آزمون برازش توزیع لجستیک در تعیین توانایی و رتبه‌بندی در امتحان‌ها

کمال عقیق[1]

چکیده

در این مقاله ابتدا با استفاده از توزیع لجستیک سه پارامتری در پاسخ‌گویی به پرسش‌های چهارگزینه‌ای در امتحانات ورودی، میزان توانایی هر یک از داوطلبان برای هر پرسش در نتیجه هر امتحان با روش نظریه‌ی پرسش- پاسخ (IRT) مورد اندازه‌گیری قرار گرفته و بر اساس توانایی نمره او تعیین شده و مبتنی بر این اطلاعات رتبه بندی داوطلبان انجام پذیرفته است. همچنین با استفاده از روش متداول پرسش‌های چهار گزینه‌ای، توزیع های تجربی و نظری نیز محاسبه و رتبه بندی شده است. دو نتیجه حاصل مورد قیاس قرار گرفته و با استفاده از آزمون نکویی برازش تناظر بین رتبه بندی‌ها مورد مطالعه و بحث قرار گرفته‌اند. براین اساس در یکی از آزمون‌های سازمان سنجش آموزش کشور که با استفاده از نرم افزارهای کامپیوتری تهیه شده است، توانایی هر یک از امتحان دهندگان برآورد شد و برای هر آزمودنی مقدار توانایی، نمره بر اساس توانایی، نمره خام (روش کلاسیک)، رتبه بندی براساس نمره‌های توانایی و براساس نمره‌های خام، و تفاوت دو روش رتبه بندی بررسی شده است .

کلمات کلیدی: نظریه پرسش- پاسخ، IRT ، رتبه بندی، توزیع لجستیک، توانایی

۱– مقدمه

شکوفایی و گسترش فن‌آوری کامپیوتر و توسعه روزافزون کاربرد آن در روان‌سنجی سبب شده است تا برخی از محدودیت‌های نظریه‌ی کلاسیک آشکارتر و برجسته‌تر شود. کاربرد نظریه‌ی کلاسیک در موقعیت‌های سنتی آزمون‌گری، اعم از گروهی یا فردی بسیار مناسب است. در این موقعیت‌ها است که در عمل نفس رتبه‌بندی اهمیت پیدا می‌کند و از تعیین توانایی افراد غفلت می‌شود. وقتی از توانایی در زمینه‌های علمی صحبت می‌کنیم، می‌توانیم از اصطلاحات توصیف کننده‌ای همچون توانایی خواندن یا توانایی ریاضی استفاده کنیم. هر یک از این اصطلاحات درست به همان چیزی اشاره دارد که متخصصان روان سنجی از آن با عنوان صفت نامشهود[1] یا صفت مکنون[2] یاد می‌کنند.

مقاله در تاریخ ۸۷/۹/۲۵ دریافت و در تاریخ ۸۷/۱۰/۲۳ به تصویب نهایی رسید.

[1] **، مهندسی عمران دانشگاه خواجه نصیرالدین طوسی

با وجود این که متخصصان روان سنجی چنین اصطلاحی را به راحتی توصیف می‌کنند و افراد مطلع می‌توانند خواص آن را برشمارند، اما اندازه‌گیری این مقولات مفهومی، به آسانی اندازه‌گیری خصوصیات فیزیکی و جسمانی نیست. یکی از هدف‌های اولیه اندازه‌گیری‌های روانی و تربیتی این است که تعیین کند افراد چه مقداری از این صفت‌ها یا توانایی‌ها را دارند.

از اوایل دهه ۱۹۵۰ مدل هایی از اندازه‌گیری‌های روانی عرضه شدند که بر این تنگناهای نظریه‌ی کلاسیک انگشت گذاردند. امروزه رایج ترین و پخته‌ترین این مدل‌ها، مجموعه‌ای را تشکیل می‌دهند که به شناخت و تعیین ویژگی‌ها و مشخصه‌های ریاضی پاسخ های امتحان دهندگان به آزمون کمک می‌کند. این مدل‌ها به عنوان مدل‌های **نظریه‌ی پرسش- پاسخ یا مدل‌های IRT** شناخته شده‌اند. نظریه‌ی پرسش - پاسخ رابطه مشخصه‌ها یا پارامترهای هر پرسش و ویژگی‌ها یا توانایی افراد (صفت

