فصلنامه روانشناسی تربیتی
دانشگاه آزاد اسلامی واحد تنکابن
سال اول، شماره ۱، پاییز ۱۳۸۸
صص۱۰۱-۱۱۲

تحلیل علمی آزمون ها بر اساس نظریه کلاسیک و نظریه صفت مکنون

محمد علی احمدوند[1*]

۱) استادیار. عضو هیأت علمی دانشگاه آزاد اسلامی واحد تهران جنوب

* نویسنده مسئول: Ahmadvandma@yahoo.com

تاریخ دریافت مقاله ۸۸/۵/۱۵ تاریخ آغاز بررسی مقاله ۸۸/۵/۲۰ تاریخ پذیرش مقاله ۸۸/۸/۱۱

چکیده

تحلیل علمی آزمون ها، متداولترین وسیله برای ارزشیابی یادگیری و پیشرفت تحصیلی و دانشگاه ها است. برای بررسی کیفیت ارزشیابی در دانشگاه تربیت معلم، چهار آزمون روانسنجی، انگیزش و هیجان، احساس و ادراک و روانشناسی رشد از بین مجموعه آزمون ها بطور تصادفی انتخاب، و پاسخنامه های آنان تحلیل گردیده اند شاخص های آماری بکار رفته عبارتند از : درجه دشواری، قدرت تمیز، واریانس و مجذور خی ۱۲۹ سؤال که روایی و پایایی آزمون ها تعیین گردید و به سؤال های پژوهش پاسخ داده شد و فرضیه های آن مورد آزمون قرار گرفت. از ۲۵ تا ۷۵ درصد سؤال های طرح شده در آزمون های نا مناسب شناخته شدند . میانگین نمرات و پاسخ نامه های دانشجویان تقریباً ۱۲ الی ۱٤ بود، در حالی که میانگین نمره های داده شده ب ه آنان خیلی بالاتر از میانگین های مذکور بوده است . ضرایب دشواری هر ٤ آزمون خارج از سطح بهینه و ضریب تمیز در آزمون انگیزش و هیجان بالاترین و روانشناسی رشد پایینترین بود. در هر ٤ آزمون این شاخص ها خارج از سطح بهینه قرار داشت . آزمون ها از روایی و پایایی مناسب برخوردار نبودند . نتیجه پژوهش ض رورت برگزاری کارگاه های تهیه آزمون و ارزیابی را تأیید می کند.

کلید واژه گان: ارزشیابی، ضریب دشواری، ضریب تمیز، نظریه صفت مکنون.

مقدمه

در دانشکده تربیت معلم دانشگاه آزاد اسلامی واحد تهران جنوب، میزان یادگیری و پیشرفت تحصیلی دانشجویان از طریق برگزاری امتحانات متمرکز در پایان هر نیم سال انجام می گیرد . اندازه گیری [1] و ارزشیابی [2] یادگیری بصورت آزمون های چهار گزینه ای و یا تشریحی رخ می دهد.نمرات دانشجویان اعلام می شود و اوراق امتحانی به دایره امتحانات جهت بایگانی سپرده می شود وهیچگونه تحلیلی پیرامون این امتحانات صورت نمی گیرد تا سهولت و دشواری آزمون ها و کیفیت تک تک سؤال ها از نظر شاخص های آماری تعیین گردد وروشن شود که آیا هدف های ارزیابی تحقق یافته اند یا خیر؟ با مساعدت و کمک مسؤ لین آموزشی و پژوهشی دانشکده، تعدادی از سؤال های آزمون ها ی اجراشده بر اساس نظریه کلاسیک [3] و نظریه مکنون [4] در قالب طرح پژوهشی مصوب، تحلیل گردید.

