کنار گذاشته شد. ازآنجاکه تعداد اوراق تصحیح‌شده توسط هر یک از مصححان برابر نبود و مصححان اوراق امتحانی متفاوتی را تصحیح کرده بودند، امکان استفاده از طرح‌های کاملاً متقاطع، برای داده‌های موجود امکان‌پذیر نبود. ازاین‌رو، در این پژوهش از طرح‌های نسبتاً آشیانه‌ای استفاده گردید. همچنین، تعداد سطوح رویهٔ آشیانه‌شده برای هر سطح از رویه‌ای که در آن آشیانه کرده است، برابر در نظر گرفته شد. لذا تعداد اوراق مصححی که کمتر از اوراق سایر مصححان بود، به‌عنوان ملاک لحاظ شد. به‌طوری‌که از ٦٠٠ ورق امتحانی که به‌طور تصادفی انتخاب شده بود، ٦٠ ورقهٔ امتحانی (۳۰ ورقه برای هر جنسیت) برای هر مصحح به‌طور تصادفی (در صورت زیاد بودن اوراق هر مصحح) انتخاب گردید که در نهایت تعداد اوراق برای درس ادبیات فارسی ٥٤٠ و برای درس زیست‌شناسی ٤٢٠ ورقه شد. در این پژوهش، برای هر دانش‌آموز و در هر درس، نمرهٔ تک‌تک سؤالات و همچنین ریزبارم هر سؤال (در صورت موجودبودن) که توسط مصحح اول داده شده بود به‌عنوان داده‌های پژوهش در فرم‌های مخصوص محقق‌ساخته ثبت شد و سپس در نرم‌افزار SPSS[19] (بستهٔ آماری برای علوم اجتماعی) وارد و در ادامه جهت تحلیل داده‌ها، از نرم‌افزار EDUG (کارگروه انجمن سوئیسی برای پژوهش‌های آموزشی[20]، ۲۰۱۰) استفاده گردید. لازم به ذکر است که هر ورقهٔ امتحان نهایی، دو مرتبه و توسط دو مصحح به‌طور جداگانه تصحیح می‌شود. نکته‌ای که باید به آن اشاره شود، این است که در این پژوهش، نمرات مصححان دوم لحاظ نگردیده است.

تحلیل داده‌ها: طبق اصل تقارن[21] کاردینت، تورنر و الال[22] (١٩٧٦)؛ هر یک از رویه‌های موجود در یک مطالعهٔ اندازه‌گیری می‌توانند به‌عنوان هدف اندازه‌گیری انتخاب شوند و در یک مطالعه، تعریف عملیاتی یک رویه را می‌توان برای رویهٔ دیگر به کار برد. این اصل، منجر به تمایز میان ٤ مرحله از یک مطالعهٔ اندازه‌گیری شده است که عبارت‌اند از: ۱. طرح مشاهده[23]، ۲. طرح برآورد[24]، ۳. طرح اندازه‌گیری، ٤. طرح بهینه‌سازی[25] (شیولسون و وب[26]، ١٩٨١).

در GT دو نوع مطالعه وجود دارد؛ مطالعهٔ تعمیم‌پذیری[27] (مطالعهٔ G) و مطالعهٔ تصمیم (مطالعهٔ D). مطالعهٔ G، سه مرحلهٔ اول را شامل می‌شود و مطالعهٔ D، همان مرحلهٔ چهارم است. در این پژوهش،

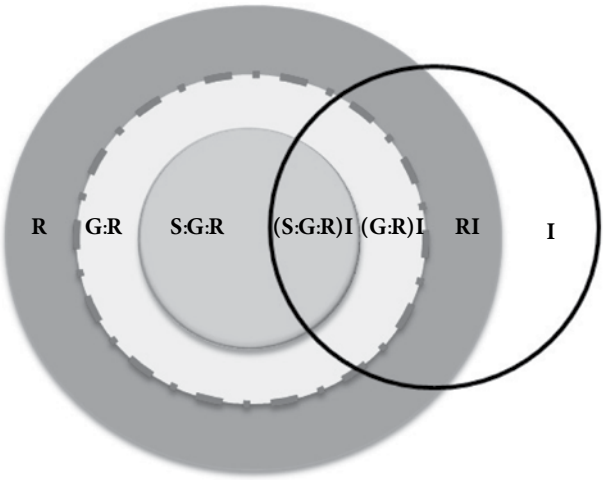
برآورد ضرایب پایایی داده‌ها، مطابق با مراحل مذکور صورت گرفت.

اولین مرحلهٔ یک مطالعهٔ اندازه‌گیری مشخص‌کردن طرح مشاهده است که شامل انتخاب رویه‌ها، مشخص کردن روابط میان رویه‌ها با یکدیگر، تعیین تعداد سطوح[۲۸] و محاسبهٔ میانگین مجذورات است (شیولسون و وب، ۱۹۸۱؛ کاردینت، جانسون و پینی[۲۹]، ۲۰۱۰). یک رویه مجموعه‌ای از سطوح مشابه اندازه‌گیری است (برنان، ۲۰۰۱). برای مثال، هر یک از سؤالات، یک سطح اندازه‌گیری قابل‌قبول برای رویهٔ سؤال را می‌سازد. همچنین، هر مصحح یک سطح اندازه‌گیری قابل‌قبول برای رویهٔ مصحح است. کاردینت و همکاران (۲۰۱۰) معتقدند که یک رویه در GT مترادف با عامل در آنوا است. ازاین‌رو، دو طرح اندازه‌گیری این پژوهش، شامل رویه‌های دانش‌آموزان (S)، سؤالات (I)، مصححان (R) و جنسیت دانش‌آموزان (G) می‌باشد که رویه‌های مذکور در درس ادبیات فارسی به‌ترتیب ۳۰، ۶۱، ۹ و ۲ سطح دارند و در درس زیست‌شناسی به‌ترتیب شامل ۳۰، ۷۸، ۷ و ۲ سطح می‌باشند.

