Quarterly Journal of
Research in School and Virtual Learning
Year 3- Number 10 (P 47-56)

فصلنامه علمی – پژوهشی
پژوهش در یادگیری آموزشگاهی و مجازی
سال سوم – شماره دهم (ص 47 - 56)

برآورد پارامترهای سؤال‌های چندگزینه‌ای در ارزشیابی نظام‌های آموزشی مجازی

محمود اکرامی [*1]؛ نگار الهامیان[2]؛ سمیه رجب‌زاده[3]
1. دانشیار، مدیریت آموزشی، دانشگاه پیام نور
2. دانشجوی دکتری، مطالعات برنامه درسی، دانشگاه خوارزمی
3. دانشجوی دکتری، برنامه‌ریزی از دور، دانشگاه پیام نور
تاریخ دریافت: 1394/04/09 **تاریخ پذیرش**: 1394/09/18

Estimating the Parameters of Multiple - Choice Questions in the Evaluation of the Educational Systems Virtual

M. Ekrami[*1]; N. Elhamian[2]; S. Rajabzadeh[3]
1. Associate Professor, Educational Administration, Payame Noor University
2. Ph.D Student, Curriculum Studies, Kharazmi University
3. Ph.D Student, Distance Planning, Payame Noor University
Received: 2015/06/30 **Accepted:** 2015/12/09

چکیده

مطالعهٔ حاضر با عنوان «برآورد پارامترهای سؤال‌های چندگزینه‌ای در ارزشیابی نظام‌های آموزشی مجازی» روی آزمون‌های ریزشی برای ممانعت از تکرار آزمون انجام پذیرفت. برای برآورد پارامترهای سؤال‌ها، پاسخ‌نامه‌های 49 آزمون نهایی دانشجویان ریزشی یک نیمسال تحصیلی بررسی گردید. در مدل کلاسیک، درجه دشواری هر سؤال (p) نسبت پاسخ‌های درست به کل آزمودنی‌ها و قدرت تشخیص (R_{pbis}) ضریب همبستگی دو رشته‌ای نقطه‌ای با کل آزمون در نظر گرفته می‌شود، محاسبات نشان داد: 1- در 89/7% آزمون‌ها حتی یک سؤال خیلی آسان و روا مشاهده نگردید. 2- در 63/26% آزمون‌ها حتی یک سؤال خیلی سخت و روان مشاهده نگردید. 3- فقط 2/04% آزمون‌ها بدون سؤال ناروا است. پس از حذف سؤال‌های ناروا، دوباره نمرهٔ دانشجویان تعیین و حدود 15/11% افراد از کمند ریزش جستند. به این ترتیب با حذف ماده‌های ناروا، درصد آزمون‌های ریزشی به شدت کاهش می‌یابد، از اتلاف منابع جلوگیری می‌شود، ملاک پیشرفت تحصیلی دانشجویان واقعی‌تر و مبنای تصمیم‌گیری‌های بهتر را برای دانشگاه فراهم می‌آورد.

واژه‌های کلیدی: مدل کلاسیک، آزمون چندگزینه‌ای، درجهٔ دشواری سؤال، قدرت تشخیص سؤال، مرکز آزمون دانشگاه پیام نور.

Abstract

Present study, titled "Estimating the parameters of multiple-choice questions in the evaluation of the educational systems virtual" was performed on failed test to prevent the repetition of tests. For parameters estimation of the questions, responses of 49 final test of failed students was investigated in an academic semester. In the classic model, the degree of difficulty of each question (p) the proportion of correct responses in all subjects, and judgment (Rpbis) double-stranded correlation with the test point considered, the calculations showed: (1) In-89.7% of tests there wasn't even a very easy and reliable question. (2) In 63.26% of tests wasn't observed even a very difficult and reliable question. (3) Only 2.04% of tests is without unreliable question. After deleting the unreliable questions, the students' marks were determined again and about 15.11% of students passed the test. Thus, by deletion unreliable questions, percent of failed tests reduced remarkably, the waste of resources is prevented, the criteria of students' academic achievement became real more and provided better basis for decision-making in university.

Keywords: Classical Model, Multiple-Choice Test, The Degree of Question Difficulty, The Power of Determining Answer, PNU Test Center.

* **نویسنده مسئول**: محمود اکرامی

*Corresponding Author: m.ekrami@pnu.ac.ir

مقدمه

ارزشیابی پیشرفت تحصیلی دانشجویان یکی از اهداف عمده مراکز آموزش عالی کشور، دانشگاه‌ها و به ویژه دانشگاه پیام نور است زیرا به این وسیله میزان یادگیری دانشجویان و در نهایت میزان دستیابی به اهداف آموزشی اندازه‌گیری می‌شود. یک ارزشیابی مؤثر نه تنها در تمایز دانشجویان نقش به سزایی دارد بلکه به ارزیابی فعالیت‌های مدرس نیز کمک می‌کند زیرا که اگر نتوانیم چیزی را مورد سنجش قرار دهیم قادر به بهبود آن نخواهیم بود. به گفتهٔ کرگ یراقی و همکاران (1389) ارزشیابی یکی از جنبه‌های مهم در فرایندهای آموزشی بوده و این امکان را فراهم می‌سازد تا بر اساس نتایج آن نقاط قوت و ضعف آموزش شناسایی شده و با تقویت جنبه‌های مثبت و رفع نارسایی‌ها در ایجاد تحول و اصلاح نظام آموزشی گام‌های مناسبی برداشته شود.

دلاور و همکاران (1385) بیان می‌کنند که تلاش‌های نخستین برای تکوین نظریهٔ کلاسیک اندازه‌گیری در دههٔ 1890 آغاز شد. این نظریه، از روش‌های دیرینه برای ساخت و توسعهٔ آزمون‌ها در حوزهٔ علوم انسانی است که از اوایل دههٔ 1900 برای توسعهٔ ابزارهای اندازه‌گیری و تعیین میزان هم‌خوانی آزمون‌ها با نظریه و نمره‌گذاری امتحانات استفاده شده است.

هومن (1375) معتقد است فرایند آزمون‌سازی در طی روند تاریخی تکوین و تکامل خود با دو جنبهٔ کاملاً متمایز و مجزای ساخت و پرورش آزمون از یک‌سو و نظریهٔ آماری برای تجزیه و تحلیل آزمون از سوی دیگر مواجه بوده است. سهولت نمره‌گذاری، اجرا و عینیت آزمون‌های چندگزینه‌ای سبب شده که به عنوان ابزار اصلی در سنجش‌های وسیع مورد استفاده قرار می‌گیرد. ثرندایک (نقل از هومن، 1375) عقیده دارد نوشتن سؤال‌های چندگزینه‌ای با وجود همه تلاش‌هایی که در جهت مکانیزه و کامپیوتری کردن آن به عمل آمده است همچنان به عنوان یک هنر تلقی می‌شود.