توجیه علمی موضوع . متداولترین ابزار اندازه گیری که در تعدادی از گروه های آموزشی نظیر روان شناسی، مشاوره، آموزش ابتدایی، مدیریت و برنامه ریزی و... بکار می رود، آزمون های چها ر گزینه ای است . در چند دهه اخیر ارزشیابی آموزشی را سرمایه گذاری برای انسانها و پیشرفت آنان تلقی کرده اند و از آنان به عنوان فرایندی منظم برای توصیف کردن و هدایت نمودن کیفیت یادگیری و پیشرفت تحصیلی و اطمینان یابی از چگونگی فعالیت های آموزشی بهره برده اند (بلوم، مگان، ۱۳۸۱). برای ارزشیابی شایسته لازم است طراحان سؤال های آزمون ها با هدف های آموزشی آشنا باشند و بتوانند به کمک بودجه بندی مطالب، هدف ها را در قالب های مشخص شده، تعیین کنند و در حوزه شناختی با توجه به طبقه های دانش، فهمیدن، کاربستن، تحلیل، ترکیب و ارزشیابی، سؤال های مناسبی را تهیه نمایند (شریفی، ۱۳۸۳، گنجی و ثابت ۱۳۸۳). برای تهیه سؤال های چند گزینه ای، قواعدی تدوین شده است که لازم است طراحان سؤال های امتحانی از آنها سود جویند . آزمون ها باید از روایی [5] و پایایی [6] برخوردار باشند. آزمون، زمانی روایی دارد که بـرای اندازه گیری آنچه مورد نظر است، مناسب باشد . روایی محتوایی نشان می دهد که نمونه سؤال های مورد استفاده در یک آزمون تا چه حد معرف جامعه سؤال هایی است که می توان از محتوای مورد نظر تهیه کرد. از این رو، باید سؤال های آزمون، نمونه کاملی از هدف ها و محتوای درسی باشند(گنجی، ۱۳۸۱، برنی، ۱۹۸۸). سؤال هایی که طبق قواعد درست تهیه شده باشند به روایی آزمون می افزایند . هر سؤال که از هدف اصلی آزمون جدا افتد به سهم خود از روایی کل آزمون می کاهد. آزمون های متشکل از سؤال های بسیار دشوار یا بسیار ساده از سطح روایی خوبی برخور دار نیستند(کوهن و همکاران، ۱۹۹۶). یک آزمون زمانی پایایی دارد که اگر آن را در فاصله زمانی کوتاه چندین بار به گروه واحدی از دانشجویان بدهیم نتایج حاصل نزدیک به هم باشد . از روش های ساده تعیین پایایی، یکی استفاده از روش دو نیمه کردن آزمون می باشد (بری من، ۱۹۹۹، ویک فیلد، ۱۹۹۶).

تحلیل سؤال ها و مراحل تحلیل . هدف از تحلیل سؤال های آزمون وارسی تک تک سؤال ها و تعیین میزان دقت و نارسایی های آنها است . در تحلیل سؤال های آزمون، نقاط قوت و ضعف یک آزمون و کیفیت همه سؤال های آن تعیین می شود . طراح سؤال پی می برد که چگونه دانش جویان به آزمون و همچنین به تک تک سؤال ها واکنش نشان داده اند و می داند که کدام گزینه نقش خاصی را در ارزشیابی نداشته است . متداول ترین مورد استفاده اطلاعات بدست آمده از تحلیل سؤال ها،

[1] . Measurement
[2] . Evaluation
[3] . Classic theory
[4] . Latent trait theory
[5] . Validity
[6] . Reliability

انتخاب سؤال های بهتر و مناسب تر برای تشکیل فرم نهایی آزمون است. تحلیل سؤال ها وضعیت دانشجویان را از لحاظ درک مطالب و مفاهیم درس و نقاط ضعف تدریس به خوبی روشن می سازد و می تواند در بهبود روش تدریس مؤثر باشد (ویرسما[1] و دیگران ۱۹۹۰-کرای گید[2]،۲۰۰٤). در روش کلاسیک تعیین می شود که چند نفر از دانشجویان گزینه درست سؤال را انتخاب کرده اند و هر یک از گزینه های انحرافی چند نفر را از گروه قوی و گروه ضعیف به خود جلب کرده اند . دوانی[3] در سال ۱۹٦۷ روشی برای تجزیه و تحلیل سؤال ها ارایه داد. وی ابتدا اوراق را از بالاترین نمره به پایین ترین نمره مرتب کرد. آنگاه یک سوم اوراق را به ترتیب از بالا ان تخاب کرده و آنها را گروه بالا نامید و یک سوم اوراق را از کمترین نمره به نام گروه پایین در نظر گرفت و بقیه برگه های امتحانی راکنار گذاشت . در مرحله بعد برای هر سؤال، تعداد گزینه هایی را که شاگردان هر دو گروه انتخاب کرده اند جداگانه شمارش و نتایج را در کارتی درج نمود و دو شاخص دشواری و قوه تمیز را برای هر سؤال تعیین نمود (سیف، ۱۳۸۲؛ نفیسی و زند پارسا، ۱۳۷٦). درصد کل آزمون شوندگان ی که به یک سؤال جواب درست می دهند ضریب دشواری [4] آن سؤال را بدست می دهد . هر اندازه ضریب دشواری یک سؤال بزرگتر باشد آن سؤال آسان تر است . سؤال هایی بهتر هستند که ضریب دشواری آنها از ۱ کمتر و از صفر بیشتر و به عدد ۰/۵ نزدیک باشند. آلن و ین[5] (۱۹۷۹) معتقدند سطح بهینه دشواری برای سؤال های چهار گزینه ای در وسط ۰/۲۵ و ۱ یعنی در حدود ۰/٦ قرار دارد (سالکانید[6]، (سالکانید[6]، ۲۰۰۸). به طور کلی ضرایب دشواری بین ۰/۳ ۰/۳ تا ۰/۷ حداکثر اطلاع را درباره تفاوت بین آزمودنی ها به دست می دهند (سلیمی زاده، ۱۳۷۷). ضریب تمیز[7] قدرت سؤال را در تمایز گذاری یا تشخیص بین گروه قوی و گروه ضعیف آزمون شوندگان مشخص می کند . هر قدر این ضریب بزرگتر باشد قوه تمیز آن بیشتر است(پیر اون لاین نت[8]، ۲۰۰۹).