مکنون) را با احتمال ارایه پاسخ درست بررسی می‌کند. فرض بر آن است که مقدار این صفت (توانایی) همیشه در امتداد پیوستاری تک بعدی[3] تغییر می‌پذیرد که به طور معمول آن را با حرف θ نشان می‌دهند. جای پرسش j روی محور θ که به طور معمول با b_j نشان داده می‌شود، به دشواری پرسش تعبیر می‌شود. تمام مدل های IRT احتمال پاسخ درست به پرسشی از آزمون‌های چند گزینه‌ای با یک پاسخ درست را به عنوان تابعی از θ مشروط به یک یا چند پارامتر پرسش نشان می‌دهند. برای هر پرسش می‌توان احتمال دادن پاسخ درست یا موافقت با طبقه ویژه‌ای از پاسخ را روی نمودار نشان داد. این تابع‌ها معرف **رگرسیون** غیرخطی احتمال وقوع پاسخ درست هستند. در این تابع‌ها احتمال وقوع پاسخ درست برحسب وجود مقدار معینی از یک توانایی یا یک صفت مکنون نظیر هشیاری یا توانایی کلامی در نزد پاسخ دهنده تعریف می‌شود.

احتمال وقوع هر پاسخ به عنوان تابعی از پارامترها و به طور جداگانه برای هر پرسش و مشخصه‌های هر شخص، مدل‌سازی می‌شود. پارامترهای پرسش معرف خواصی از پرسش است که آن را به عنوان میزان دشواری وتوانایی تشخیص پرسش می‌شناسیم و مشخصه‌های شخص معرف میزان توانایی آزمون شونده است. اگر احتمال وقوع پاسخ به عنوان تابعی از مشخصه‌های (توانایی) فرد معرفی شود، تابع مورد نظر در IRT به عنوان تابع پاسخ شناخته می‌شود.

در این مقاله ابتدا به تحلیل و بررسی مدل نظریه پرسش – پاسخ[4] (IRT) مبتنی بر مدل لجستیک سه پارامتری می‌پردازیم. در این تحلیل: برآورد پارامترهای هر یک از پرسش‌ها (دشواری b , تشخیص a و حدس c)، برآورد احتمال ارایه پاسخ درست در هر یک از سطوح توانایی به آنها به وسیله نرم‌افزار ترسیم منحنی مشخصه‌های سؤال‌ها و نمودار تابع اطلاع آنها، برآورد توانایی هر یک از امتحان دهندگان، مقایسه نتیجه‌ها بر اساس نمره‌های خام و نمره‌های مبتنی بر توانایی با استفاده از **آزمون کای اسکور** خواهیم پرداخت و برای مثال، یکی از آزمون‌های تخصصی زبان انگلیسی سازمان سنجش آموزش کشور که به صورت آزمون‌های چهار گزینه‌ای است مورد بررسی قرار گرفته است.

۲– تعاریف، اصطلاح‌ها و مفاهیم نظری

۲–۱ تابع لجستیک به عنوان مبنای نظریه‌ی پرسش– پاسخ

اگر بتوان با متغیر تصادفی x مشخصه‌های رفتار یک سیستم را در مقابل یک عمل یا آزمایش بیان کرد، قانون توزیع x، توزیعی است که به نام توزیع لجستیک خوانده می‌شود. بنابراین مدل استاندارد ریاضی برای نظریه‌ی پرسش – پاسخ و برای منحنی مشخصه‌های پرسش، **مدل تابع لجستیک**[5] است. نظریه‌ی پرسش- پاسخ با فرض وجود رابطه‌ای ریاضی بین توانایی‌ها (یا دیگر صفت‌های مفروض) و احتمال جواب گویی به سؤال‌ها، به مطالعه و بررسی نمره‌های پرسش و امتحان می‌پردازد [۱و۵].

۲–۲ مدل‌های یک، دو و سه پارامتری

وقتی در تحلیل هر پرسش برآورد دو مشخصه آن یعنی میزان دشواری و تشخیص آن مورد نظر باشد مدل دو پارامتری با تابع توزیع احتمال زیر را انتخاب می‌کنیم.

$$p(\theta)=\frac{1}{1+e^{-a(\theta-b)}}$$

اگر در معادله بالا مقدار a را برابر ۱ قرار دهیم، مدل بالا به مدل لجستیک یک پارامتری تبدیل خواهد شد.

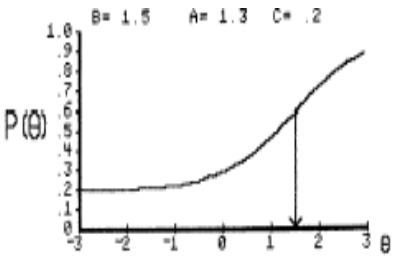
برین باوم[6] با اندکی دستکاری در تابع لجستیک، مدل سه پارامتری را ارایه داد [۲و۳]. وقتی که عامل شانس در پاسخ گویی دخیل باشد، مدلی سه پارامتری خواهیم داشت. برای مدل سه پارامتری می‌توان از فرمول زیر استفاده کرد:

$$p(\theta)=c+(1-c)\frac{1}{1+e^{-a(\theta-b)}}$$

یعنی در این مورد تنها ضریب حدس ($0<c<1$) وارد می‌شود که به طور معمول حدس بالای ۰/۳۵قابل قبول نیست.