به اعتقاد سیف (1386) چند دلیل عمده از آزمون‌های چندگزینه‌ای بیش از سایر انواع آزمون‌ها در حوزه تعلیم و تربیت استفاده می‌شود: اول به علت آنکه آزمودنی قادر است در زمان معین تعداد زیادی سؤال را پاسخ دهد و به عبارتی دیگر در یک زمان محدود تعداد زیادی از هدف‌های آموزشی و بخش مهمی از محتوای درس را اندازه بگیرد .دوم اینکه آزمون‌های چندگزینه‌ای نسبت به آزمون‌های صحیح و غلط و دوگزینه‌ای کمتر امکان حدس زدن کورکورانه را به آزمون‌شونده می‌دهند. دلیل سوم برای استفاده بیشتر از این نوع آزمون سهولت در نمره‌گذاری و تصحیح و تفسیر آن و در نتیجه صرفه‌جویی در نیروی انسانی و وقت و هزینه است.

در مقابل شریفی (نقل از کریمی و فلسفی‌نژاد، 1390) این نوع آزمون‌ها معایبی نیز دارد از جمله اینکه ساختن این آزمون‌ها بسیار دشوار است و در مقایسه با آزمون‌های صحیح - غلط خواندن این آزمون‌ها و پیدا کردن گزینه درست مستلزم وقت زیادتری است.

امروزه آزمون‌های چندگزینه‌ای استعداد و پیشرفت تحصیلی در حوزهٔ تعلیم و تربیت بیش از سایر موقعیت‌ها مورد استفاده قرار می‌گیرد اما اینکه یک آزمون چندگزینه‌ای چه تعداد گزینه باید داشته باشد تا از حداکثر پایایی برخوردار باشد، همواره مورد بحث است. (کریمی، فلسفی‌نژاد، درتاج، 1390).

نحوهٔ ارزیابی متداول برای دانشجویان پیام نور طرح سؤالات چندگزینه‌ای است. آزمون‌های چندگزینه‌ای که به آزمون‌های تستی شهرت دارند یکی از انواع آزمون عینی هستند که اخیراً به دلیل محاسبه ویژگی‌های روان‌سنجی آن مورد توجه متخصصان آزمون‌سازی قرار گرفته‌اند. این نوع آزمون‌ها اگر با دقت و مهارت کافی ساخته شوند، برای اندازه‌گیری تمام اهداف درس در تمام سطوح فکری و در تمام رشته‌های تحصیلی مناسب هستند. اکثر آزمون‌های استاندارد که در سطح بین‌المللی که برای گزینش دانشجو در سطوح مختلف تحصیلی انجام می‌پذیرد، از نوع چندگزینه‌ای است ولی بهترین شرایط استفاده از این روش همراه کردن آن با روش‌های دیگر ارزیابی است.

اجرای آزمون‌های نهایی پایان ترم در دانشگاه پیام نور، افزون بر فراهم آوردن امکان ارزشیابی‌های دانشجویان و دوره‌های آموزشی، آیا می‌تواند راهی برای تقلیل هزینه‌های ناشی از نقص ابزارهای سنجش موجود در بانک سؤال و هزینه‌های تکرار سنجش و از همه مهم‌تر پیش‌گیری از هزینه‌های غیرمادی ناشی از ریزش دانشجویان و ارائهٔ تصویری نابجا از افت تحصیلی دانشجویان دست کم در مقطع تحصیلی کارشناسی در همه واحدها و مراکز آموزشی دانشگاه پیام نور در سراسر کشور باشد.

هدف از انجام این مطالعه، ارتقای کیفیت آزمون‌های پیشرفت تحصیلی مورد استفاده برای ارزشیابی دانشجو در آزمون‌های نهایی پایان ترم از طریق بهره‌گیری از نظریه‌های جدید در تجزیه و تحلیل سؤالات مانند نظریهٔ تستی کلاسیک و نیز استفاده از فناوری در افزایش دقت و صحت محاسبات است

که ماحصل آن به کارگیری نرم‌افزار تجزیه و تحلیل سؤالات چندگزینه‌ای مبتنی بر نظریهٔ تستی کلاسیک برای دانشجویان مقطع کارشناسی است. تجزیه و تحلیل سؤالات به منظور کاهش منابع خطا، تعیین سؤالات مناسب، تعیین سؤالاتی که نیاز به ویرایش دارند، تعیین شاخص افتراق بین دانشجویان قوی و ضعیف و نیز تعیین پایایی آزمون انجام می‌گیرد.

پرسش‌های پژوهش

1. درجهٔ دشواری[1] هریک از ماده‌های آزمون‌های نهایی چه میزان است؟
2. روایی[2] هریک از ماده‌های آزمون‌های نهایی چه میزان است؟

ارزشیابی آموزشی به یک فعالیت رسمی گفته می‌شود که برای تعیین کیفیت اثربخشی و یا ارزش یک برنامه، فرآورده، پروژه، فرایند، هدف یا برنامه درسی به اجرا در می‌آید. هدف اصلی ارزشیابی آموزشی، تعیین قدر و ارزش پدیده مورد ارزشیابی است تا اینکه به افراد علاقه‌مند و مسئول کمک کند تا دربارهٔ آن پدیده تصمیم‌های درستی اتخاذ نمایند.

از لحاظ اصطلاحی، ارزشیابی یا ارزیابی به اعتقاد کرم‌دوست (1383) عبارت از آزمون و قضاوت دربارهٔ ارزش، کیفیت، اهمیت، میزان، درجه یا شرایط یک پدیده است.

«آزمون وسیله و ابزاری است برای اندازه‌گیری غیرمستقیم صفات و ویژگی‌های روانی». تریتسچلر (2000) نیز معتقد است که آزمون وسیله یا روشی است که پاسخ‌های قابل مشاهده و درخور توجهی را برای فراهم کردن اطلاعات درباره ویژگی‌های شخص یا اشخاص فرا می‌خواند. آناستازی (1997) معتقد است که آزمون وسیله کم یا بیش اسرارآمیزی نیست، بلکه کوششی است برای اندازه‌گیری نمونه‌ای از رفتار فرد به صورتی عینی و پیوسته (دلاور و زهراکار، 1387: صص 21 و 22).

شریفی (به نقل از مقدمی، 1392) می‌گوید اصطلاح استاندارد شده بدین معناست که آزمون قبلاً در مورد گروه نمونه‌ای از افراد موردنظر در بوتهٔ آزمایش گذاشته شده و نتایج پژوهش‌های مربوط به آن از راه روش‌های آماری مورد تجزیه و تحلیل قرار گرفته و روایی و اعتبار آن تعیین شده است.