در تحلیل سؤال های آزمون، علاوه بر تعیین ضریب های دشواری و تمیز برای هر سؤال، بررسی نحوه پراکندگی پاسخ های مربوط به گزینه های انحرافی هر سؤال نیز ضروری است. گزینه های انحرافی سؤال ها باید طوری تهیه شوند که بتوانند افراد ضعیف را به خود جلب کنند. هدف از طرح این سؤال این است که افراد گروه قوی گزینه درست و افراد گروه ضعیف یکی از گزینه های غلط را انتخاب کنند. مشکل ترین قسمت در ساخت یک سؤال گزینه های غلط آن است که با وجود غلط بودن، به پاسخ درست بیشتر شباهت داشته باشند . در صورتی یک سؤال به خوبی عمل می کند که افراد ضعیف بیشتر از افراد گروه قوی گزینه های انحرافی آن سؤال را انتخاب نمایند (سیف، ۱۳۸۲؛ ردی[9]، ۲۰۰۹).

تحلیل سؤال بر اساس نظریه صفت مکنون . در نظری کلاسیک، ضریب دشواری، قدرت تمیز، واریانس و شاخص های دیگر آماری تعیین می گردند و بر اساس این شاخص ها می توان سؤال امتحانی را تحلیل نمود . گرچه به روش کلاسیک، ایرادهایی وارد شده است. اما می تواند برای طراحان سؤال راهگشا باشد و آنها را به سوی ساخت سؤال های مناسب تر رهنمون شود . منتقدین روش کلاسیک مدعی اند که ویژگی های سؤال ها به گروه آزمونی وابسته هستند . به طور مثال، اگر سؤال ها ساده باشند آزمودنی ها قوی تر و اگر سؤال ها دشوارتر باشند آزمودنی ها ضعیف تر ارزشیابی می شوند . اگر آزمون برای افراد قوی اجرا شود، دشواری کمتری را نشان

[1] . Wiersma
[2] . Craighead
[3] . Downie
[4] . Difficulty index
[5] . Allen and yen
[6] . Salkind
[7] . Discrimination index
[8] . Www.Pareonline.net
[9] . Redie

می دهد تا زمانی که همین آزمون برای افراد ضعیف اجرا شود. همچنین نمره گذاری در روش کلاسیک غیر واقع ی است و به هر سؤال ارزش یکسانی تعلق می گیرد. در حالی که سؤال های مختلف از نظر محتوایی که مورد پرسش قرار می دهند و از نظر دشواری یکسان نیستند که بتوان برای آنها ارزشی برابر قایل شد . در حال حاضر روش تحلیل صفت مکنون [1] یا نظریه سؤال-پاسخ بر این فرض متکی است که یک ویژگی زیربنایی وجود دارد که به شخص امکان می دهد تا در یک تکلیف شناختی معین، موفقیت کسب کند. افزون بر این، این پندار وجود دارد که هر چه شخص از این صفت بیشتر برخوردار باشد، در آزمون مربوط عملکرد بهتری خواهد داشت . در نظریه سؤال-پاسخ هر چه توانایی فرد از می زان دشواری سؤال بیشتر باشد، احتمال پاسخ درست فرد به سؤال نیز بیشتر می شود و هر چه جایگاه فرد بر روی محور توانایی ها از میزان دشواری سؤال کمتر بشود احتمال پاسخ درست این فرد به سؤال نیز کمتر و کمتر می شود.

با استفاده از نظریه سؤال-پاسخ می توان برخی ویژگی های سؤال را تعیین کرد. این ویژگی سؤال برای بسیاری مقاصد آزمون سازی از جمله تحلیل سؤال های آزمون مفیدند. ویژگی های هر سؤال به صورت نمودار نشان می دهند و آن را منحنی ویژگی سؤال [2] (آی سی سی) می نامند. منحنی ویژگی سؤال، احتمال پاسخ درست دادن به هر سؤال را به توانا یی آزمون شونده ربط می دهد . به سخن دیگر، منحنی یا خم ویژگی سؤال یک بازنمایی نموداری از رابطه بین احتمال پاسخ درست دادن به یک سؤال و موقعیت آزمون شون ده در صفت مورد اندازه گیری توسط آزمون است (گلاورز، خرازی ۱۳۷۸). از روی منحنی ویژگی سؤال می توان ضریب دشواری و تمیز سؤال را تعیین کرد. ضریب دشواری عبارت است از نمره معیاری که در آن ۵۰ درصد آزمون شوندگان سؤال را درست پاسخ داده اند و ضریب تمیز برابر با شیب منحنی ویژگی سؤال است (راسیا[3]، ۲۰۰۶).