دو نوع متفاوت از روابط میان رویه‌ها وجود دارد. رویه‌ها یا با هم متقاطع‌اند[۳۰] یا در یکدیگر آشیان شده‌اند[۳۱]. اگر هر سطح از هر رویه برای هر سطح از رویهٔ دیگر تکرار شود، دو رویه متقاطع‌اند و اگر مجموعه‌های متفاوتی از سطوح اندازه‌گیری اولین رویه، در ترکیب با همهٔ سطوح اندازه‌گیری دومین رویه رخ دهد، یک رویه درون رویهٔ دوم «آشیانه» کرده است (کاردینت و همکاران، ۲۰۱۰). در مطالعهٔ حاضر، ارتباط رویه‌های اندازه‌گیری بدین قرار است: دانش‌آموزان درون جنسیت و جنسیت درون مصححان آشیانه کرده‌اند که با نماد S:G:R نشان داده شده است. همچنین، رویهٔ سؤال متقاطع با رویه‌های مذکور I × S:G:R می‌باشد. در ادامه، واریانس کل در این پژوهش، به ۷ منبع واریانس تقسیم شده است که منابع تغییرپذیری عبارت‌اند از: مصححان، جنسیت (درون مصححان)، دانش‌آموزان (درون جنسیت)، سؤالات، اثرات تعاملی مصححان- سؤالات، جنسیت- سؤالات، دانش‌آموزان- سؤالات و اثر باقی‌مانده.

مشخص کردن طرح برآورد، دومین مرحلهٔ یک مطالعهٔ اندازه‌گیری است. در این مرحله، وضعیت نمونه‌گیری از رویه‌ها باید مشخص شود. به بیان دیگر، این مرحله شامل تصمیم‌گیری در مورد رویه‌هاست که آن‌ها محدود یا نامحدود[۳۲] و تصادفی یا ثابت در نظر گرفته شوند. همچنین، در این مرحله مؤلفه‌های واریانس برآورد می‌شوند (شیولسون و وب، ۱۹۸۱، کاردینت و همکاران، ۲۰۱۰). GT، اساساً نظریهٔ اثرهای تصادفی[۳۳] است. یک رویهٔ تصادفی از طریق نمونه‌گیری سطوح یک رویه به‌طور تصادفی ایجاد می‌شود. حتی اگر سطوح یک رویه به‌طور تصادفی نمونه‌گیری نشده باشند، رویه ممکن است تصادفی در نظر گرفته شود، اگر سطوح مشاهده نشده در مطالعهٔ G قابل‌جابه‌جایی با سطوح مشاهده شده باشد (وب، شیولسون و هرتل[۳۴]، ۲۰۰۷). همچنین، یک رویه ثابت است اگر سطوح آن، تمام سطوح در مجموعه مرجع باشد که پژوهشگران قصد تعمیم به آن را دارند (شیولسون و وب، ۱۹۹۱). در مطالعهٔ حاضر، رویه‌های مصححان، دانش‌آموزان و سؤالات به‌عنوان رویه‌های تصادفی با مجموعه مرجع نامحدود در نظر گرفته شده‌اند. به‌عبارت‌دیگر، مصححان موردمطالعه یک نمونهٔ تصادفی از تمام مصححان مشابه

ممکن محسوب می‌شوند و نیز دانش‌آموزان و سؤالات. جنسیت دانش‌آموزان یک رویهٔ ثابت است. نمودار طرح برآورد پژوهش حاضر، در قسمت زیر ارائه شده که رویه ثابت جنسیت در آن با نقطه‌چین نشان داده شده است.



نمودار ۱. تقسیم‌بندی واریانس برای طرح برآورد I(S:G:R)

در مرحلهٔ سوم یک مطالعهٔ اندازه‌گیری، باید مشخص شود که کدام رویه‌ها تفکیکی هستند و کدام‌یک ابزاری. کاردینت و همکاران (۲۰۱۰) مطرح می‌کنند که منظور از رویهٔ تفکیکی[۳۵]، رویه‌ای است که هدف و تمرکز اندازه‌گیری قرار گرفته است و واریانس حاصل از این رویه مترادف با مفهوم واریانس نمرهٔ واقعی در CTT است. دیگر رویه‌های موجود در مطالعهٔ اندازه‌گیری، رویه‌های ابزاری[۳۶] محسوب می‌شوند. کاردینت و همکارانش (۱۹۷۶) با مطرح کردن اصل تقارن در GT، بیان کردند که برخلاف تمرکز سنتی روی افراد، هدف اندازه‌گیری[۳۷] (رویهٔ تفکیکی) ممکن است بسته به هدف خاص تصمیم‌گیرنده تغییر کند و تفاوت‌های فردی ممکن است به‌عنوان منبع خطا (رویهٔ ابزاری تصادفی) در نظر گرفته شوند.

در پژوهش حاضر، رویه‌های دانش‌آموزان و سؤالات هر کدام در تحلیل‌های جداگانه به‌عنوان رویه‌های تفکیکی در نظر گرفته شدند. گفتنی است، دو طرح اندازه‌گیری به‌کاررفته در این پژوهش از نوع طرح‌های آمیخته[۳۸] –طرح‌هایی که در آن‌ها ترکیبی از رویه‌های ثابت و تصادفی وجود دارد– نسبتاً آشیانه‌ای –طرح‌هایی که شامل هر دو رویهٔ متقاطع و آشیانه‌ای هستند– و متعادل می‌باشند.

برای ادامهٔ تحلیل در این مرحله، نیاز است که نوع تصمیم (نسبی یا مطلق) مشخص شود و به دنبال آن واریانس‌های خطا و ضرایب تعمیم‌پذیری برآورد شوند. در یک وضعیت اندازه‌گیری، نوع تفسیر نمره (هنجارمرجع در مقابل ملاک‌مرجع) تعیین می‌کند که کدام تصمیم مناسب است و واریانس خطا به‌طور متفاوتی برای هر نوع از تصمیم تعریف می‌شود. یک تصمیم نسبی[۳۹] است، اگر جایگاه سطوح رویهٔ تفکیکی، درون توزیع نمرات و در ارتباط با یکدیگر تعیین شود. واریانس خطا برای این نوع تصمیم را واریانس خطای نسبی[۴۰] می‌نامند که شامل همهٔ مؤلفه‌های واریانس تعاملی است که رویهٔ

تفکیکی را در برمی‌گیرد. همچنین، یک تصمیم مطلق[41] است اگر جایگاه دقیق هر سطح از رویهٔ تفکیکی بر روی مقیاس اندازه‌گیری تعیین شود. واریانس خطا برای این نوع تصمیم را واریانس خطای مطلق[42] می‌نامند که شامل همه مؤلفه‌های واریانس موجود در طرح به جز مؤلفهٔ واریانس مربوط به رویهٔ تفکیکی است. ضریب پایایی و محاسبهٔ آن بستگی به نوع واریانس خطا دارد. برای هر یک از واریانس‌های خطا (نسبی یا مطلق)، ضرایب پایایی جداگانه‌ای برآورد می‌شود. به بیان دیگر، دو نوع ضریب پایایی برآورد می‌شود: ضریب تعمیم‌پذیری[43] نسبی و ضریب تعمیم‌پذیری مطلق (وب و شیولسون، ۲۰۰۵؛ کاردینت و همکاران، ۲۰۱۰؛ فن و سان، ۲۰۱۳). در پژوهش حاضر، هر دو نوع ضریب به همراه خطای استاندارد مربوطه گزارش شده است.