به اعتقاد امین و همکاران (1389) سؤالات چند گزینه‌ای معمول‌ترین نوع سؤالات برای سنجش دانش در سطوح مختلف (یادآوری، تفسیر و حل مسئله) می‌باشند. گرچه بیشتر سؤالات چندگزینه‌ای معمولی برای سنجش سطح پایین دانش به کار گرفته می‌شوند ولی اگر این سؤالات خوب طرح شوند، قادرند سطح بالای دانش، فهم، ادراک، کاربرد، اطلاعات و حل مسئله را بسنجند.

دانشگاه پیام نور و مؤسسات آموزش عالی از دور، نوع پیشرفته‌تر و نظام‌یافته‌تری از مؤسسات آموزش مکاتبه‌ای است که در کنار دانشگاه‌های سنتی ایران ایجاد شده است، به گفته آقازاده (1380) هدف این بزرگ‌ترین مرکز آموزش عالی مجازی عالم اسلام به شرح زیر است.

1. توسعهٔ نظام آموزش عالی به‌گونه‌ای مؤثر و با شیوه‌های جدید علمی و عملی تا بتواند گروه‌های بیشتری از علاقه‌مندان به تحصیلات دانشگاه‌ها را زیر پوشش آموزش عالی قرار دهد.
2. فراهم کردن فرصت آموزش عالی برای گروهی از افراد شاغل و زنان خانه‌دار که به دلیلی نمی‌توانند در دانشگاه‌های عادی و سنتی حضور داشته باشند.
3. برقراری فضای علمی و آموزشی در داخل خانه‌ها و در قلمروی حرفه‌ای و شغلی افراد در شهرها و روستاها.
4. برقراری روش‌های جدید و قابل‌انعطاف در سطح آموزش عالی که با اوقات آزاد دانشجویان هماهنگی داشته باشد.

در توسعهٔ کیفی ساختار آموزشی، یکی از اقدامات مهم اشاره شده دانشگاه پیام نور در دورهٔ ارتقای کیفی، سامان‌دهی نظام سنجش و آزمون است. موارد پنج‌گانه بازنگری و سامان‌دهی نظام سنجش و آزمون دانشگاه پیام نور در دوره ارتقای کیفی از این قرار است:

- استانداردسازی ضوابط طراحی و ویرایش سؤال
- بهبود فرایند طرح و ویراستاری سؤال
- بهره‌گیری از فناوری‌های نوین آزمون بر خط
- نظام‌مند کردن فعالیت بانک سؤال
- تثبیت و تقویت جایگاه مرکز آزمون

مطالعهٔ حاضر از نوع ارزشیابی و پس‌رویدادی و روش آن توصیفی - تحلیلی است. در مطالعهٔ توصیفی متغیری اعمال نمی‌گردد، متغیری کنترل نمی‌گردد و مطالعه به بیان آنچه هست می‌پردازد. این مطالعه با هدف پالایش (حذف

1. Difficulty Index
2. Validity

سؤال‌های ناروا) آزمون‌های نهایی چهارگزینه‌ای در دانشگاه پیام نور به منظور پیش‌گیری از تکرار آزمون در آزمون‌های ریزشی است. تعیین پارامترهای سؤال (درجه دشواری و روایی ماده‌های هر یک از آزمون‌های ریزشی) مطالعه را در زمرهٔ ارزشیابی قرارمی دهد.

روش پژوهش

در این مطالعه در ظاهر واحد تحلیل، ماده‌های آزمون‌های نهایی (درمجموع 1625 ماده) است، ولی به جهت ماهیت مدل کلاسیک و محاسبات انجام شده از جمله همگونی درونی (KR20)، درجهٔ دشواری سؤال (p نسبت پاسخ‌های درست به کل آزمون‌شوندگان در هر آزمون) و روایی هر سؤال (R_{pbis} همبستگی هر سؤال با کل آزمون) به گونهٔ مستقیم به تعداد آزمون‌شوندگان در هر آزمون بستگی دارد، بنابراین واحد تحلیل، پاسخنامه‌های آزمون‌شوندگانی است که مرکز آزمون دانشگاه پیام نور به عنوان آزمون‌های ریزشی معرفی می‌نماید. در مجموع پاسخنامه‌های تعداد 1434 نفر از آزمون‌شوندگان و 1625 ماده از 49 آزمون نهایی یک نیمسال تحصیلی از طریق مرکز آزمون دانشگاه در اختیار پژوهشگر قرار گرفت. به این ترتیب جامعهٔ آماری مطالعهٔ حاضر پاسخنامه‌های 1434 نفر دانشجویان آزمون‌های ریزشی است. از سوی دیگر مطالعه از نوع ارزشیابی است، بنابراین نمی‌توان نمونه‌برداری نمود و لازم است تمام شماری گردد. اگرچه در خلال پاسخ به پرسش‌های پژوهش مشخص گردید تعداد آزمون‌های ریزشی بیش از 49 درس و آزمون‌شوندگان آزمون‌های ریزشی بیش از 1434 نفر بوده و به دلایل مختلف در اختیار پژوهشگر قرار نگرفت و محاسبات بر پایهٔ حجم پاسخنامه‌های دریافت شده انجام گردید (این امر در محدودیت‌های پژوهش منعکس گردیده است). لازم به یادآوری است با موافقت و اقدام مرکز آزمون، ابتدا تعداد کل آزمون‌شوندگان و درصد ریزش هر آزمون در سطح کشور تعیین گردید، پس از تعیین و حذف سؤال‌های ناروا، اصلاح نمره‌های کل آزمون‌شوندگان در سطح کشور به وسیلهٔ مرکز آزمون اعمال و دوباره درصد ریزش تعیین و در اختیار مجری طرح قرار گرفت. حجم گروه نمونه در دسترس هر یک از آزمون‌ها برای مقاصد پژوهشی برحسب نام درس، شمارهٔ درس، حجم نمونه و تعداد ماده‌های هر آزمون در جدول 1 نشان داده شده است.

چنان که اشاره گردید در ارزیابی آزمون‌ها مطالعه از نوع ارزشیابی است و به گفته ثرندایک (1982) لازم است تمام شماری صورت گرفته و نتیجه‌گیری شود. از سوی دیگر گزارش مرکز آزمون دانشگاه نشان می‌دهد در آزمون‌های ریزشی 49 گانه مورداشاره بالغ بر 55767 نفر آزمون‌شونده شرکت داشته‌اند.

همان‌گونه که بیان شد مطالعه از نوع پس رویدادی و ابزار سنجش، آزمون‌های چهارگزینه‌ای است که به وسیلهٔ اعضای هیئت‌علمی متخصص در هر رشته تحصیلی تهیه و همراه با کلید آن در اختیار بانک سؤال قرار گرفته، در آزمون‌های نهایی و پایان ترم مورد بهره‌برداری قرار می‌گیرد و برای پژوهشگر در حکم استفاده از اسناد و مدارک است.