برای تهیه منحنی ویژگی سؤال نسبت یا درصد آزمون شوندگانی که آن سؤال را دست ج واب داده اند در رابطه با نوعی ملاک، مثلاً نمره کل آزمون آنها رسم می شود. بر روی محور افقی نمره کل آزمون و بر روی محور عمودی نسبت آزمون شوندگانی که به سؤال پاسخ درست داده اند، مشخص می شود . در این نظریه می تواند ضریب دشواری سؤال را مستقیماً به سطح زیر منحنی ویژگی سؤال ربط داد. هر چه زیر منحنی ویژگی سؤال بیشتر باشد، ضریب دشواری سؤال بزرگ تر است . هر چه منحنی ویژگی سؤال حالت پلکانی بیشتری داشته باشد همبستگی بین آن سؤال و کل آزمون بیشتر است (سالکانید، ۲۰۰۸؛ سیف، ۱۳۸۲).

سؤال های پژوهش

۱. توزیع گزینه درست در بین ٤ گزینه به چه صورت است؟

۲. میانگین نمره های دانشجویان در هر آزمون چقدر است؟

۳. شاخص های آماری هر سؤال در آزمون ها چه وضعی دارند؟

٤. چند درصد سؤال های هر آزمون نا مناسب هستند؟

۵. آزمون ها با توجه به سطح بهینه هر شاخص چه وضعی دارند؟

٦. منحنی ویژگی سؤال های انتخابی چه شکلی دارند؟

فرضیه های تحقیق

[1] . Latent trait theory
[2] . ICC(Item Characteristic Curve)
[3] . Rasiah

۱. تفاوت دشواری یا سهولت آزمون ها قابل توجه است.

۲. شاخص های آماری آزمون ها در سطح بهینه قرار ندارند.

۳. درصد قابل توجهی از سؤال های هر آزمون نا مناسب هستند.

٤. پایایی و روایی آزمون روانسنجی بهتر از آزمون های دیگر است.

۵. طراحان با طرح سؤال های مناسب و علمی آشنایی کافی ندارند.

هدف پژوهش، پاسخگویی به سؤال ها و آزمون فرضیه ها است . نتایج پژوهش در اختیار طراحان آزمون ها گذاشته می شود تا در بهبود بخشیدن به کیفیت سؤال های خود و نگهداری سؤال های مناسب در بانک سؤال و حذف سؤال های نا مناسب اقدام کنند.

روش تحقیق

از بین گروه های آموزشی دانشکده تربیت معلم، گروه روان شناسی با داشتن بالاترین آزمون های چهار گزینه ای انتخاب شد و از بین ۲۰ آزمون اجرا شده ٤ آزمون احساس و ادراک، انگیزش و هیجان، روان سنجی. روان شناسی رشد به صورت نمونه و به قید قرعه برگزیده شدند. پاسخنامه های امتحانی که توسط ۲۹۷ نفر از دانشجویان تکمیل شده بود از اداره امتحانات دریافت گردید و با تدوین جدول مشخصات برای هر سؤال و نحوه پاسخدهی دانشجویان، کارت تحلیل سؤال تهیه گردید.

همه سؤال ها از نظر ساخت (تنه سؤال، گزینه انحرافی، گزینه کلید) و رعایت نکات استاندار د مورد توجه قرار گرفت و از روش های آماری تحلیل سؤال بهره برده شد و شاخص های آماری نظیر ضریب دشواری، ضریب تمیز، واریانس و مجذورخی برای هر سؤال محاسبه گردید و سؤال ها از حیث گزینه های انحرافی و شاخص ها در سه دسته مناسب، قابل اصلاح و نا مناسب تفکیک شدند. میانگین آزمون ها در گروه های قوی، میانه و ضعیف برای آزمون ها محاسبه گردیدند . توزیع نمرات و نمای آنها به دست آمد . پایایی از راه دو نیمه کرد ن آزمون ها تعیین شد و رابطه بین ضریب دشواری و قدرت تمیز با استفاده از نظریه صفت مکنون برای آزمون ها با شکل، نشان داده شد و منحنی ویژگی سؤال نیز در چند مورد رسم گردید.

یافته ها

برای تحلیل سؤال های آزمون ها از کارت تحلیل سؤال استفاده شد که در اینجا اولین سؤال روانشناسی رشد به عنوان نمونه درج می گردد:

جدول شماره ۱. کارت تحلیل سؤال

عنوان آزمون: روان شناسی رشد	**شماره سؤال: ۱**
متن سؤال	
مرحله چهارم رشد اخلاقی کلبرگ کدام است؟	
الف. جهت گیری هدف- وسیله ای	
ب. جهت گیری حفظ کردن نظم اجتماعی*	
ج. جهت گیری اصول اخلاقی همگانی	
د. جهت گیری قرارداد اجتماعی	

گروهها	الف	ب	ج	د	سفید	جمع	ملاحظات
گروه بالا		۱۸		۲		۲۰	
گروه پایین	۱	۱٤	٤	۱		۲۰	

شاخص های محاسبه شده	$N_1=14$	$N_U=18$		$N=40$	
$d=0/20$	$X^2=5$	$pq=0/16$	$q=0/20$	$P=0/80$	$P_1=70\%$	$P_U=90\%$

تحلیل: ۱۸ نفر از گروه قوی و ۱٤ نفر از گروه ضعیف گزینه درست را انتخاب کرده اند.
سؤال آسان است اما با توجه به پراکندگی خوب و معنا دار بودن مجذور خی بعنوان سؤال آسان، می توان در اول آزمون از آن استفاده کرد.