مرحلهٔ چهارم در یک مطالعهٔ اندازه‌گیری، طرح بهینه‌سازی است که مطالعهٔ D را شامل می‌شود. یک مطالعه D از مؤلفه‌های واریانس برآورد شده مطالعهٔ G به‌منظور برآورد اثرات سطوح مختلف رویه‌ها (رویه‌های ابزاری تصادفی یا منابع خطا) بر اتکاپذیری اندازه‌ها استفاده می‌کند تا بدین‌وسیله بهترین روش اندازه‌گیری (کمینه‌سازی منبع خطا – بیشینه‌سازی پایایی) را طراحی کند. همچنین در طراحی چنین مطالعه‌ای، مجموعه مرجع تعمیم[44]، مجموعهٔ مرجعی که پژوهشگر قصد دارد نتایج یک روش اندازه‌گیری خاص را به آن تعمیم دهد، تعریف می‌شود (شیولسون و وب، ۱۹۸۱؛ برنان، ۲۰۰۱). باید در نظر داشت که افزایش یا کاهش سطوح رویه‌ها و یا تغییر ماهیت آن‌ها برای دستیابی به یک طرح اندازه‌گیری مطلوب مستلزم در نظر گرفتن یک رشته ملاحظات منطقی و عملی است. در این پژوهش برای هر طرح اندازه‌گیری، مطالعهٔ D با تغییر سطوح رویه‌ها- نه تغییر ماهیت آن‌ها- و در قالب یک سؤال انجام گرفت.

■ یافته‌های پژوهش

درس ادبیات فارسی

اطلاعات کلی مربوط به طرح برآورد و مشاهدهٔ این مطالعه، در جدول زیر ارائه شده است.

جدول ۱. طرح مشاهده و برآورد مطالعهٔ اندازه‌گیری برای درس ادبیات فارسی

رویه	برچسب	سطوح	مجموعه مرجع
مصححان	R	۹	INF
جنسیت	G:R	۲	۲
دانش‌آموزان	S:G:R	۳۰	INF
سؤالات	I	۶۱	INF

براســاس طرح برآورد و مشــاهده در جدول ۱، ۹ مصحح، هر کدام اوراق امتحانی ۳۰ دانش‌آموز دختر و ۳۰ دانش‌آموز پســر را در درس ادبیات فارسی که دارای ٦١ سؤال بوده است تصحیح کرده‌اند. لازم به ذکر اســت که مصححان به دانش‌آموزان متفاوتی نمره داده‌اند. در این بین رویه‌های مصححان، دانش‌آموزان و سؤالات به‌عنوان رویه‌های تصادفی با مجموعه مرجع نامحدود و جنسیت دانش‌آموزان به‌عنوان رویهٔ ثابت در نظر گرفته شده‌اند. هر کدام از رویه‌های مذکور، به‌عنوان منبع تغییر در نظر گرفته می‌شوند. بنابراین برآورد واریانس آن‌ها در مطالعهٔ اندازه‌گیری اهمیت دارد.

سهم درصدی هر مؤلفهٔ واریانس از واریانس نمرهٔ کل (مجموع مؤلفه‌های تصحیح‌شده واریانس) در جدول ۲ نشان داده شده است. بزرگ‌ترین سهم (٤٣/٨٪) از واریانس نمرهٔ کل به واریانس بین سؤالات (I) تعلق دارد و این بدان معناســت که ســؤالات از نظر سطح دشواری متفاوت هستند. در مرتبهٔ دوم، بزرگ‌ترین سهم (٣٩/٣٪) اثر تعاملی سؤالات با دانش‌آموزان است که در جدول با نماد SI:G:R نشان داده شده است. سهم این مؤلفه با دیگر منابع ناشناختهٔ واریانس و خطای تصادفی آمیخته و بیانگر این مطلب اســت که دشواری نسبی سؤال از دانش‌آموزی به دانش‌آموز دیگر متفاوت است. همچنین رویهٔ دانش‌آموزان، ۸/۳ درصد از واریانس نمره کل را به خود تخصیص داده است. دیگر مؤلفه‌های واریانس برآورد شده، سهم ناچیزی از واریانس نمرهٔ کل دارند.

جدول ۲. جدول منابع تغییر درس ادبیات فارسی برای طرح‌های اندازه‌گیری SGR/I، I/SGR

منبع تغییرات	مجموع مجذورات	درجه آزادی	میانگین مجذورات	مؤلفه‌ها			درصد	خطای استاندارد
				تصادفی	مختلط	تصحیح‌شده		
R	۲۸/۲۴	۸	۳/۵۳	-۰/۰۰۰۲۷	۰/۰۰۰۸۹	۰/۰۰۰۸۹	۲/۱	۰/۰۰۰۶۸
G:R	۴۰/۶۶	۹	۴/۵۲	۰/۰۰۲۳۱	۰/۰۰۲۳۱	۰/۰۰۱۱۶	۲/۷	۰/۰۰۱۰۵
S:G:R	۱۲۱/۷۱	۵۲۲	۰/۲۳	۰/۰۰۳۵۵	۰/۰۰۳۵۵	۰/۰۰۳۵۵	۸/۳	۰/۰۰۰۲۴
I	۶۰۷/۸۰	۶۰	۱۰/۱۳	۰/۰۱۸۶۵	۰/۰۱۸۶۵	۰/۰۱۸۶۵	۴۳/۸	۰/۰۰۳۳۷
RI	۲۸/۳۹	۴۸۰	۰/۰۶	-۰/۰۰۰۱۹	۰/۰۰۰۷۱	۰/۰۰۰۷۱	۱/۷	۰/۰۰۰۱۰
GI:R	۳۸/۰۷	۵۴۰	۰/۰۷	۰/۰۰۱۷۹	۰/۰۰۱۷۹	۰/۰۰۰۹۰	۲/۱	۰/۰۰۰۱۴
SI:G:R	۵۲۴/۰۸	۳۱۳۲۰	۰/۰۲	۰/۰۱۶۷۳	۰/۰۱۶۷۳	۰/۰۱۶۷۳	۳۹/۳	۰/۰۰۰۱۳
Total	۱۳۸۸/۹۵	۳۲۹۳۹					۱۰۰	

- آیا نمرات ادبیات فارسی دانش‌آموزان از پایایی لازم برخوردار است؟

برای پاســخ‌گویی به این ســؤال، دانش‌آموزان (درون جنســیت)، جنســیت (درون مصححان) و مصححان به‌عنوان رویه‌های تفکیکی، و ســؤالات به‌عنوان رویه ابزاری در نظر گرفته شــدند که طرح اندازه‌گیری آن SGR/I است.