در پژوهش حاضر، شواهد مربوط به روایی محتوا گردآوری شده است، به این معنا که ابزار سنجش یا آزمون‌های نهایی از طریق بانک سؤال به وسیلهٔ اعضای هیئت‌علمی متخصص در هر رشته ساخته و تألیف گردیده و زیر نظر مرکز آزمون دانشگاه بارها در دوره‌های مختلف اجرا و اساساً ملاک ارزیابی آزمون‌شوندگان قرار گرفته و نتایج آن در پژوهش حاضر نیز به عنوان اسناد و مدارک مورد استفاده قرار گرفته، بنابراین از روایی کافی برخوردار است.

در این مطالعه می خواهیم بدانیم که به وسیله یک سؤال تا چه حدمی توان اندازه ملاک[1] را پیش‌بینی کرد، یا یک سؤال با نمره محدود خود (1 یا 0) تا چه حد می‌تواند در نمرهٔ کل سهیم باشد. بنابراین از ضریب همبستگی دو رشته‌ای نقطه‌ای استفاده می‌کنیم. ضریب همبستگی دو رشته‌ای نقطه‌ای شاخص رابطه دو متغیر است که در یکی از آنها فقط دو طبقه نمره وجود دارد و توزیع نمره‌های متغیر دیگر پیوسته است و برخلاف ضریب همبستگی دو رشته‌ای مفروضه دیگری ندارد.

در پژوهش حاضر، به منظور برآورد ضریب اعتبار آزمون‌های 49 گانه آزمون‌شوندگان از فرمول 20 کودر- ریچاردسون (هومن، 1389) استفاده گردید.

روایی و اعتبار ابزار سنجش

در پژوهش حاضر، به منظور برآورد ضریب اعتبار آزمون‌های 49 گانه آزمون‌شوندگان از فرمول 20 کودر - ریچاردسون (هومن، 1389) استفاده گردید. هنگامی که نمره‌گذاری همهٔ سؤال‌ها، دو ارزشی به صورت 0 یا 1 انجام می‌شود، از فرمول شماره 20 کودروریچاردسون[2] استفاده شده است. این روش ضمن آنکه مستلزم تنها یک بار اجرای ابزار سنجش است،

1. Criterion measure
2. Kuder and Richardson

ضریبی که به دست می‌دهد در واقع شاخص هماهنگی درون[1] یا همگونی[2] یعنی میزان تداخل مجموعه ماده‌ها از لحاظ سنجش یک سازه یا ویژگی مشخص است. فرمول 20 کودر- ریچاردسون عبارت است از:

$$\text{اعتبار} = \frac{n}{n-1}\left(1 - \frac{\sum p_i q_i}{S_t^2}\right)$$

در فرمول مورد اشاره n= طول آزمون، p_i= نسبت افرادی که به ماده i ام پاسخ درست داده‌اند، q_i= نسبت افرادی که به ماده i ام پاسخ نادرست داده‌اند، و S_t^2= واریانس کل آزمون است. به این ترتیب مقادیر اعتبار آزمون‌های 49 گانه محاسبه و این نتایج حاصل شد: اعتبار(KR20) آزمون‌های "آشنایی با رایانه 1" ردیف 22 و "روش‌های نمونه‌گیری 1" ردیف 43 و"زبان تخصصی 3" ردیف 35 به ترتیب به مقدار 0/949و 0/923 و 0/911بالاترین مقدار را داشته، و مقادیر 0/6 تا 0/5 کمینه مقدار قابل‌قبول برای آزمون‌های آمار در علوم اجتماعی ردیف 47، الکتریسیته ردیف 13، جغرافیای شهری ردیف 25، متون اختصاصی انگلیسی1 ردیف 23، فیزیک ردیف 16، ایستایی 1 ردیف 30، مبانی جغرافیایی سیاسی ردیف 34، فنون یادگیری زبان ردیف 48 به ترتیب مقادیر 0/597، 0/592، 0/586، 0/584، 0/539، 0/534، 0/534 و 0/528 به دست آمده است. در مقابل اعتبار آزمون‌های کمتر از 0/5 امکان سنجش دانش آزمون‌شوندگان را به شدت تهدید می‌نماید، به این معنا که اگر با قرعه‌کشی (برای مثال با پرتاب سکه) آزمون‌شوندگان به صورت قبول و رد اعلام می‌گردید، اعتبار روش پرتاب سکه 0/5 می‌بود و این معتبرتر از آزمونی است که با صرف هزینه‌های پیدا و پنهان اعتبار کمتر از 0/5 دارد. تعداد 14 آزمون از 49 آزمون‌های نهایی فاقد اعتبار کافی بوده و قادر به ارزشیابی آزمون‌شوندگان نیست.

اگرچه هدف پژوهش ارزیابی پیشرفت تحصیلی آزمون‌شوندگان در 49 آزمون برگزارشده است و تلاش می‌شود برای متوسط هر آزمون در جامعه یک فاصله اطمینان 99 درصدی اولیه ارائه دهد، اما مطالعهٔ حاضر برای هر یک از آزمون‌ها دو متغیر (پارامترهای سؤال) را معرفی می نماید: 1- درجهٔ دشواری سؤال که با نماد p نشان داده شده و نسبت افرادی که به هر سؤال پاسخ درست داده است را مشخص می‌کند. در این مطالعه برای مقادیر $p>0.9$ سؤال خیلی آسان و برای مقادیر $p<0.1$ سؤال خیلی سخت ارزیابی می‌شود. 2- روایی سؤال، قدرت تشخیص سؤال، ضریب همبستگی (دو رشته‌ای نقطه‌ای) سؤال با کل آزمون که با نماد R_{pbis} نشان داده شده و در شاخص‌های روایی سؤال به آن اشاره گردیده، برای مقادیر منفی و نیز مقادیر $R_{pbis}<0.1$ سؤال ناروا تشخیص داده شده و لازم است حذف شود. در صورت لزوم پس از حذف سؤال‌های ناروا مجدداً در دروس منتخب برای متوسط پیشرفت تحصیلی فاصله اطمینان 99 درصدی تعیین می‌گردد. چنان‌که قبلاً اشاره گردید برای پاسخ به پرسش (پارامترهای هر ماده در هریک از آزمون‌های 49 گانه نهایی چقدراست؟) از ضریب همبستگی دو رشته‌ای نقطه‌ای[3] و درجه دشواری سؤال[4] استفاده می‌گردد. فرمول ضریب همبستگی دو رشته‌ای نقطه‌ای (هومن، 1389) چنین است:

$$r_{pbis} = \frac{M_{p\_}M_q}{S_t}\sqrt{pq}$$

M_p= میانگین نمره‌های متغیر ملاک افرادی که در توزیع دو ارزشی به سؤال پاسخ درست داده‌اند.

M_q= میانگین نمره‌های ملاک افرادی است که در طبقهٔ صفر قرار داشته به بیان دیگر به سؤال پاسخ نادرست داده‌اند.

P= نسبت افرادی که به سؤال پاسخ درست داده‌اند. q= نسبت افرادی که به سؤال پاسخ نادرست داده‌اند.