۱۲۹ سؤال ٤ آزمون بر اساس محاسبه شاخص ها ارزیابی شد و مناسب بودن، قابل اصلاح و ویژگی نا مناسب بودن را دارد.

طبق بررسی های انجام شده، آزمون انگیزش و هیجان از لحاظ درصد سؤال های مناسب نسبت به ۳ آزمون دیگر از کیفیت بهتری برخوردار بود . ۷۵ درصد سؤال های این آزمون مناسب ارزیابی شد . آزمون روانسنجی رتبه دوم و آزمون حساس و ادراک رتبه سوم و آزمون روان شناسی رشد رتبه چهارم را به دست آورد.

جدول شماره ۲. وضعیت آزمون ها بر حسب سؤال های مناسب، قابل اصلاح و نا مناسب

درصد سؤال های مناسب	تعداد سؤال	سؤال های نا مناسب	سؤال های قابل اصلاح	سؤال های مناسب	آزمون
۷۵ درصد	۲۹	٤	۳	۲۲	انگیزش و هیجان
۵۷ درصد	۳۰	٦	۷	۱۷	روانسنجی
۵۵ درصد	٤۰	۱۰	۸	۲۲	احساس و ادراک
٤۳ درصد	۳۰	۱۲	۵	۱۳	روان شناسی رشد

آزمون روانشناسی رشد از لحاظ انتخاب سؤال، ضعیف است و ۵۷ درصد سؤال قابل اصلاح و نا مناسب دارد.

جدول شماره ۳. مقایسه میانگین های نمرات دانشجویان در ٤ آزمون

نما	میانگین گروه ضعیف	میانگین گروه میانه	میانگین گووه قوی	میانگین کل	آزمون ها
۱٤	۱۰/۵۰	۱٤/۵۲	۱۷/۱۷	۱٤/۱۸	روان شناسی
۱٤	۱۰/۳۵	۱٤/۳۵	۱۷/٦٤	۱٤/۱۱	انگیزش و هیجان
۱٤	۱۰/۸۰	۱٤/۱۰	۱٦/۹۰	۱۳/۹۳	احساس و ادراک
۱٤	۹/۸۷	۱۲/۱۲	۱٤/۸۳	۱۲/۲۷	روانشناسی رشد

آزمون روانشناسی رشد دارای کمترین میانگین و آزمون روانسنجی دار ای بالاترین میانگین است . با توجه به میانگین کل نمرات در ٤ آزمون، روانسنجی بالاترین میانگین و روانشناسی رشد پایین ترین میانگین را به خود اختصاص داد. نمای نمرات در هر ٤ آزمون نمره ۱٤ بود. طراحان سؤال با آگاهانه و یا نا آگاهانه گزینه درست را در بین ٤ گزینه (الف، ب، ج، د) به نسبت تقریباً متوازی جای نداده اند و این خود یک نقص تلقی می شود.

جدول شماره ٤. مقایسه محل گزینه درست در بین چهار گزینه آزمون ها

د	ج	ب	الف	آزمون ها
۸	۱٦	۱۰	٦	احساس و ادراک
۹	۱۱	٦	۳	انگیزش و هیجان
۵	۷	۱۰	۸	روان سنجی
۷	٦	۱۱	٦	روانشناسی رشد

گزینه درست باید به نسبت ۲۵ درصد بین ٤ گزینه توزیع شود . در آزمون احساس و ادراک، گزینه کلید بیشتر در قسمت ج قرار گرفته است (٤٠ درصد). در آزمون انگیزش و هیجان نیز، گزینه کلید بیشتر در قسمت ج قرار گرفته است (۳۸ درصد). در آزمون روان سنجی و روان شناسی رشد، گزینه کلید بیشتر در قسمت ب قرار گرفته است(به ترتیب ۳۳ درصد و ۳۷ درصد).

ضرایب دشواری آزمون ها متفاوت هستند . برخی از آنها دارای سادگی زیاد و برخی تا حدی دشواری دارند . آزمون احساس و ادراک در مقایسه با آزمون های دیگر ساده تر و آزمون رشد دشوارتر ارزیابی شد . قدرت تمیز آزمون انگیزش و هیجان بیشتر از بقیه بود.

جدول شماره ۵. مقایسه شاخص های ضریب دشواری و ضریب تمیز و واریانس در ٤ آزمون

آزمون ها	ضریب دشواری	ضریب تمیز	واریانس
احساس و ادراک	۰/۷۲	۰/۳۱	۰/۱٦
انگیزش و هیجان	۰/٦۸	۰/۳٦	۰/۱۸
روان سنجی	۰/٦۸	۰/۳۳	۰/۱۸
روانشناسی رشد	۰/٦٥	۰/۲٦	۰/۱۸

از بین ۳ شاخص فقط واریانس آزمون ها در حد بهینه قرار داشت.