بر اساس نتایج جدول ۳، رویهٔ سؤالات با ۵۰/٤٪، بزرگ‌ترین اثر منفی را در دقت اندازه‌گیری مطلق نمرات دانش‌آموزان دارد و اثر تعاملی سؤالات با دانش‌آموزان (درون جنسیت) در هر دو نوع خطای اندازه‌گیری نسبی و مطلق، سهم بزرگی را به خود اختصاص داده است. همچنین، واریانس تفکیکی بیش از ۱۸ برابر واریانس نسبی و بیش از ۹ برابر واریانس مطلق است.

در ادامه، دو شاخص کلی از دقتِ روشِ اندازه‌گیری، در قالب ضرایب تعمیم‌پذیری، برای اندازه‌گیری نسبی (۰/۹۵) و برای اندازه‌گیری مطلق (۰/۹۰) به‌دست آمده است که نشان می‌دهد پایایی اندازه‌گیری نمرات دانش‌آموزان بالا بوده است. اگر هدف مقایسهٔ نمرات ادبیات فارسی دانش‌آموزان با یکدیگر باشد، ضریب تعمیم‌پذیری نسبی نشان می‌دهد که با ضریب پایایی بالا (۰/۹۵)، امکان تفکیک نمرات ادبیات فارسی دانش‌آموزان وجود دارد. حتی اگر هدف مشخص کردن جایگاه یک دانش‌آموز در مقیاس نمره ادبیات فارسی باشد، مقدار ضریب مطلق (۰/۹۰) نشان می‌دهد که با شرایط حاضر به‌صورت پایا، امکان مشخص کردن جایگاه فرد در مقیاس نمره ادبیات فارسی وجود دارد.

جدول۳. مطالعهٔ G برای طرح اندازه‌گیری SGR/I

منبع واریانس	واریانس تفکیکی	منبع واریانس	واریانس خطای نسبی	درصد نسبی	واریانس خطای مطلق	درصد مطلق
R	۰/۰۰۰۸۹		–		–	
G:R	۰/۰۰۱۱۶		–		–	
S:G:R	۰/۰۰۳۵۵		–		–	
	–	I	–		۰/۰۰۰۳۱	۵۰/۴
	–	RI	۰/۰۰۰۰۱	۳/۹	۰/۰۰۰۰۱	۱/۹
	–	GI:R	۰/۰۰۰۰۱	۴/۹	۰/۰۰۰۰۱	۲/۴
	–	SI:G:R	۰/۰۰۰۲۷	۹۱/۳	۰/۰۰۰۲۷	۴۵/۲
مجموع واریانس	۰/۰۰۵۵۹		۰/۰۰۰۳۰	۱۰۰	۰/۰۰۰۶۱	۱۰۰
انحراف استاندارد	۰/۰۷۴۷۹		خطای استاندارد نسبی ۰/۰۱۷۳۴		خطای استاندارد مطلق ۰/۰۲۴۶۲	
ضریب تعمیم‌پذیری نسبی	۰/۹۵					
ضریب تعمیم‌پذیری مطلق	۰/۹۰					

– *دقت سؤالات ادبیات فارسی در برآورد توانایی دانش‌آموزان تا چه حد است؟*

به‌منظور پاسخ به این سؤال، طرح اندازه‌گیری I/SGR آزمون شد که در آن، سؤالات به‌عنوان رویهٔ تفکیکی و رویه‌های دیگر، شامل مصححان، دانش‌آموزان و جنسیت آن‌ها، به‌عنوان رویه‌های ابزاری

در نظر گرفته شـــدند. اگر قصد این باشـــد که بررسی شود، سؤالات درس ادبیات فارسی چقدر خوب می‌توانند نســـبت به یکدیگر در مقیاس دشواری قرار بگیرند، آنگاه تنها دو منبع بالقوه از واریانس خطا وجود دارد، یکی تعاملات سؤالات با مصححان و دیگری تعاملات سؤالات با دانش‌آموزان، یعنی RI و SI:G:R. در این صورت ضریب تعمیم‌پذیری نسبی مدنظر است؛ و اگر هدف این باشد که بدانیم آیا می‌توان اندازه‌گیری دقیقی از دشواری سؤال را برای سؤالات مختلف ادبیات فارسی فراهم کرد یا خیر، در این صورت تفسیر نتایج متمرکز بر ضریب تعمیم‌پذیری مطلق خواهد بود.

در جدول ٤، نتایج مربوط به مطالعهٔ G نشـــان داده شـــده است. در این طرح اندازه‌گیری، ضرایب تعمیم‌پذیری به‌دســـت‌آمده مشابه و خیلی نزدیک به یک است. دلیل مشابه بودن ضرایب تعمیم‌پذیری در این وضعیت اندازه‌گیری این است که، دو منبع واریانس S:G:R (۰/۰۰۰۰۱) و R (۰/۰۰۰۱) که تنها در ضریب تعمیم‌پذیری مطلق ســـهیم هســـتند، تأثیرشان نسبتاً ناچیز است، به‌طوری‌که این دو منبع تنها در حـــدود ۰/٦٪ از واریانـــس کل را به خود اختصاص داده‌اند. رویهٔ G نیز یک رویهٔ ثابت اســـت و در ضریب سهمی ندارد. کاردینت و همکاران (۲۰۱۰) بیان می‌کنند که رویهٔ ثابت که قسمتی از رویه ابزاری را تشـــکیل می‌دهد، تحت تأثیر نوسانات نمونه‌گیری قرار نمی‌گیرد، پس نمی‌تواند بر دقت اندازه‌گیری تأثیری داشـــته باشـــد. ازاین‌رو، واریانس مربوط به جنسیت و اثرات تعاملی آن با دیگر رویه‌ها با مقادیر صفر در داخل پرانتز نشان داده می‌شود.