S_t= انحراف استاندارد نمره‌های کل.

بار دیگر یادآوری می‌گردد که در مطالعهٔ ارزشیابی آموزشی، نمونه‌برداری نمی‌توان انجام داد، بنابراین آزمون فرضیه هم صورت نمی‌گیرد و تمام‌شماری انجام شده و لازم است با استفاده از آمار توصیفی به پرسش‌های پژوهش پاسخ داده شود. اگرچه در عمل همهٔ داده‌های خام آزمون‌شوندگان به دلایل مختلف از سوی مرکز آزمون دانشگاه در اختیار مجری طرح قرار نگرفت، ولی با موافقت و اقدام مرکز آزمون، تعداد کل آزمون‌شوندگان و درصد ریزش در هر آزمون در سطح کشور تعیین و در اختیار پژوهشگر قرار گرفت.

در جهت پاسخ به پرسش‌های پژوهش به کمک پاسخنامه‌های دریافت شده، سؤال‌های ناروای هریک از آزمون‌های 49 گانه تعیین و نتایج در اختیار مرکز آزمون قرارداده شد، اصلاح نمره‌های کل آزمون‌شوندگان در سطح کشور به وسیلهٔ کارشناسان مرکز آزمون اعمال و مجدداً درصد ریزش تعیین و در اختیار مجری طرح به شرحی که دریافت‌های پژوهش اشاره خواهد شد قرار گرفت. به این ترتیب ضمن عدم نیاز به

1. Internal Consistency
2. Homogeneity
3. Point biserial
4. Item difficulty

اجرای آزمون فرضیه، برای پاسخ به پرسش‌های پژوهش با استفاده از آمار توصیفی از برنامه‌های موجود به ویژه بسته نرم‌افزاری SPSS استفاده می‌گردد.

یافته‌های پژوهش

برای پاسخ به پرسش پژوهش (پارامترهای هر ماده در هر یک از آزمون‌های 49 گانه نهایی چقدراست؟) از دو شاخص روایی سؤال و درجهٔ دشواری سؤال استفاده می‌شود. برای تعیین روایی سؤال چنان که اشاره گردید از ضریب دو رشته‌ای نقطه‌ای استفاده می‌گردد. ضریب همبستگی دو رشته‌ای نقطه‌ای، شاخص رابطه دو متغیر است که دریکی از آنها فقط دوطبقه نمره (صحیح – غلط) وجود دارد و توزیع نمره‌های دیگر، فاصله‌ای و پیوسته است و مفروضه دیگری ندارد.

معنادار بودن تفاوت ضریب همبستگی دو رشته‌ای نقطه‌ای از صفر در سطح α=0/05 برابر$1/96\div\sqrt{n}$ در سطح α=0/01 برابر $2/58\div\sqrt{n}$ است .از سوی دیگر نسبت افرادی را که به سؤالی پاسخ درست داده‌اند درجه دشواری آن سؤال گویند، در این مطالعه اگر به سؤالی بیش ازاز 90 آزمودنی‌ها پاسخ درست بدهند آن سؤال خیلی آسان و اگر به سؤالی کمتر از 10% آزمودنی‌ها پاسخ درست داده باشند آن سؤال خیلی سخت و اگر ضریب همبستگی دو رشته‌ای نقطه‌ای سؤالی کمتر از $1/64\div\sqrt{n}$ =0/259 باشد سؤال ناروا[1] تلقی می‌شود. گاهی همبستگی‌های ضعیف‌تر نیز برای از دست ندادن سؤال، روا تلقی شده است، به همین منظور آستانهٔ نارواییِ سؤال به کمتر از 0/1 تنزل داده شد. برای تعیین پارامترهای سؤال‌های آزمون یکم (آمار مقدماتی) به تفصیل و هریک از آزمون‌های 49 گانه به اختصار در جدول 4 نشان داده شده است.

تحلیل داده‌ها

آزمون نهایی آمار مقدماتی (شماره درس 1117001) دارای 28 ماده 4 گزینه‌ای است که درکل روی 2450 نفر اجرا گردیده ولی با روشی که در فصل سوم اشاره گردید، پاسخنامه 24 نفر مورد بررسی و محاسبه قرار گرفت. به این ترتیب پاسخ درست هر ماده، نسبت افرادی که به هر سؤال پاسخ درست داده و بالاخره ضریب همبستگی دو رشته‌ای نقطه‌ای هر سؤال در جدول 2 نشان داده شده است.

چنان‌که در جدول 2 مشاهده می‌شود در ستون سمت چپ شماره سؤال‌ها (با نماد ITNO) از شماره 1 تا 28، در ستون‌های دوم تا پنجم تعداد افرادی است که در هر سؤال به هریک از گزینه‌های A تا D پاسخ داده‌اند، ستون ششم با نماد key پاسخ درست هر سؤال را مشخص می‌نماید، ستون هفتم با نماد P نسبت افرادی که به هر سؤال پاسخ درست داده‌اند و ستون آخر درجه روایی هر سؤال یا همبستگی دو رشته‌ای نقطه‌ای هر سؤال با کل آزمون با نماد R_{pbis} رانشان می‌دهد. در ستون هفتم دامنه P از 0/000 (مربوط به سؤال 26 و بسیار سخت) تا 0/833 (مربوط به سؤال 20) و در ستون آخر دامنه R_{pbis} از 0/158- (مربوط به سؤال 16 و ناروا) تا 0/524 (مربوط به سؤال 8 و کاملاً روا) قرار دارد. به این ترتیب ماده‌های خیلی سخت، خیلی آسان و ناروای آزمون آمار مقدماتی تعیین و در جدول 3 نشان داده می‌شود.

چنان‌که در جدول 3 مشاهده می‌شود تعداد 7 ماده از آزمون آمار مقدماتی (ماده‌های 3 و 6 و 7 و 15 و 16 و 18 و 26) از روایی کافی برخوردار نبوده و لازم است از مجموعه سؤال‌های آزمون آمار مقدماتی حذف شود، سؤال 28 نیز خیلی سخت ارزیابی گردید. قابل‌ذکر است 76/45% آزمودنی‌ها قبل از حذف سؤال‌های ناروا ریزش داشتند و پس از حذف سؤال‌های ناروا ریزش آزمودنی‌ها به 58/08% تقلیل یافت، به‌عبارت‌دیگر حدود 451 نفر از کمند ریزش رهایی یافتند. نتایج حاصل از تعین ماده‌های ناروا در هریک از آزمون‌های 49 گانه و تعداد و درصد ریزش قبل و بعد از حذف ماده‌های ناروا در جدول 4 نشان داده شده است.