در معادله بالا θ نمایان‌گر متغیر توانایی و a نمایان‌گر پارامتر تشخیص و b نمایشگر پارامتر دشواری پرسش است. واضح است که پارامترهای a و b برای هر پرسش ثابت‌اند و دامنه تغییرات آن‌ها از $-\infty$ تا $+\infty$ است ولی در عمل $-۲/۸۰ < a < ۲/۸۰$ و $-۳ < b < ۳$ و در نظر گرفته می‌شود (شکل ۱). b نقطه‌ای روی محور توانایی است که

احتمال پاسخ‌گویی متناظر با آن ۰/۵است. باید توجه داشت که بر اساس مفروضات نظریه‌ی پرسش – پاسخ، متغیر توانایی از پارامترهای تشخیص و دشواری پرسش مستقل است.



شکل ۱ منحنی مشخصه های پرسش برای مدل سه پارامتری

با توجه به آنچه گفته شد در واقع می‌توان برای توانایی هر امتحان دهنده یک مقدار عددی یا یک نمره منظور کرد که جای او را روی محور توانایی مشخص می‌کند. این مقدار توانایی را با حرف θ و احتمال این که امتحان دهنده با این سطح توانایی به این پرسش پاسخ درست بدهد را با $p(\theta)$ نشان می‌دهیم.

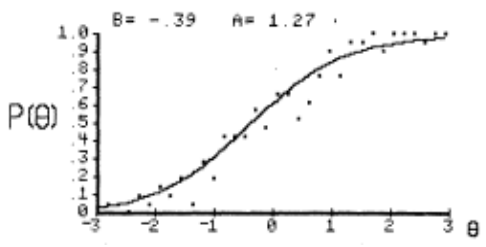
۲–۳ برآورد پارامترهای یک پرسش

فرض کنیم در آزمونی، یک نمونه مرکب از n تا امتحان شونده به یک پرسش موجود در امتحان پاسخ داده‌اند. سطح توانایی این امتحان دهندگان روی محور توانایی توزیع شده است. این امتحان دهندگان را به t گروه توانایی در طول محور توانایی تقسیم می‌کنیم. فرض کنیم n_i امتحان دهنده در گروه i ، $i=1,2,3,...,t$ ، با سطح توانایی یکسان θ_i قرار دارند و فرض کنیم r_i امتحان دهنده از این گروه به پرسش، پاسخ درست داده‌اند. بنابراین برآورد احتمال پاسخ های درست در سطح توانایی θ_i، با محاسبه نسبت پاسخ‌های درست مشاهده شده به کل پاسخ‌ها در همین زیرگروه امکان پذیر است.

$$p(\theta_i)=\frac{r_i}{n_i}$$

کار اساسی ما این است که منحنی مشخصه‌های پرسش را که بهترین **برازش**[۷] برای نسبت پاسخ‌های مشاهده شده است، به دست آوریم. فرایند برازش منحنی، با استفاده از روش برآورد حداکثر درست نمایی[۸] انجام می‌گیرد.

شکل ۲ یک منحنی مشخصه پرسش را نشان می‌دهد که با نسبت پاسخ‌های درست مشاهده شده برازش شده است.



شکل ۲ منحنی مشخصه پرسش

نتیجه‌ی حاصل آن است که در مدل IRT پارامترهای a و b محاسبه شده در هر زیر گروه از توانایی، همیشه یکسان خواهد بود. این در حالی است که اندیس دشواری پرسش در مدل کلاسیک از یک زیر گروه به زیر گروه دیگر متغیر است. به همین علت، تعبیر و تفسیر دشواری پرسش آن طور که در نظریه‌ی IRT تعریف شده، ساده تر است.

در مدل لجستیک سه پارامتری، بهترین روش برآورد پارامترهای پرسش، برآورد حداکثر درست نمایی کناری[۹] است. در این روش برآورد پارامترها از طریق انتگرال گیری از تابع درست نمایی روی توزیع توانایی به دست می‌آید. یعنی:

$$L(\beta|x)=\prod_i\int p(x_i|\theta,\beta)f(\theta)d\theta$$

که $f(\theta)$تابع چگالی توانایی است. برآوردهای حداکثر درست نمایی کناری، مقدارهایی از β هستند که معادله بالا را ماکسیمم می‌کند.