جدول ۴. مطالعهٔ G برای طرح اندازه‌گیری I/SGR

منبع واریانس	واریانس تفکیکی	منبع واریانس	واریانس خطای نسبی	درصد نسبی	واریانس خطای مطلق	درصد مطلق
	–	R	–		۰/۰۰۰۱۰	۴۶
	–	G:R	–		(۰/۰۰۰۰۰)	۰
	–	S:G:R	–		۰/۰۰۰۰۱	۳/۱
I	۰/۰۱۸۶۵		–		–	
	–	**RI**	۰/۰۰۰۰۸	۷۱/۷	۰/۰۰۰۰۸	۳۶/۵
	–	**GI:R**	(۰/۰۰۰۰۰)	۰	(۰/۰۰۰۰۰)	۰
	–	**SI:G:R**	۰/۰۰۰۰۳	۲۸/۳	۰/۰۰۰۰۳	۱۴/۴
مجموع واریانس	۰/۰۱۸۶۵		۰/۰۰۰۱۱	۱۰۰	۰/۰۰۰۲۱	۱۰۰
انحراف استاندارد	۰/۱۳۶۵۶		**خطای استاندارد نسبی** ۰/۰۱۰۴۷		**خطای استاندارد مطلق** ۰/۰۱۴۶۶	
ضریب تعمیم‌پذیری نسبی	۰/۹۹					
ضریب تعمیم‌پذیری مطلق	۰/۹۹					

– با تغییر تعداد سطوح رویه‌ها، چه تغییری در اندازهٔ ضرایب تعمیم‌پذیری به وجود می‌آید؟

در طرح اندازه‌گیری SGR/I با کاهش سطوح رویهٔ ابزاری (سؤالات) به یک دوم، یک سوم و یک چهارم، اندازهٔ ضریب تعمیم‌پذیری نسبی در مطالعهٔ G (۰/۹۵) به‌ترتیب به ۰/۹۰، ۰/۸٦ و ۰/۸۲ می‌رسد و همچنین اندازهٔ ضریب تعمیم‌پذیری مطلق در مطالعهٔ G (۰/۹۰) به‌ترتیب به ۰/۸۲، ۰/۷۵ و ۰/٦۹ کاهش پیدا می‌کند. برای رسیدن به سطح قابل‌قبولی از هر دو ضرایب و با توجه به اینکه افزایش سؤالات، زمان زیادی را جهت پاسخ‌گویی می‌طلبد که با خستگی و از بین رفتن تمرکز دانش‌آموزان همراه خواهد بود، بهتر است تعداد سؤالات از ٦۱ به ۳۰ برسد؛ البته به شرطی که این کاهش، روایی سؤالات را به مخاطره نیندازد.

در طرح اندازه‌گیری I/SGR، اگر در سطوح رویهٔ دانش‌آموزان تغییری ایجاد نشود و سطوح مصححان به دو نفر کاهش پیدا کند، اندازهٔ ضریب تعمیم‌پذیری نسبی (۰/۹۹) و ضریب تعمیم‌پذیری مطلق (۰/۹۹) به ترتیب ۰/۰۲ و ۰/۰٤ کاهش می‌یابد که همچنان در سطح بسیار مطلوبی قرار می‌گیرد. در صورتی‌که تعداد سطوح رویه‌های ابزاری طرح مذکور به یک سوم فعلی کاهش یابد، می‌توان همچنان به ضرایب تعمیم‌پذیری بالای ۰/۹۵ دست پیدا کرد.

درس زیست‌شناسی

– آیا نمرات زیست‌شناسی دانش‌آموزان از پایایی لازم برخوردار است؟

اطلاعات کلی مربوط به طرح مشاهده و برآورد این مطالعه، در جدول ۵ ارائه شده است.

جدول ۵. طرح مشاهده و برآورد مطالعهٔ اندازه‌گیری برای درس زیست‌شناسی

رویه	برچسب	سطوح	مجموعه مرجع
مصححان	R	۷	INF
جنسیت	G:R	۲	۲
دانش آموزان	S:G:R	۳۰	INF
سؤالات	I	۷۸	INF

هر کدام از ۷ مصحح، به ۳۰ دانش‌آموز دختر و ۳۰ دانش‌آموز پسر برای درس زیست‌شناسی، که دارای ۷۸ سؤال بوده است، نمره داده‌اند. در این بین، رویه‌های مصححان، دانش‌آموزان و سؤالات به‌عنوان رویه‌های تصادفی با مجموعهٔ مرجع نامحدود و جنسیت دانش‌آموزان به‌عنوان رویهٔ ثابت در نظر گرفته شد.

جدول ۶. منابع تغییر درس زیست‌شناسی برای طرح‌های اندازه‌گیری SGR/I، I/SGR

منبع تغییرات	مجموع مجذورات	درجه آزادی	میانگین مجذورات	مؤلفه‌ها			درصد	خطای استاندارد
				تصادفی	مختلط	تصحیح‌شده		
R	۴/۳۷	٦	۰/۷۳	-۰/۰۰۰۱۳	۰/۰۰۰۰۹	۰/۰۰۰۰۹	۰/۶	۰/۰۰۰۱۶
G:R	۹/۴۸	۷	۱/۳۵	۰/۰۰۰۴۶	۰/۰۰۰۴۶	۰/۰۰۰۲۳	۱/۴	۰/۰۰۰۲۷
S:G:R	۱۱۵/۰۷	٤۰٦	۰/۲۸	۰/۰۰۳۵۰	۰/۰۰۳۵۰	۰/۰۰۳۵۰	۲۱/۹	۰/۰۰۰۲۵
I	۵۷/۶۷	۷۷	۰/۷۵	۰/۰۰۱۷۵	۰/۰۰۱۷۵	۰/۰۰۱۷۵	۱۰/۹	۰/۰۰۰۲۸
RI	۵/۹۷	٤٦۲	۰/۰۱	-۰/۰۰۰۰۳	۰/۰۰۰۰۴	۰/۰۰۰۰۴	۰/۳	۰/۰۰۰۰۲
GI:R	۸/۰۳	۵۳۹	۰/۰۱	۰/۰۰۰۱۵	۰/۰۰۰۱۵	۰/۰۰۰۰۸	۰/۵	۰/۰۰۰۰۳
SI:G:R	۳۲۲/۲۷	۳۱۲٦۲	۰/۰۱	۰/۰۱۰۳۱	۰/۰۱۰۳۱	۰/۰۱۰۳۱	۶۴/۴	۰/۰۰۰۰۸
Total	۵۲۲/۸۷	۳۲۷۵۹	-	-	-	-	۱۰۰	-