چنان‌که در سطر آخر جدول 4 مشخص گردیده است در مطالعه حاضر با توجه به درصد ریزش قبل و بعد از حذف ماده‌های ناروا از هر یک از آزمون‌های 49 گانه، درمجموع از میان 55768 نفر آزمون‌شونده، تعداد 8426 نفر (15/10%) از کمند ریزش رهایی یافته و به این ترتیب از اتلاف وقت و سرمایه مادی دانشگاه و دانشجویان جلوگیری می‌شود.

تفسیر نتایج

ارزشیابی آموزشی از مهم‌ترین ارکان دورهٔ آموزشی است. به گفتهٔ هومن (1375: ص 16) فعالیت‌های مرتبط با ارزشیابی عبارت از: اندازه‌گیری[2]، پژوهش[3] و ارزیابی[4] یادگیرنده است.

1. Invalid

2. Measurement
3. Research
4. Appraisal

دو نوع آزمون کلی در ارزشیابی آموزشی به کار می‌رود: آزمونی که برای ارزشیابی گام‌به‌گام پیشرفت‌های دانشجویان، مقایسه دانشجویان، تشخیص و بهبود نقاط ضعف و در آخر به منظور رفع نواقص آموزشی به کارمی رود و نه نمره دادن و صدور گواهینامه، به بیان هومن (1389: ص 192) این نوع آزمون را سازنده[1] و یا نرم- مرجع[2] گویند، در مقابل آزمون‌های جامع یا نهایی[3] است که معرف دانش و مهارت‌های دانشجویان است و به وسیلهٔ بهترین داوران متخصص رشته مورد نظر ساخته و اجرا می‌شود. این نوع آزمون‌ها را ملاک- مرجع[4] نیز می‌گویند.

در پژوهش تقی‌زاده (1390) 6 آزمون در بین کارکنان و مدیران دفتر آموزش و پژوهش استانداری البرز که هر کدام شامل 50 ماده چهارگزینه‌ای بود اجرا گردید و درجه دشواری و قدرت تشخیص ماده‌های خیلی سخت، خیلی آسان و ناروای هریک از آزمون‌ها مشخص گردید. در آن مطالعه متغیر اخلاق فردی دارای 50 ماده چهارگزینه‌ای است و در آن دامنه p از 0/000 تا 0/889 و دامنه r_{pbis} از 0/130- تا 0/422 قرار دارد و تعداد 24 ماده از پرسشنامه اخلاق فردی ناروا ارزیابی گردیده و لازم است حذف شوند. دومین متغیر یعنی متغیر تربیت قرآنی و راهکارهای آن است که دامنه p از 0/000 تا 1/000 و دامنه r_{pbis} از 0/228- تا 0/638 است و تعداد 17 ماده آن از روایی کافی برخوردار نبوده، متغیر دستیابی به اطلاعات هم دارای 50 مادهٔ چهارگزینه‌ای است و در آن دامنه p از 0/000 تا 0/952 و دامنه r_{pbis} از 0/609- تا 0/231 قرار دارد و تعداد 12 مادهٔ پرسش‌نامه ناروا ارزیابی گردیده، در متغیر بعدی یعنی متغیر روابط اجتماعی دامنه p از 0/037 تا 1/000 و دامنه r_{pbis} از 0/247- تا 0/370 قرار دارد و تعداد 6 ماده آن ناروا ارزیابی گردید، پنجمین متغیر، متغیر روابط خانوادگی بود که در آن دامنه p از 0/667 تا 1/000 و دامنه r_{pbis} از 0/184- تا 0/511 قرار دارد و تعداد 16 ماده آن از روایی کافی برخوردار نبود، آخرین متغیر یعنی متغیر سیری در اندیشهٔ سیاسی غرب و اسلام که در آن دامنه P از 0/000 تا 0/947 بوده و دامنه r_{pbis} از 0/238- تا 0/620 قرار دارد و تعداد 32 ماده آن ناروا ارزیابی گردید و لازم است از مجموعه سؤالات پرسش‌نامه حذف گردند. نتایج نشان داد آزمون سیری در اندیشهٔ سیاسی غرب و اسلام ضعیف و فاقد اعتبار بوده است.

در پژوهش اکرامی (1389) در استانداری تهران با استفاده از ضریب همبستگی دو رشته‌ای نقطه‌ای پارامترهای هر ماده ازجمله روایی و سطح دشواری سؤال (نسبت افرادی که پاسخ درست به سؤال داده‌اند) مورد تجزیه و تحلیل قرار گرفت. نتایج نشان داد آزمون دوره‌های غیر حضوری «اصول و مبانی ارتباط» دارای 2 سؤال ناروا و 14 سؤال خیلی آسان؛ دوره «علم و دین» دارای 6 سؤال ناروا، یک سؤال خیلی سخت و 4 سؤال خیلی آسان؛ دوره «شناسایی و استفاده از ظرفیت ذهنی» دارای 2 سؤال ناروا، 9 سؤال خیلی آسان، دورهٔ «سیری در اندیشهٔ سیاسی اسلام و غرب» دارای 4 سؤال ناروا، 4 سؤال خیلی آسان؛ دوره «مدیریت و سازمان از دیدگاه قرآن و نهج‌البلاغه» بدون سؤال ناروا و دارای دشواری مناسب؛ دوره «ارتباط مؤثر با محیط و دیگران» دارای 4 سؤال ناروا و 13 سؤال خیلی آسان بود. در میان‌دوره‌ها سؤال‌های دوره «مدیریت و سازمان از دیدگاه قرآن و نهج‌البلاغه» بهترین طراحی را داشت.

امیریان (1391) با استفاده از مدل IRT پارامترهای سه‌گانه a,b,c و نیز نمره پیشرفت تحصیلی (θ) را در آزمون‌های دستیاری، کارورزی و کارآموزی دانشگاه علوم پزشکی مشهد تعیین نمود. نتایج مطالعه امیریان نشان داد در آزمون یکم دستیاری (10 درصد)، آزمون دوم دستیاری (6/06 درصد)، آزمون سوم کارورزی (0)، آزمون چهارم دستیاری (1/39 درصد)، آزمون پنجم کارآموزی (0)، آزمون ششم کارآموزی (1/67 درصد) سؤال‌ها ناروا تشخیص داده شد و باید از مجموعه سؤال‌ها حذف شود. مقایسه نتایج مطالعه حاضر و دیگران به اختصار در جدول 5 نشان داده شده است.