برای ۳۰ سؤال روان شناسی رشد، میانگین کل نمرات دانشجویان ۱۲/۲۷ و نمای نمرات آزمون ١٤ شد. طرح سؤال، تمایل بیشتری در نهادن گزینه کلید در قسمت ب داشته است(۳٦/٦ درصد).

از ۳۰ سؤال، تعداد ۱۳ سؤال مناسب، ۵ سؤال قابل اصلاح و۱۲ سؤال نا مناسب، تشخیص داده شد. با توجه به منحنی ویژگی سؤال برای سؤال شماره ۱، نسبت پاسخدهی در هر دو گروه بالا و پایین زیاد است . به منظور اجتناب از طولانی شدن مقاله به درج منحنی ویژگی سؤال شماره ۱ از آزمون روان شنا سی رشد اکتفا می شود.

نمودار شماره ۱. منحنی ویژگی سؤال شماره ۱ روان شناسی رشد

تحلیل: دانشجویان دارای نمرات ١٦ و ١٤ در حد بالایی به سؤال پاسخ درست نداده اند . دانشجویان دارای نمره ۱۸ به سؤال جواب نداده اند و از نمره ۲۰ به بعد نسبت پاسخدهی به سؤال افزایش یافته است . با توجه به سطح زیر منحنی و پلکانی نبودن نمودار، سؤال آسان شناخته می شود (ملاک نمره ۳۰ – ۰ است).

ضریب دشواری آزمون روان شناسی رشد (P=0/65) کمی بیشتر از سطح بهینه بود . ضریب تمیز آزمون (d=0/26) پایین تر از سطح بهینه و واریانس آزمون (V=0/18) قابل قبول تلقی شد . با توجه به رابطه بین ضریب دشواری و ضریب تمیز آزمون و همچنین شاخص های دیگر، آزمون های روانشناسی رشد در حد نسبتاً دشواری با قدرت تمیز پایین ارزیابی شد.

نمودار شماره۲. رابطه بین ضریب دشواری و قدرت تمیز آزمون رشد

با در نظر گرفتن میانگین ضریب دشواری (P=0/65) می توان ادعا کرد ضریب تمیز آزمون بین ٦٢ – ٠ + تا ٠/٦٢ – خواهد بود.

برای ٤٠ سؤال آزمون احساس و ادراک، میانگین کل نمرات دانشجویان برابر ١٣/٩٣ و نمای آزمون نمره ١٤ شد. طراح سؤال، تمایل بیشتری در جای دادن گزینه کلید در قسمت (ج) داشته است. از ٤٠ سؤال مناسب تشخیص داده شد . ضریب همبستگی سؤال های آزمون ٠/١٦ محاسبه شد که در سطح ٥/٠٥ معنی دار بود و ضریب همبستگی بین دو گروه همتا ٠/٩٧ محاسبه شد.

ضریب دشواری آزمون احساس و ادراک (P=0/72) و خارج از سطح بهینه بود . ضریب تمیز (d=0/31) خارج از سطح بهینه بود . واریانس آزمون قابل قبول تلقی گردید. با توجه به رابطه بین ضریب دشواری و ضریب تمیز آزمون و همچنین شاخص های دیگر، آزمون احساس و ادراک در حد مناسب ارزیابی شد . برای ٢٩ سؤال آزمون انگیزش و هیجان، میانگین کل ٤١/١١ و نمای ١٤ بدست آمد. طراح سؤال، گزینه کلید را با درصد بالاتری در قسمت ج جای داده است (٣٧ درصد). از ٢٩ سؤال آزمون، ٤ سؤال نا مناسب، ٣ سؤال قابل اصلاح و ٢٢ سؤال مناسب شناسایی گردید . ضریب همبستگی سؤال های آزمون با نمره کل آزمون در سطح ٠/٠١ معنی دار بود ضریب همبستگی بین دو گروه همتا ٠/٩٤ محاسبه گردید. با توجه به منحنی ویژگی سؤال، سؤال شماره ١ آزمون، نسبتاً ساده و قابل استفاده و سؤال شماره ١٣ نسبتاً ساده ارزیابی شد . ضریب دشواری آزمون انگیزش و هیجان (P=0/68) خارج از سطح بهینه و سطح تمیز (d=0/36) خارج از بهینه و واریانس (V=0/18) قابل قبول تعیین گردید. با توجه به رابطه بین ضریب دشواری و ضریب تمیز آزمون و همچنین شاخص های دیگر، آزمون انگیزش و هیجان در سطح مناسب توصیف گردید.