اگر پارامترهای پرسش در مدل IRT معلوم باشند آنگاه برآورد توانایی برای نمونه‌ای از امتحان دهندگان با استفاده از روش برآورد حداکثر درست نمایی[۱۰] آسان‌تر و صریح‌تر خواهد بود. در این مقاله خواهیم دید که چگونه می‌توان میزان توانایی یا θ را با استفاده از برآورد پارامترهای پرسش که در جریان **کالیبره** کردن پرسش‌ها حاصل می‌شود، برآورد کرد. برآورد هم‌زمان توانایی و پارامترهای پرسش در مدل‌های متفاوت (اعم از یک، دو یا سه پارامتری) کاری دشوار و پر درد سراست. البته تهیه نرم افزارهای کامپیوتری

یا نرم افزارهای موجود، این کار را به نسبت آسان ساخته است.

۴-۲ برآورد توانایی امتحان دهنده

در نظریه پرسش- پاسخ، هدف اولیه ارایه یک آزمون به امتحان دهنده این است که جای او روی مقیاس توانایی مشخص شود.

برای اندازه‌گیری توانایی، مبتنی بر تعدادی پرسش (N) خواهد بود که هر کدام جنبه‌ای از آن توانایی را اندازه می‌گیرد. در بحث پارامترهای پرسش و برآورد آن‌ها فرض می شود که پارامتر توانایی امتحان دهندگان معلوم است. بر عکس برای برآورد (پارامتر) توانایی نامعلوم یک امتحان دهنده، فرض بر این خواهد بود که مقدارهای عددی پارامترهای پرسش‌های آزمون معلوم‌اند. وقتی امتحانی اجرا می‌شود، هر امتحان دهنده به هر یک از پرسش‌های آن پاسخی می‌دهد و به پاسخ‌ها به صورت دو وجهی[11] نمره داده می‌شود. پس برای هر یک از پرسش‌های آزمون، نتیجه نمره‌ای برابر با یک یا صفر خواهد بود. در نظریه‌ی پرسش- پاسخ برای برآورد توانایی امتحان دهنده، از روش حداکثر درست نمایی استفاده می‌شود بعد این فرایند برای هر یک از امتحان دهندگانی که در آزمون شرکت داشته‌اند به طور جداگانه تکرار می‌شود. به هر حال این روش مبتنی بر رویکردی است که با هر امتحان دهنده، جداگانه برخورد می‌کند. از این رو، موضوع اصلی آن این است که چطور می شود توانایی یک فرد واحد را برآورد کرد. معادله‌ی برآورد در فرایند بر آورد حداکثر درست نمایی به صورت زیر است:

$$\hat{\theta}_{s+1}=\hat{\theta}_s+\frac{\sum_{i=1}^{N}-a_i[u_i-p_i(\hat{\theta}_s)]}{\sum_{i=1}^{N}a_i^2[p_i(\hat{\theta}_s)Q(\hat{\theta}_s)]}$$

که در آن $\hat{\theta}_s$ توانایی برآورد شده برای امتحان دهنده با تکرار s است، a_i پارامتر تشخیص پرسش i برای $i=1,2,3,\ldots,k$ است، $u_i=1$ برای پاسخ درست به پرسش i و $u_i=0$ برای پاسخ نادرست به پرسش i است، $p_i(\hat{\theta}_s)$ احتمال پاسخ درست به پرسش i در مدل منحنی مشخصه‌های پرسش داده شده در سطح توانایی $\hat{\theta}$ با تکرار s است، $Q_i(\hat{\theta}_s)=1-p_i(\hat{\theta}_s)$ احتمال پاسخ نادرست به پرسش i در مدل منحنی مشخصه‌های پرسش داده شده در سطح توانایی $\hat{\theta}$ با تکرار s است.

خطای استاندارد برآورد را نیز می‌توانیم طبق فرمول زیر محاسبه کنیم.

$$SE(\hat{\theta})=\frac{1}{\sqrt{\sum_{i=1}^{N}a_i^2[p_i(\hat{\theta}_s)Q(\hat{\theta}_s)]}}$$

مرحله‌های برآورد حداکثر درست نمایی در حالتی که امتحان دهنده به همه پرسش‌ها پاسخ غلط و یا در حالتی که امتحان دهنده به همه پرسش‌ها پاسخ درست داده باشد قابل محاسبه نیست. یعنی برآورد توانایی در حالت نخست $-\infty$ و در حالت دوم $+\infty$ می‌شود.

۵-۲ مربع خی

در تحلیل‌های IRT مسأله مهم این است که آیا یک منحنی مشخصه پرسش معین با داده‌های مشاهده شده، خوب برازش[12] شده است یا نه؟ روش‌های متفاوتی برای آزمون نکویی برازش پیشنهاد شده است. برای بررسی نکویی برازش از آزمون آماری مجذور خی[13] استفاده شد.