در پژوهش مقدمی (1392) تعداد شش آزمون دستیاری از مجموعه آزمون‌های دستیاری دانشگاه علوم پزشکی مشهد که دارای 150 ماده سؤال چهارگزینه‌ای است و تعداد 206 دستیار با سطوح مختلف و رشته‌های تحصیلی در آن شرکت داشتند وجود دارد که حد نصاب قبولی با توجه به سطح تحصیلی آنها متفاوت است. به این صورت که کمینه نمره قبولی از سال یک تا چهار به ترتیب نمره 65 و 75 و 85 و 95 و کمینه درصد قبولی از سال یک تا چهار به ترتیب 43/33 و 50 و 56/67 و 63/33 است. برای مثال در آزمون پوست، فرد شماره هفت که در دوره ارتقا 1 به 2 قرار دارد با توجه به اینکه از 150 نمره موفق به کسب نمره قبولی (72 و با درصد 48) شده است، اما پس از حذف سؤالات ناروا و محاسبه درصد موفقیت نهایی (36/73) این دستیار از کسب حد نصاب قبولی

1. Formative test
2. Norm-referenced test
3. Summative test
4. Criterion-referenced test

باز می‌ماند و قادر نخواهد بود به دوره بعدی راه یابد. از سوی دیگر درصد موفقیت افراد شماره 11 و 8 و 6 و 5 و 2 به گونه مؤثر افزایش یافته، برای کسب موقعیت‌های شغلی در شرایط واقعی و مساعدتری قرار می‌گیرند. در دومین گروه یعنی آزمون روان‌پزشکی کلیه دستیاران در آزمون 150 نمره موفق به کسب حد نصاب قبولی در مقطع تحصیلی خود شده‌اند و به دوره بالاتر راه یافته‌اند، اما پس از حذف سؤالات ناروا و محاسبه درصد موفقیت نهایی افراد شماره 1 و 2 و 6 و 21 و 23 و 25 (با کسب به ترتیب 35/955، 56/180، 39/326، 37/079، 38/202 و 47/191 درصد) متأسفانه به دلیل اینکه نتوانسته‌اند نمره قبولی را کسب کنند از ورود به ارتقای بالاتر باز می‌مانند. در آزمون چشم‌پزشکی هم با توجه به محاسبات اولیه فرد شماره 39 که در سطح تحصیلی 1 به 2 قرار دارد موفق به کسب حد نصاب قبولی (45/333) شده است اما پس از حذف سؤالات ناروا و محاسبه درصد موفقیت نهایی از کسب نمره قبولی باز می‌ماند (38/655) و موفق به ارتقا به دوره بالاتر نمی‌شود. از سویی دیگر افراد شماره 25 و 28 که طبق درصد موفقیت اولیه نمرات یکسانی داشتند (84/667) ولی پس از حذف سؤال‌های ناروا و محاسبه درصد موفقیت نهایی فرد شماره 25 نمره بالاتر را کسب می‌کند (89/076) و نفر برتر گروه می‌شود. همچنین در آزمون کودکان، افراد شماره 5 و 46 که طبق درصد موفقیت اولیه (65/333 و 52/000) قبول محسوب می‌شدند پس از محاسبه درصد موفقیت نهایی (62/500 و 47/115) در آزمون دستیاری کودکان مردود و از ارتقا به دوره بالاتر باز ماندند. افراد شماره 37 و 23 هم که اجازه شرکت در 2 ارتقا را دارند فقط توانسته‌اند حد نصاب قبولی در ارتقای 2 به 3 را کسب کنند (56/667 و 53/333) اما از دریافت گواهینامه محروم می‌شوند (52/885 و 50/962) و برای کسب گواهینامه باید در آزمون سال آینده ارتقای دستیاری شرکت کنند. در آزمون دستیاری زنان و زایمان بر اساس نمره 150 ماده و درصد موفقیت اولیه، کلیه دستیاران موفق به کسب حد نصاب شده و به دوره بالاتر راه یافته‌اند اما پس از حذف سؤالات ناروا و محاسبهٔ درصد موفقیت نهایی افراد شماره 1، 3، 8، 34 و 54 نتوانسته‌اند حد نصاب قبولی را کسب کنند (با کسب به ترتیب 36/441، 42/373، 49/153، 33/898 و 42/373 درصد) و قادر به ارتقا به دوره بالاتر نیستند. از سویی دیگر فرد شماره 11 و 36 طبق درصد موفقیت اولیه نمرات برتر (86/000) گروه را دریافت کرده بودند اما پس از حذف سؤالات ناروا فرد شماره 25 نمره بالاتر (89/831) را در گروه به دست آورد. همچنین در گروه ارولوژی طبق محاسبات اولیه بر روی 150 ماده، همهٔ دستیاران به‌جز فرد شماره 5 (با درصد 46/667 که اجازه شرکت در ارتقای 3 به 4 و گواهینامه را داشته و موفق نشده در هیچ سطحی نمره قبولی را کسب کند) موفق به کسب حد نصاب قبولی شدند، اما پس از حذف سؤالات ناروا که در این آزمون تعداد آنها 57 ماده بود محاسبات نهایی روی 93 ماده انجام شد و درصد موفقیت نهایی دستیاران نشان می‌دهد که افراد شماره 7 و 11 (49/462 و 33/333) با توجه به این که در محاسبات اولیه قبول شدند پس از حذف سؤالات، مردود و از راه‌یابی به دوره بالاتر باز ماندند. از سویی دیگر درصد موفقیت افراد شماره 1، 2، 6، 8، 9، 10، 12، 15، 16، 18، 19 و 20 به گونه مؤثر افزایش یافته است.

در مطالعهٔ حاضر با توجه به درصد ریزش قبل و بعد از حذف ماده‌های ناروا از هر یک از آزمون‌های 49 گانه، درمجموع از میان 55768 نفر آزمون‌شونده، تعداد 8426 نفر (15/10%) از کمند ریزش رهایی یافته و به این ترتیب از اتلاف وقت و سرمایه مادی دانشگاه و دانشجویان جلوگیری می‌شود.

بنا بر مدارک موجود در بانک سؤال، پاسخنامه‌های 430000 (چهارصد و سی هزار) نفر افراد ریزشی در نیم سال اول سال تحصیلی 92-1391 جمع‌آوری و آماده تحویل، به منظور تجزیه و تحلیل ماده‌های هر آزمون گردید؛ ولی به سبب پیش‌گیری از ایجاد اختلال در نظام متداول اعلام نمرات دانشگاه و فراهم نبودن جواز اعمال نمرات جدید پس از حذف سؤالات ناروا، فقط بخشی از این تعداد در اختیار مجری طرح قرار گرفت. می‌توان تخمین زد

$$\frac{15/109\times 430000}{100} = 64968 \approx 65000$$

یعنی قریب 65 هزار نفر از مجموع 430000 نفر افراد ریزشی، به قبولی‌های نیم سال مورد نظر اضافه می‌گردد؛ اگر هزینه‌های طرح سؤال، کاغذ، چاپ و تکثیر و هزینه‌های قرنطینه، مراقب و اجرای آزمون و اعلام نتایج برای حدود 50 نفر در هر آزمون به طور متوسط 000 000 10 ریال معادل یک میلیون تومان در نظر گرفته شود، محاسبات اولیه نشان می‌دهد $\frac{65000\times 1000000}{50} = 1{,}300{,}000{,}000$ معادل یک میلیارد و سیصد میلیون تومان صرفه‌جویی در هزینه‌های مرکز آزمون و امتحانات دانشگاه می‌شود. افزون بر آن پالایش سؤال‌های متجاوز از 5000 آزمون سراسری برگزار شده در هر ترم تحصیلی زیر نظر مرکز آزمون دانشگاه و بانک سؤال، در کمتر از 10 روز پس از برگزاری هر آزمون

صرفه‌جویی‌های نجومی برای دانشگاه از یک سو و دست‌مایه بودجه تلاش پژوهشی عُشری از اعشار هزینه‌های پیدا و پنهان مورد اشاره، در جهت پیش‌گیری و ممانعت از ریزش و تکرار آزمون‌های نهایی برای گروه پژوهشی متخصص و مجری طرح خواهد بود.