همان‌طور که نتایج تحلیل واریانس در جدول ٦ نشـان می‌دهد، بزرگ‌ترین سـهم از واریانس نمره کل ٦٤/٤ درصد است در نتیجهٔ تعامل سؤالات با دانش‌آموزان (SI:G:R) است. این مؤلفه با دیگر منابع واریانس ناشـناخته و خطای تصادفی آمیخته است. در مرتبهٔ دوم، واریانس بین دانش‌آموزان S:G:R با ۲۱/۹ درصد و در مرتبهٔ سـوم، واریانس بین سـؤالات I با ۱۰/۹ درصد قرار دارند. مؤلفه‌های واریانس مرتبط با مصححان، جنسیت دانش‌آموزان و تعاملات این دو با سؤالات، کوچک است و سهم اندکی در واریانـس نمره کل دارند به‌طوری‌که در مجموع کمتر از ۳ درصد از واریانس کل را به خود اختصاص داده‌اند.

- آیا نمرات زیست‌شناسی دانش‌آموزان از پایایی لازم برخوردار است؟

برای پاسخ‌گویی به این سـؤال، رویهٔ دانش‌آموزان (درون جنسیت)، جنسیت (درون مصححان) و مصححان به‌عنوان رویه‌های تفکیکی محسوب می‌شوند. براساس نتایج جدول ۷، اثر تعاملی سؤالات با دانش‌آموزان، بزرگ‌ترین اثر منفی را در دقت اندازه‌گیری مطلق و نسـبی نمرات دانش‌آموزان دارد. این مؤلفه ۹۸/۸٪ از واریانس خطای نسـبی و ۸٤/٦٪ از واریانـس خطای مطلق را به خود اختصاص داده اسـت. همچنین، سهم رویهٔ سؤالات از واریانس خطای مطلق ۱٤/٤٪ است. واریانس تفکیکی بیش از ۲۹ برابر واریانس نسبی و بیش از ۲۳ برابر واریانس مطلق است.

جدول ۷. مطالعهٔ G برای طرح اندازه‌گیری SGR/I

منبع واریانس	واریانس تفکیکی	منبع واریانس	واریانس خطای نسبی	درصد نسبی	واریانس خطای مطلق	درصد مطلق
R	۰/۰۰۰۰۹		–		–	
G:R	۰/۰۰۰۲۳		–		–	
S:G:R	۰/۰۰۳۵۰		–		–	
	–	I	–		۰/۰۰۰۰۲	۱۴/۴
	–	RI	(۰/۰۰۰۰۰)	۰/۴	(۰/۰۰۰۰۰)	۰/۴
	–	GI:R	(۰/۰۰۰۰۰)	۰/۷	(۰/۰۰۰۰۰)	۰/۶
	–	SI:G:R	۰/۰۰۰۱۳	۹۸/۸	۰/۰۰۰۱۳	۸۴/۶
مجموع واریانس	۰/۰۰۳۸۲		۰/۰۰۰۱۳	۱۰۰	۰/۰۰۰۱۶	۱۰۰
انحراف استاندارد	۰/۰۶۱۸۴		خطای استاندارد نسبی ۰/۰۱۱۵۶		خطای استاندارد مطلق ۰/۰۱۲۵۰	
ضریب تعمیم‌پذیری نسبی	۰/۹۷					
ضریب تعمیم‌پذیری مطلق	۰/۹۶					

در ادامه، نتایج مطالعهٔ G نشان می‌دهد که اندازهٔ هر دو نوع ضریب تعمیم‌پذیری برای اندازه‌گیری نسبی و مطلق، بیش از ۰/۹۵ است؛ پس که می‌توان نتیجه گرفت که اندازه‌گیری نمرات دانش‌آموزان از پایایی بالایی برخوردار است. اگر هدف مقایسهٔ نمرات دانش‌آموزان در درس زیست‌شناسی با یکدیگر باشد، ضریب تعمیم‌پذیری نسبی نشان می‌دهد که با ضریب پایایی (۰/۹۷) بالا، امکان تفکیک نمرات زیست‌شناسی دانش‌آموزان وجود دارد. حتی اگر هدف مشخص کردن جایگاه یک دانش‌آموز در مقیاس نمره زیست‌شناسی باشد، مقدار ضریب مطلق (۰/۹۶) نشان می‌دهد که این امر به‌صورت پایا امکان‌پذیر است.

– *دقت سؤالات زیست‌شناسی در برآورد توانایی دانش‌آموزان تا چه حد است؟*

برای پاسخ‌گویی به این سؤال، می‌توان طرح اندازه‌گیری I/SGR را آزمون کرد که رویهٔ سؤالات به‌عنوان رویهٔ تفکیکی در نظر گرفته شده است. جدول ۸، نتایج مطالعهٔ G مربوط به این طرح را نشان می‌دهد.

جدول ۸. مطالعهٔ G برای طرح اندازه‌گیری I/SGR

منبع واریانس	واریانس تفکیکی	منبع واریانس	واریانس خطای نسبی	درصد نسبی	واریانس خطای مطلق	درصد مطلق
	–	R	_		۰/۰۰۰۰۱	۲۵/۷
	–	G:R	_		(۰/۰۰۰۰۰)	۰/۰
	–	S:G:R	_		۰/۰۰۰۰۱	۱۵/۸
I	۰/۰۰۱۷۵		_		–	
	–	RI	۰/۰۰۰۰۱	۲۰/۲	۰/۰۰۰۰۱	۱۱/۸
	–	GI:R	(۰/۰۰۰۰۰)	۰	(۰/۰۰۰۰۰)	۰
	–	SI:G:R	۰/۰۰۰۰۲	۷۹/۸	۰/۰۰۰۰۲	۴۶/۶
مجموع واریانس	۰/۰۰۱۷۵		۰/۰۰۰۰۳	۱۰۰	۰/۰۰۰۰۵	۱۰۰
انحراف استاندارد	۰/۴۱۸۶		خطای استاندارد نسبی ۰/۰۰۵۵۵		خطای استاندارد مطلق ۰/۰۰۷۲۵	
ضریب تعمیم‌پذیری نسبی	۰/۹۸					
ضریب تعمیم‌پذیری مطلق	۰/۹۷					