برای سؤال ٣٠ سؤال روانسنجی میانگین کل نمرات ١٤/١٨ و نمای نمرات آزمون ٤ شد. طراح سؤال، تمایل بیشتری در نهادن گزینه کلید در قسمت ب داشته است (٣٣ درصد) از ٣٠سؤال، تعداد ١٧ سؤال مناسب، ٧ سؤال قابل اصلاح و ٦ سؤال مناسب، شناخته شد. با توجه به منحنی ویژگی سؤال، سؤال شماره ١، نسبت پاسخدهی در گروه پایین کم و در گروه بالا زیاد است و این نسبت از نمره ١٢ به بالا افزایش می یابد . سؤال از حیث نظریه صفت مکنون مناسب بود. برای سؤال شماره ٣٠، از نمره ١٣ به بالا پاسخدهی افزایش یافت . از حیث نظریه صفت مکنون، سؤال مطلوب و قابل نگهداری در بانک سؤال است.

ضریب دشواری آزمون روان سنجی (P=0/68) کمی بیشتر از سطح بهینه و ضریب تمیز (d=0/33) نیز پایین تر از سطح بهینه و واریانس آزمون (V=0/18) قابل قبول بود. با توجه به رابطه بین ضرایب و همچنین شاخص های دیگر، آرمون روانشناسی در حد نسبتاً آسان و قدرت تمیز نسبتاً پایین ارزیابی گردید.

نمودار شماره ٣. محل آزمون ها روی سطوح بهینه و خارج از بهینه ضریب دشواری

نمودار شماره ٤. محل آزمون ها روی سطوح بهینه و خارج از بهینه

نمودار شماره ٥. محل آزمون ها روی سطوح بهینه و خارج از بهینه

بحث و نتیجه گیری

برای تهیه و تدوین سؤال چهار گزینه ای به آموزش و مهارت نیاز هست. در طرح هر سؤال باید قواعد آزمون سازی رعایت گردد و طراح با توجه به هد فهای آموزشی و طبقات چندگانه شناختی، سؤال خود را تدوین نماید و دقت لازم را در تهیه گزینه های انحرافی بکار ببرد. آزمون های معلم ساخته، پس از اجرا نیازمند تحلیل و بازنگری هستند تا طراح از میان روایی و پایایی و شاخص های آماری آزمون خودآگاه گردد.

درس سنجش و اندازه گیری در گرایش های رشته علوم تربیتی ارایه می شود و دانشجویان کارشناسی ارشد نیز در این گرایش ها، درس ارزشیابی آموزشی را می گذرانند و با شیوه های تهیه، اجرا و استاندارد کردن آزمون ها آشنا می شوند و بقیه دانشجویان و همچنین استادان غیر علوم تربیتی از این شیوه ها نا آگاه هستند. از این رو رشته های دیگو به گونه ای تجربی به تهیه آزمون چهار گزینه ای می پردازند و پس از امتحان نمره را اعلام می کنند و گاهی اوقات نیز، همان سؤال ها را برای نیم سالهای بعد مورد استفاده قرار می دهند.

در این میان تلاشها و زحمات دانشجویان با تعدادی سؤال غیر استاندارد ارزشیابی می شود و بین دانش آموزان تمیز و تشخیص درست بر اساس میزان یادگیری و پیشرفت درسی آنان به عمل نمی آید و از برگذاری امتحانات متمرکز در یک دوره حدود ۱۰ روزه، نتیجه درستی عاید نمی گردد.

برای نتایج این پژوهش آزمون های احساس و ادراک، انگیزش و هیجان، روان سنجی و همچنین روان شناسی رشد از سطح بهینه دشواری دور هستند و به سمت سادگی سوق یافته اند . فرضیه پژوهش مبتنی بر ضریب دشواری مناسب آزمون ها مورد تأیید قرار گرفت.

با توجه به درصد بالای سؤال های نا مناسب، در آزمون ها، روایی و پایایی آنها برابر فرضیه پژوهش مورد تأیید قرار نگرفته و بر خلاف انتظار آزمون روان سنجی روایی و پایایی بالاتری از آزمون های دیگر نداشت . در حالی که طراح آزمون که سنجش و اندازه گیری را تدریس می کند آزمونی برگزار کرده که از کیفیت لازم برخوردار نیست. آزمون های چهارگانه از لحاظ ضر یب تمیز نیز خارج از سطح بهینه هستند و نمی توانند به خوبی دو گروه دانشجویان قوی و ضعیف را از هم جدا سازند. فرضیه دیگر پژوهش نیز مبتنی بر ضریب تمیز پایین آزمون ها تأیید گردید. آزمون ها از لحاظ واریانس در سطح قابل قبولی قرار دارند و فرضیه پژوهش در این زمینه تأیید نگردید. چنانچه سؤال های قابل اصلاح آزمون ها، مورد بررسی قرار گیرند و سؤال های بهتری جایگزین سؤال های نا مناسب شوند می توان آزمون ها را به سوی کیفیت بهتر برای ارزشیابی علمی توان آموزشی دانشجویان هدایت کرد.

تحلیل سؤال های آزمون بر اساس صفت م کنون تا حدودی با تحلیل کلاسیک سؤال ها همخوانی دارد و دانشجویان دارای نمرات بالاتر، بهتر توانسته اند به سؤال ها پاسخ درست بدهند.