اگر مقدار به دست آمده از یک مقدار معین بزرگ‌تر باشد، منحنی مشخصه پرسش که با برآورد پارامترهای پرسش به دست آمده، برازنده داده‌ها نیست. این امر ممکن است به دو علت اتفاق بیافتد. نخست این که مدل نادرستی برای ترسیم و تحلیل منحنی مشخصه‌های پرسش انتخاب شده باشد. دوم این که مقدار احتمال‌های مشاهده شده پاسخ‌های درست آن قدر متفرق باشند که نتوان منحنی برازش شده‌ای را برای مدل مورد انتظار به دست آورد.

تحت مدل مشخص IRT فراوانی مورد انتظار انتخاب k توسط پاسخ دهندگان با استفاده از فرمول زیر محاسبه می‌شود.

$$E_i(k)=N\int p\left(\nu_i=k|\theta=t\right)f(t)dt$$

$f(t)$ در این فرمول عبارت است از چگالی توانایی که به طور معمول استاندارد نرمال فرض می‌شود زیرا تابع های پرسش- پاسخ بر مبنای این توزیع مقیاس سازی می‌شود. **فون** در **والنبرگ**[14] نشان داده است که آماره مربع خی برای هر یک از تک پرسش‌ها در بسیاری از موارد به فرض

تک بعدی بودن آزمون، حساس نیست. برای اجتناب از این مسأله پیشنهاد شده است که مجذور خی برای پرسش‌های در دسته‌های دوتایی و سه‌تایی محاسبه شود. در صورتی که این دسته پرسش‌ها از نابرازندگی مشابهی[15] برخوردار باشند، مقدارهای بزرگی از مربع خی را نشان خواهند داد [۴].

۳- تحلیل نتیجه‌های آزمون تولیمو

۳-۱ تحلیل پرسش‌های آزمون تولیمو

براین اساس آنچه تا اینجا گفته شد، برای تمام پرسش‌های آزمون تولیمو ۱۲، پارامترهای پرسش‌ها (یعنی دشواری b، تشخیص a و حدس c) برآورد شد. چون پرسش‌ها از نوع چند گزینه‌ای بودند، امکان پاسخ‌گویی به پرسش‌ها از راه حدس وجود داشت، بنابراین، پارامتر حدس نیز برای تعیین احتمال پاسخ‌گویی درست به سؤال وقتی که دانش یا توانایی در کار نیست، برآورد شد. منحنی مشخصه‌های پرسش‌ها نیز براساس مقدارهای برآورده شده‌ی پارامترها رسم شد. برای این کار به دلیل حجم بالای عملیات و این که هیچ‌گاه نمی‌توان این محاسبات را با دست انجام داد از نرم‌افزارهای زیادی استفاده شد که مهم‌ترین آن‌ها نرم‌افزاری بود که از دانشگاه ایلینویز آمریکا خریداری شد و برای پیشبرد کار نرم افزارهای واسطه زیادی نوشته و مورد استفاده قرار گرفت. سرانجام پرسش‌ها بر اساس مشابهت مقدار پارامترهای آن‌ها و نیز بر اساس مقایسه منحنی‌های مشخصه پرسش‌ها، به ۷ گروه زیر تقسیم شدند:

جدول ۱ طبقه بندی پرسش‌های آزمون تولیمو

گروه	
اول	این گروه که مشتمل بر ۵۰ پرسش است دارای ویژگی‌های زیر است: پارامتر دشواری (b=۳)، پارامتر تشخیص بالاتر از یک (a<۱)، و پارامتر حدس کمتر از ۰/۲ (c<۰/۲). همان‌طور که پارامتر دشواری و منحنی مشخصه‌های پرسش‌ها نشان داد، پرسش‌های موجود در این گروه، پرسش‌هایی بسیار دشوار هستند به گونه‌ای که احتمال ارایه پاسخ درست در بالاترین سطح توانایی یعنی ($\theta = ۳$) فقط ۰/۵ بوده است و این احتمال برای افراد با توانایی متوسط و پایین هم بسیار ناچیز بوده است. با این پرسش ها تنها تا حدودی می‌توان تواناترین افراد را از سایرین تشخیص داد و می‌توان گفت که پرسش‌ها اصلاً قادر به شناسایی افراد با توانایی متوسط و پایین و تمیز آن‌ها از یکدیگر نیست.
دوم	این گروه که مشتمل بر ۱۳ پرسش می باشد دارای ویژگی‌های زیر است: پارامتر دشواری b =۳ ، پارامتر تشخیص بالاتر از یک (a >۱) و پارامتر حدس بالاتر از ۰/۲ (c>۰/۲). در این گروه ۱۳ سؤال جای گرفته است که همانند پرسش‌های گروه قبلی بسیار دشوارند، طوری که احتمال ارایه پاسخ درست به این پرسش‌ها از سوی افراد با سطح توانایی بالا ($\theta = ۳$) تنها ۰/۵ بوده است و افراد با توانایی متوسط و پایین چندان قادر به پاسخ‌گویی درست به پرسش‌ها نبودند. البته با این تفاوت که در این دسته از پرسش‌ها احتمال پاسخ‌گویی درست با تکیه بر شانس برای آن‌ها بیشتر بوده یا می‌توان گفت در حد معقولی بوده است. با این گروه از پرسش‌ها نیز تنها تا حدودی می‌توان تواناترین افراد را مشخص کرد.
سوم	این گروه که مشتمل بر ۱۵ پرسش می‌باشد دارای خصوصیات زیر است: پارامتر دشواری (b=۳)، پارامتر تشخیص بین ۰/۵ تا یک (۰/۵<a<۱)، پارامتر حدس کمتر از ۰/۱۵ (c< ۰/۱۵). این گروه از پرسش‌ها شامل ۱۵ پرسش می‌شود. با توجه به منحنی مشخصه‌های پرسش‌ها می‌توان گفت از سطح توانایی یک (θ〉۱) به بالا، به موازات افزایش توانایی احتمال دادن پاسخ درست به این پرسش‌ها هم افزایش می‌یابد. اما برای افرادی که روی مقیاس توانایی در سطح متوسط و پایین قرار گرفته‌اند احتمال ارایه پاسخ درست به پرسش‌ها بسیار پایین بوده است و با توجه به دشواری پرسش‌ها می‌توان گفت که این احتمال حتی برای تواناترین افراد (($\theta = ۳$)) هم فقط ۰/۵بوده است. در نتیجه با این گروه از