در این روش اندازه‌گیری، منابع تشکیل‌دهندهٔ واریانس خطای مطلق عبارت‌اند از: اثر تعاملی سؤالات با دانش‌آموزان (۴۶/۶٪)، مصححان (۲۵/۷٪)، دانش‌آموزان (۱۵/۸٪) و اثر تعاملی سؤال با مصححان (۱۱/۸٪). همچنین اثر تعاملی سؤالات با دانش‌آموزان با ۷۹/۸٪، بزرگ‌ترین اثر منفی را در دقت اندازه‌گیری نسبی دشواری سؤالات دارد. اثر تعاملی سؤالات و مصححان نیز با ۲۰/۲٪ سهم قابل‌توجهی در واریانس خطای نسبی دارد. باوجوداین، هر دو واریانس خطای نسبی و مطلق در مجموع، کمتر از ۵ درصد از واریانس کل را به خود اختصاص داده‌اند و ضرایب تعمیم‌پذیری به‌دست‌آمده برای اندازه‌گیری نسبی و مطلق سؤال اختلاف ناچیزی با هم داشته و خیلی نزدیک به یک است. بنابراین می‌توان نتیجه گرفت که قرار دادن سؤالات زیست‌شناسی به‌طور پایا در مراتب دشواری (از آسان‌ترین به مشکل‌ترین- اندازه‌گیری نسبی) امکان‌پذیر است. به بیان دیگر، سؤالات قدرت نشان دادن موفقیت دانش‌آموزان را با درجات مختلف دارند. همچنین، به‌طور رضایت‌بخش و با دقت بالا می‌توان دشواری سؤال را برای سؤالات مختلف زیست‌شناسی فراهم کرد.

- با تغییر تعداد سطوح رویه‌ها، چه تغییری در اندازهٔ ضرایب تعمیم‌پذیری به وجود می‌آید؟

پاسخ‌گویی به این سؤال، مستلزم انجام مطالعهٔ D (طرح بهینه‌سازی) است که برای هر طرح اندازه‌گیری و به دنبال مطالعهٔ G انجام گرفته است.

در طرح اندازه‌گیری SGR/I با کاهش سطوح رویهٔ ابزاری (سؤالات) به یک دوم، یک سوم و یک چهارم، اندازهٔ ضریب تعمیم‌پذیری نسبی در مطالعهٔ G (۰/۹۷) به‌ترتیب به ۰/۹۳، ۰/۹۱ و ۰/۸۸ می‌رسد و همچنین اندازهٔ ضریب تعمیم‌پذیری مطلق در مطالعهٔ G (۰/۹۶) به‌ترتیب به ۰/۹۲، ۰/۸۹ و ۰/۸۶ کاهش پیدا می‌کند. با وجود کاهش تعداد سؤالات به ۱۵ سؤال، باز ضرایب تعمیم‌پذیری بالاتر از حداقل حد مطلوب پذیرفته شده (۰/۸۰) قرار دارند. بهتر است سؤالات را به تعدادی کاهش دهیم که هم‌زمان کل محتوای کتاب درسی نیز پوشش داده شود.

در طرح اندازه‌گیری I/SGR، اگر تعداد مصححان از ۷ به ۳ و هم‌زمان تعداد دانش‌آموزان نیز از ۳۰ به ۱۵ برسد، هر دو ضرایب تعمیم‌پذیری، با وجود کاهش در اندازه، همچنان ۰/۹۰ به بالا هستند. علاوه بر این، اگر در سطوح رویهٔ دانش‌آموزان تغییری ایجاد نشود و سطوح مصححان به دو نفر کاهش پیدا کند، اندازهٔ ضریب تعمیم‌پذیری نسبی از ۰/۹۸ به ۰/۹۴ و ضریب تعمیم‌پذیری مطلق از ۰/۹۷ به ۰/۹۰ می‌رسد.

■ بحث و نتیجه‌گیری ■

در وضعیت‌های اندازه‌گیری همچون امتحانات نهایی، منابع متفاوتی از خطا از قبیل سؤالات، مصححان، شرایط امتحان، جنسیت و... وجود دارد که نمرات مشاهده‌شده را متأثر می‌کند. در چنین شرایطی، هنگام برآورد پایایی، CTT قادر به تفکیک منابع چندگانه خطای اندازه‌گیری نیست و همهٔ آن‌ها را به‌عنوان خطای تصادفی در نظر می‌گیرد. درصورتی‌که GT منابع چندگانه خطای منظم را به‌دقت مشخص می‌کند و ضمن تفکیک آن‌ها، اثر هر یک را بر روی پایایی اندازه‌گیری تعیین می‌کند. با در نظر گرفتن منابع چندگانه خطا، نتایج به‌دست‌آمده را با دقت بیشتری می‌توان به سایر موقعیت‌های اندازه‌گیری تعمیم داد (شیولسون و وب، ۱۹۹۱؛ برنان، ۲۰۰۱). بدین منظور در این پژوهش، برای بررسی پایایی امتحانات نهایی دو درس ادبیات فارسی و زیست‌شناسی از GT استفاده گردید.

با توجه به اینکه اندازهٔ نمونه در هر دو درس موردمطالعه یکسان بوده و با در نظر گرفتن سهم واریانس دانش‌آموزان در هر دو درس، می‌توان گفت: توانایی دانش‌آموزان در درس ادبیات فارسی، در مقایسه با درس زیست‌شناسی، از تجانس بیشتری برخوردار بوده است. برای آزمون‌های نرم‌مرجع، اثر شخص باید بزرگ باشد درحالی‌که اثر سؤال باید حدود یک سوم کمتر از اثر شخص باشد. درحالی‌که در آزمون‌های ملاک‌مرجع، به دلیل اینکه دانش‌آموزان بر حسب سطح تسلطشان همگن هستند، واریانس شخص ممکن است پایین باشد. ازاین‌رو در آزمون‌های ملاک‌مرجع که سؤالات ملاک یا هدف خاصی را اندازه می‌گیرند، به‌دست‌آمدن مقدار بزرگی برای واریانس سؤال مطلوب است (براون و راس[۴۵]، ۱۹۹۶، به نقل از کومازاوا، ۲۰۰۹؛ کومازاوا،

۲۰۰۹). با توجه به ملاک‌مرجع بودن امتحانات نهایی، در درس ادبیات فارسی، سؤالات بیشترین واریانس را به خود اختصاص داده‌اند. درصورتی‌که در درس زیست‌شناسی، سؤالات سومین منبع تشکیل‌دهندهٔ واریانس هستند و اثر سؤال یک دوم اثر شخص است.