با اینکه چندین دهه از پیدایش روش تحلیل کلاسیک گذشته است هنوز اجرای آن در نظام ارزشیابی متداول نشده است. امروزه با استفاده از نظر یه صفت مکنون به گونه ای بهتر وعلمی تر می توان در مورد سؤال ها و آزمون ها به داوری پرداخت و نسبت به اصلاح آنها و بهبود کیفیت ارزشیابی همت گماشت . استفاده از سؤال و آزمون های مطلوب با روایی و پایایی مناسب، برای اندازه گیری یادگیری و پیشرفت تحصیلی دانشجویان اج تراب ناپذیراست. چون توانایی اعضای علمی گروههای آموزشی در طرح سؤال مطلوب، متفاوت است، پشنهاد می شود که از افراد قوی و آموزش دیده، در طرح سؤال های امتحانی استفاده بیشتری شود و برای افراد علاقه مند کارگاههای ارزشیابی و تهیه و اجرا و تحلیل آزمون دایر گردد.

اجرای پژوهش در سطح وسیع تر و در گروههای دیگر آموزشی با استفاده از نرم افزارهای کامپیوتری و استفاده از نظریه سؤال-پاسخ در تحلیل آزمون های اجرا شده، اطلاعات علمی ارزشمند و معتبری در اختیار طراحان سؤال و استادان قرار دهد.

منابع

- بازرگان، عباس. (۱۳۸۱). ***ارزشیابی آم وزشی.*** تهران: سمت.
- سیف، علی اکبر . (۱۳۸۲). ***روشهای اندازه گیری و ارزشیابی آموزشی.*** تهران: دوران.
- سلیمی زاده، محمد کاظم. (۱۳۷۷). ***آشنایی با تجزیه و تحلیل سؤالات و آزمون ها*** . تهران: سازمان سنجش.
- شریفی، حسن پاشا. (۱۳۸۳). ***اصول روانسنجی و روان آزمایی.*** تهران: رشد.
- گنجی، حمزه. (۱۳۷۱). ***روانشناسی عمومی*** . تهران: دانشگاه پیام نور.
- گنجی، حمزه و ثابت، مهرداد. (۱۳۸۳). ***روانسنجی.*** تهران: ساوالان.
- گلاورز، جی. ای، برونینگ، ار. اچ. (۱۳۷۸). ***روان شناسی تربیتی*** . ترجمه علینقی خرازی . تهران: مرکز نشر دانشگاهی.
- نفیسی، غلامرضا و زند پارس، علی حسن. (۱۳۷٦). ***سنجش و ارزشیابی*** . تهران: دانشگاه آزاد اسلامی واحد جنوب.
- Bryman. A. (1991). ***Quantitative data analysis***, Rout ledge. London.
- Burney. D. H. (1998). ***Research Methods***. Brooksycole Publishing Company Pacific Grove.
- Craighead. W. E. (2004). ***The Concies Corsin Encyclopedia of psychology and behavioral science***. Chien R. J. Sweden. M. E. Phillips.
- S. M (1996). ***Psychological testing and assessment,*** may feet publishing Company. California.
- http://Pareonline.Net. (2009).***Basic item analysis for multiple choice tests.*** Practical assessment Research and Evaluation Education. Springer Publishing Company.
- http://redie.Uabc. (2009)***. The level of difficulty and discrimination power of the basic knowledge and***

skills Examination. (exhcoba.Vol.2.No.l.)

- Raja Isaiah Rasiah. (2009). Relationship ***between Item Difficulty and Discrimination indices in True/False Type Multiple choice Questions of a Para-Clinical.*** Multidisciplinary paper.Vol.35No.2.
- Salkind, N. (2008). ***Encyclopedia of Educational psychology.***
- Wakefield. J. F. (1996). ***Educational psychology***. Houghton M. Elfin Company. Boston.
- William. W et al. (1990). ***Educational Measurement and Testing.*** Ally and Bacon. Boston.

Quarterly Journal of Educational Psychology
Islamic Azad University Tonekabon Branch
Vol. 1, No. 1, autumn 2009, No 1

The Scientific analysis of tests Based on Classic theory and latent trait theory

Ahmadvand. Mohammad Ali*[1]
1, 2) Assistant professor. Islamic Azad University. Tehran South Branch
* Corresponding author: Ahmadvandma@yahoo.com

Abstract
Scientific test analysis is most popular means for the evaluation of student at universities. In order to study the quality of evalition at teacher training faculty, 4fields of study. Namely psychology was randomly chosen. Students answer sheets were analyzed in those 4 teats. The aim of study was to check the reliability and validity, difficulty index and discrimination index of the tests. The hypothesis of research was that statistic indicates of reliability and validity is outside the optimized level in 4tests. The finding of research indicated that between 25 till 57% of the items related to4tests are unsuitable and the difficulty index, discriminative index, validity and reliability of the tests are outside the optimized level.
Key words: Evaluation, difficulty index, discrimination index, latent trait theory.