	پرسش‌ها هم تنها تا حدودی می‌توان توانایی‌های بالاتر از متوسط و نیز تواناترین افراد را در زمینه دانش زبان شناسایی کرد.
چهارم	۳۴ پرسش در این گروه قرار گرفته که ۱۵ پرسش دارای پارامتر دشواری b=۳ , ۱۴ پرسش پارامتر دشواری بین ۲ تا ۳ (۳ >b> ۲)، و ۵ پرسش هم پارامتر دشواری کمتر از ۲ (۲ >b) دارند. منحنی مشخص‌های پرسش‌های این گروه شیب بسیار ملایمی دارد زیرا تمام پرسش‌ها پارامتر تشخیص پایینی دارند و در نتیجه پرسش‌ها از قدرت تمیز کافی برخوردار نیستند و نمی‌تواند افرادی را که در سطح توانایی بالا قرار گرفته‌اند از افرادی که در سطح توانایی پایین قرار دارند به صراحت تفکیک کنند.
پنجم	این گروه که مشتمل بر ۱۹ پرسش است دارای ویژگی‌های زیر است: دشواری (۲/۵ >b>۱) و تشخیص بالاتر از ۰/۵(۰/۵ ⟩a). همان‌گونه که منحنی مشخصه‌های پرسش‌ها نشان می‌دهد از سطح توانایی یک به بالا (۱⟩θ)، احتمال پاسخ‌گویی درست به پرسش‌ها، روند افزایشی داشته است، یعنی با بالا رفتن توانایی، احتمال پاسخ‌گویی درست نیز افزایش یافته است. بنابراین این گروه از پرسش‌ها به خوبی توانسته است افرادی را که در زمینه دانش زبان دارای توانایی بالاتر از متوسط بوده‌اند، شناسایی کند. اما در سطوح توانایی متوسط و پایین‌تر از آن قادر به تفکیک و تشخیص تفاوت ها نبوده است.
ششم	سه پرسش در این گروه جای دارد با پارامتر دشواری بین ۰/۵ تا ۱ (۱ >b> ۰/۵)، پارامتر تشخیص بزرگ‌تر از ۰/۵ (۰/۵ >a) و پارامتر حدس بزرگ‌تر از ۰/۲. با توجه به منحنی مشخصه‌های پرسش‌ها از سطح توانایی متوسط به بالا، احتمال پاسخ‌گویی درست به پرسش‌ها رو به افزایش بوده است و با افزایش توانایی، احتمال پاسخ‌گویی درست نیز افزایش یافته است. پارامتر تشخیص پرسش‌ها نیز مطلوب است. با این گروه می‌توان افراد با توانایی متوسط و بالا را در زمینه دانش زبان به خوبی مشخص نمود. اما این پرسش‌ها قادر به تفکیک افراد در سطوح پایین‌تر از متوسط نمی‌باشد.
هفتم	پنج پرسش در این گروه جای دارد با پارامتر دشواری کمتر از ۰/۵ (۰/۵>b)، پارامتر تشخیص بین ۰/۵ تا ۱ (۱ >a> ۰/۵) است. با توجه به پارامترهای دشواری این پرسش‌ها می‌توان گفت که احتمال پاسخ‌گویی درست به این پرسش‌ها برای کسانی که در سطوح پایین توانایی قرار داشته‌اند کمتر بوده و هر چقدر توانایی افزایش یافته این احتمال نیز افزایش می‌یافته است. البته این موضوع تا سطح توانایی ۲ (۲ = θ) صادق بوده و احتمال پاسخ‌گویی درست برای سطوح متفاوت توانایی بالاتر از آن یکسان بوده است. به عبارت دیگر پارامتر تشخیص پرسش‌ها نیز تا حد زیادی مناسب است یعنی این گروه از پرسش‌ها تا سطح توانایی (۲ = θ) به خوبی افراد را از یکدیگر متمایز می‌کند، اما از این سطح توانایی به بالا دیگر خوب عمل نمی‌کند. در واقع این گروه از پرسش‌ها نمی‌تواند تواناترین افراد را از افراد توانا تفکیک کند. اما برای سنجش و تمیز سطوح توانایی ضعیف تا بالاتر از متوسط کارایی دارد.