استفاده از GT و خاصیت تقارن‌پذیری آن، به ارزشیابان و پژوهشگران آموزشی این امکان را می‌دهد که هر کدام از اجزای یک سیستم آموزشی را می‌توانند به‌عنوان هدف اندازه‌گیری خود انتخاب کنند. به بیان دیگر، در پژوهش‌های آموزشی علاوه بر نمرات دانش‌آموزان، سایر ابعاد آموزشی نیز از قبیل برنامه‌های آموزشی، اهداف آموزشی، محیط آموزشی، سال تحصیلی، روش‌های تدریس، مصححان، معلمان، حجم کتاب، فصول کتاب، سؤالات و غیره می‌تواند به‌عنوان رویهٔ تفکیکی (هدف اندازه‌گیری) انتخاب شود. در پژوهش حاضر با به‌کارگیری اصل تقارن، رویه‌های دانش‌آموزان و سؤالات در قالب دو طرح جداگانه و برای هر کدام از دروس موردمطالعه به‌عنوان هدف اندازه‌گیری در نظر گرفته شدند.

نتایج مطالعهٔ G مربوط به طرح‌های اندازه‌گیری SGR/I و I/SGR نشان داد که هم نمرات دانش‌آموزان و هم سؤالات از پایایی بالایی برخوردارند. به‌طوری‌که دامنهٔ ضرایب تعمیم‌پذیری برای هر دو نوع اندازه‌گیری نسبی و مطلق، (۰/۹۰ تا ۰/۹۹) است. همان‌طور که وب و همکاران (۲۰۰۷) مطرح کرده‌اند، برای گرفتن تصمیم‌هایی در مورد افراد مبتنی بر نمرات مشاهده‌شده‌شان، ضریب پایایی ۰/۸۰ و بالاتر غالباً به‌قدر کافی پایا تلقی می‌شود و درصورتی‌که تصمیمات پیامدهای چشمگیری داشته باشند، مقادیر ۰/۹۰ به بالاتر ترجیح داده می‌شود.

ازآنجاکه طرح سؤال و روند تصحیح مستلزم صرف هزینه و زمان است با انجام مطالعات D و با در نظر گرفتن محدودیت‌های منطقی و عملی، می‌توان ترکیب مناسبی از سؤالات و مصححان را، با توجه به‌اندازه پایایی دلخواه، به‌دست آورد و طرح اندازه‌گیری مطلوبی برای برآورد پایایی این امتحانات طراحی کرد. نتایج امتحانات نهایی به دلیل تشریحی بودن، وابسته به‌دقت تصحیح مصححان است. ازاین‌رو ضروری است به‌منظور ارتقای کیفیت تصحیح، در پژوهش‌های جداگانه‌ای با به‌کارگیری دیگر طرح‌های متنوع، به‌خصوص طرح‌های متقاطع این مسئله بررسی شود. همچنین، می‌توان سایر ویژگی‌های مصححان از قبیل سابقهٔ تصحیح، سن، جنسیت و دیگر عوامل را در طرح‌های اندازه‌گیری مناسبی وارد نمود و سهم واریانس آن‌ها را برآورد کرد.

در این پژوهش، از طرح‌های نسبتاً آشیانه‌ای استفاده گردید که از آن می‌توان به‌عنوان محدودیت این پژوهش نام برد. همان‌طور که شیولسون، وب و رولی[۴۶] (۱۹۸۹) بیان کرده‌اند؛ در مطالعات G باید تا جایی که امکان‌پذیر است از طرح‌های متقاطع استفاده کرد. زیرا با این طرح‌ها، امکان برآورد جداگانهٔ همهٔ مؤلفه‌های واریانس وجود دارد. درحالی‌که در دیگر طرح‌ها، اثر مستقیم مؤلفهٔ واریانس مربوط به رویهٔ آشیانه‌ای به‌صورت جداگانه برآورد نمی‌شود.

منابع

- عمادی، عبدالرسول. (۱۳۹۲). *کنکور به‌طور کامل حذف نشده بلکه سهم سوابق تحصیلی در کنکور بسیار بالا می‌رود* [اخبار]. بازیابی شده در ٤ آذر ۱۳۹۲، برگرفته از http://www.aee.medu.ir/IranEduThms/theme2/cntntpge.php?pgid=22&rcid=130
- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.
- Brennan, R. L. (2010). Generalizability theory. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 61-68). New York, NY: Springer- Verlag.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*(2), 119-135.
- Fan, X., & Sun, S. (2013). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *The Journal of Early Adolescence, 34*(1), 38-65.
- Kumazawa, T. (2009). Revision of a Criterion-Referenced Vocabulary Test Using Generalizability Theory. *JALT Journal, 31*(1), 81-100.
- Miller, M. D. (2010). Classical Test Theory Reliability. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 27 – 30). New York, NY: Springer -Verlag.
- Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury, 42*(3), 236-240.
- Suen, H. K., & Lei, P. W. (2007). Classical versus Generalizability theory of measurement. *Educational measurement, 4*, 1-13.
- Shavelson, R. J., & Webb, N .M. (1981). Generalizability theory: 1973 – 1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133- 166.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932.
- Shavelson, R. J., & Webb, N .M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Swiss Society for Research in Education Working Group (2010). *EDUG user guide*. Neuchatel, Switzerland: IRDP.
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. *Handbook of statistics, 26*, 81-124.
- Webb, N.M., & Shavelson, R. J. (2005). Generalizability theory: Overview. *Encyclopedia of statistics in behavioral science, 2*, 717-719.

پی‌نوشت‌ها

1. reliability
2. Scholtes, Terwee & Poolman
3. classical Test Theory (CTT)
4. generalizability Theory (GT)
5. item response Theory
6. Miller
7. multiple sources of error
8. facet
9. Fan & Sun
10. Suen & Lei
11. Brennan
12. analysis of variance (ANOVA)
13. fixed & random
14. decision study
15. criterion- referenced
16. norm- referenced
17. Kumazawa
18. measurement designs
19. statistical package for social science
20. Swiss Society for Research in Education Working Group
21. symmetry
22. Cardinet, Tourneur & Allal
23. observation design
24. estimation design
25. optimization design
26. Shavelson & Webb
27. generalizability study
28. conditions
29. Johnson & Pini
30. crossed
31. nested
32. restricted or infinite
33. random effects
34. Haertel
35. differentiation facet
36. instrumentation facets
37. object of measurement
38. mixed
39. relative decision
40. relative error variance
41. absolute decision
42. absolute error variance
43. generalizability coefficient
44. universe of generalization
45. Brown & Ross
46. Rowley