منابع

اکرامی، محمود (1389). "ارزیابی دوره‌های آموزشی غیرحضوری ضمن خدمت کارکنان دولت در استان تهران"، طرح پژوهش با استفاده از اعتبارات استانداری نهران انجام شده است. چاپ نشده.

آلن، مری جی.؛ ین، وندی ام (1387). مقدمه‌ای بر نظریه‌های اندازه‌گیری (روان‌سنجی) (ترجمه علی دلاور)، چاپ سوم. تهران: نشر ذره.

امیریان، سحر (1391). "تعیین پارامترهای سه‌گانه آزمون‌های چهارگزینه‌ای دانشگاه علوم پزشکی مشهد بر پایه مدل IRT"، پایان‌نامه کارشناسی ارشد مدیریت آموزشی. دانشگاه پیام نور تهران، چاپ نشده.

امین، محمد مهدی؛ شایان، شهرام؛ هاشمی، حسن؛ پورصفا، پریناز؛ ابراهیمی، افشین (1389). "تجزیه و تحلیل آزمون دروس دانشجویان کارشناسی بهداشت با استفاده از برنامه نرم‌افزاری آنالیز سؤالات چندگزینه‌ای بر اساس نظریه تستی کلاسیک (CTT)"، *مجله ایرانی آموزش در علوم پزشکی (ویژه‌نامه توسعه آموزش)*، شماره 5.

ایبل، رابرت ال. فریزبی، دیوید دی (1381). اصول اندازه‌گیری در علوم تربیتی (ترجمه حسین سپاسی). اهواز: انتشارات دانشگاه شهید چمران.

تقی‌زاده، تهمینه (1390). "ارزیابی دوره‌های کوتاه‌مدت دفتر آموزش و پژوهش استانداری تهران"، پایان‌نامه کارشناسی ارشد مدیریت آموزشی، دانشگاه پیام نور تهران، چاپ نشده.

دلاور، علی؛ زهراکار، کیانوش (1387). سنجش و اندازه‌گیری در روان‌شناسی، مشاوره و علوم تربیتی، چاپ اول. تهران: نشر ارسباران.

ربانی، علی؛ فرزیان‌پور، فرشته؛ زمانی، غلامرضا (1385). "ارزیابی درونی در گروه بیماری‌های کودکان دانشکده پزشکی دانشگاه علوم پزشکی تهران"، مجله بیماری‌های کودکان ایران، دوره 16 شماره 3.

سام‌شریفی، اسماعیل؛ دلاور، علی؛ بلوکی، آزاده؛ شعبانی، سمیه (1391). ارزشیابی آزمون نظری آزمون گواهینامه رانندگی بر اساس نظریه سؤال – پاسخ و مقایسه آن با نظریه کلاسیک آزمون.

سیف، علی‌اکبر (1381). روش‌های اندازه‌گیری و ارزشیابی آموزشی، ویرایش دوم. تهران: نشر دوران.

سیف، علی‌اکبر (1386). اندازه‌گیری، سنجش و ارزشیابی آموزشی، ویرایش چهارم. تهران: نشر دوران.

شریفی، حسن‌پاشا (1381). اصول روان‌سنجی و روان آزمایی، چاپ هفتم. تهران: انتشارات رشد.

کرم دوست، نوروزعلی (1383). "بررسی رابطه ارزشیابی دانشجویان دانشکده روان‌شناسی و علوم تربیتی از تدریس استادان با میانگین نمرات آنان از درس استادان در سال تحصیلی 77-78 تا 80-79"، *مجله روان‌شناسی و علوم تربیتی*، شماره 34، ص76-57.

کیامنش، علیرضا (1382). روش‌های ارزشیابی آموزشی. چاپ دهم، تهران: انتشارات دانشگاه پیام نور.

مقدمی، نجمه (1392). "برآورد پارامترهای سؤال‌های چهارگزینه‌ای در آزمون‌های نهایی دستیاری دانشگاه علوم پزشکی مشهد"، پایان‌نامه کارشناسی ارشد مدیریت آموزشی. دانشگاه پیام نور تهران، چاپ نشده.

هومن، حیدرعلی (1370). اندازه‌گیری روانی و تربیتی و فن تهیه آزمون و پرسش‌نامه. تهران: انتشارات پیک فرهنگ.

هومن، حیدرعلی (1373). "مقایسه مدل تک پارامتری راش و مدل دو پارامتری"، پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی، چاپ نشده.

Brennan, R.L. (1989) Introduction to Problems, and, Practical Lssvesin Equating. Applied Psychological Measurement, VOL.17,NO 3

Brennan, R.L. (1989) .Introduction to Problems, and, Practical Lssvesin Equating. Applied

Cook, L.L. & Peterson, N.S. (1987) Problems Reacted to the use of conventional and Item

Hambletow, R.K. & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational measurement, 14, 75-96.

Hambletow, R.K. & Cook, L.L. (1983). The robustness of item response models effects of test length and sample size owe the precisiow of ability estimates. In D Jweis (ED) New horizons in testing (pp.31-94). New York: ACADEMIC PRESS.

MacDonald. & Paun0nen, S.V. (2002). A Monte Carlo comparison of Item and person satisfies Based on Item Response theory versus classical test theory. Educational and psychological measurement vole 62 no.6, pp.921-943. University …. Wester ontazio.

Mony, H. (2006). A multilevel Bayesian item response. Theory method for scaling socioeconomic status in international studies of Education Journal Educational and Behavioral states ties,vol,31,no.1,pp.63-79.

Oshima, T.C.O. (1994). The effect of speediness on parameter estimation in item response theory. Journal of Educational measurement, 31,200-219.

Quereshe. M.Y.T.L. Fisher (1997). The effect speediness on parameter estimation in item response theory. Journal of Educational measurement .31.200-219.

Response Theory Equating Methods in Less than optimal Circumstances. Applied Psychological Measurement, VOL.17

Stage, C. (2000). A comparison between item analgim Based o w Item Response theory and classical test theory. A study of the sweat subset ERC.

Z.JR. Slater. (1992). An analysts of the utility of item response theory for the practitioner in a university departmental seething. Source: DAI-A54:03, p.872, Sep 1993.