حال که مشخصات هر یک از گروه‌های پرسش‌ها را بیان کردیم، این مسأله مطرح می‌شود که کدام یک از انواع پرسش‌ها برای سنجش توانایی یا دانش زبان آزمودنی‌ها مناسب‌تر است. اگر هدف از اجرای آزمون تولیمو سازمان سنجش، توانایی دانش زبان و تعیین سطوح مختلف توانایی افراد باشد. اولاً: دشواری پرسش‌های آزمون باید به گونه‌ای باشند که احتمال اندکی از پاسخ‌گویی درست به پرسش برای افرادی با سطح توانایی پایین نیز وجود داشته باشد و در تمام موارد با بالا رفتن سطح توانایی، احتمال پاسخ گویی درست نیز افزایش یابد. در پرسش‌های چهار گزینه‌ای مقدار این احتمال برای سطوح پایین توانایی چیزی در حد و حدود احتمال پاسخ‌گویی تصادفی یا پاسخ‌گویی درست از روی شانس است. در واقع پرسش‌ها باید از دشواری متوسطی برخوردار باشند. ثانیاً: پرسش‌ها باید از پارامتر تشخیص مناسبی نیز برخوردار باشد تا هم بتواند افرادی با

سطح بالای توانایی را از افرادی با سطح پایین توانایی صراحتاً متمایز کند و هم سطوح مختلف بین این دو را شناسایی کند.

با مفروض دانستن هدف بالا، پرسش‌های گروه اول و دوم، پرسش‌های مناسبی نیست، چون مقدار پارامتر دشواری پرسش‌های این دو گروه خیلی بالاست و احتمال پاسخ گویی درست از سوی تواناترین افراد تنها ۰/۵ بوده است و احتمال پاسخ‌گویی درست از سوی افرادی با توانایی متوسط و پایین نیز بسیار ناچیز بوده است. در نتیجه این دو گروه برای تعیین سطوح متفاوت توانایی دانش زبان مناسب نیستند و *با این دو گروه از پرسش‌ها می‌توان تا حدی فقط تواناترین افراد را در زمینه دانش زبان* مشخص نمود.

با فرض بالا، پرسش‌های گروه سوم و چهارم نیز مناسب نیست، چون پارامتر دشواری بالایی دارند و احتمال پاسخ‌گویی درست تنها برای افراد برخوردار از سطح توانایی بالاتر از یک ($\theta\rangle ۱$)، به مقدار قابل توجهی بوده و از این سطح به بعد رو به افزایش است. افرادی با توانایی متوسط و پایین هم به سختی و با احتمال بسیار اندکی توانسته‌اند به این گروه از پرسش‌ها پاسخ درست بدهند. *پس با این گروه پرسش‌ها نیز فقط می‌توان تواناترین افراد را مشخص کرد.*

پرسش‌های گروه پنجم نیز برای تعیین سطوح متفاوت توانایی دانش زبان مناسب نیستند چون افرادی که در سطوح متوسط و پایین توانایی قرار گرفته‌اند ($\theta < ۱$)با احتمال بسیار ناچیزی توانسته‌اند به این گروه از پرسش‌ها پاسخ درست بدهند. با این گروه از پرسش‌ها تنها می‌توان اشخاصی را که در سطح بالای توانایی دانش زبان قرار گرفته‌اند، شناسایی کرد.

پرسش‌های گروه ششم نیز برای نیل به هدف مذکور مناسب نیست زیرا اشخاصی را که در مقیاس توانایی در سطوح پایین‌تر از متوسط ($\theta\rangle ۰$) قرار گرفته‌اند نمی‌توانند مشخص کنند *با این گروه تنها می‌توان افرادی را که در مقیاس توانایی در سطوح متوسط و بالا هستند مشخص کرد.* با این وجود می‌توان ادعا کرد که پرسش‌های این گروه و گروه بعدی مناسب‌ترین پرسش‌های این آزمون بوده‌اند.

پرسش‌های گروه هفتم با توجه به داشتن پارامتر تشخیص مناسب و پارامتر دشواری پایین به تقریب تا سطوح توانایی بالا یعنی ($\theta = ۲$) خیلی خوب عمل کرده‌اند و به خوبی افراد را تفکیک کرده‌اند، اما قادر به تمایز گذاری بین تواناترین افراد و افراد توانا نبوده اند. همان طور که در مورد گروه قبلی گفته شد به رغم این مسایل می‌توان گفت به لحاظ مقایسه‌ای این گروه از پرسش‌ها را می‌توان مناسب‌ترین پرسش‌های این آزمون قلمداد کرد.

آنچه که تاکنون ذکر شد در مورد مناسب بودن پرسش‌ها با فرض این موضوع بود که هدف از آزمون برگزار شده، سنجش تمام سطوح متفاوت توانایی دانش زبان بوده است.

حال اگر هدف از اجرای آزمون شناسایی و تعیین افراد با توانایی متوسط و بالا در زمینه دانش زبان باشد، پرسش‌های آزمون باید به گونه‌ای باشند که از سطوح توانایی متوسط به بالا افراد قادر به پاسخ‌گویی درست به پرسش‌ها باشند و با افزایش سطح توانایی، احتمال پاسخ‌گویی درست نیز افزایش یابد. در واقع دشواری پرسش‌ها باید متناسب با سطوح توانایی متوسط و بالا باشد تا بتوان افرادی را که در این سطوح قرار دارند مشخص کرد. افزون بر این پرسش‌ها باید از پارامتر تشخیص مناسبی نیز برخوردار باشند براین اساس پرسش‌های گروه اول و دوم و سوم مناسب نیستند، زیرا همان‌طور که گفته شد، مقدار پارامتر دشواری پرسش‌های این سه گروه بالاست و احتمال پاسخ‌گویی درست برای تواناترین افراد فقط ۰/۵ بوده است و مقدار این احتمال برای افرادی با توانایی متوسط هم بسیار اندک بوده است. بنابراین با این پرسش‌ها تا حدودی می‌توان فقط تواناترین افراد را در زمینه دانش زبان مشخص کرد.

۳-۲ برآورد توانایی آزمودنی‌ها در آزمون تولیمو

متأسفانه هیچ راهی برای از پیش دانستن پارامتر واقعی توانایی وجود ندارد و بهترین راه تنها برآورد آن است. به لحاظ منطقی می‌توان گفت مقدار متوسط برآوردهایی که به وسیله کامپیوتر محاسبه می‌شود به مقدار پارامتر توانایی آزمودنی‌ها نزدیک است. وقتی پارامتر دشواری پرسش‌ها برابر یا نزدیک به پارامتر توانایی آزمودنی باشد, میانگین توانایی برآورد شده نزدیک به مقدار توانایی وی است. نکته حائز اهمیت در برآورد توانایی آزمودنی‌ها در نظریه‌ی پرسش ـ پاسخ این است که توانایی آزمودنی نسبت به

پرسش‌ها ثابت است یعنی نسبت به پرسش‌هایی که برای سنجش توانایی مورد نظر به کار می‌روند نامتغیر است. این بدین معناست که از یک آزمون با هر جایگاهی که در سرتاسر مقیاس توانایی داشته باشد (یعنی با هر میزانی از دشواری) می‌توان برای برآورد توانایی آزمودنی استفاده کرد . براین اساس توانایی هر یک ازامتحان دهندگان آزمون تولیمو با استفاده از نرم افزارهای تهیه شده برآورد شد. برای هر آزمودنی مقدار توانایی، نمره براساس توانایی، نمره خام (روش کلاسیک)، رتبه بندی براساس نمرات توانایی و براساس نمرات خام و تفاوت دو روش رتبه بندی ارایه شده است.

در آزمون تولیمو ۷۸۴ آزمودنی شرکت کرده‌اند که دامنه توانایی آزمودنی‌ها بین ۳/۳۵۸- تا ۴/۶۶۵+ است. توزیع فراوانی توانایی آزمودنی‌ها به این قرار است که، ۹۰ آزمودنی توانایی بالاتر از ۳+ , ۴۰۷ آزمودنی توانایی بین ۲+ تا ۳+ , ۲۴۶ آزمودنی توانایی بین ۱+ تا ۲+ , ۳۱ آزمودنی توانایی بین صفر تا ۱+ و ۱۰ آزمودنی توانایی زیر صفر دارند.

۴– نتیجه گیری

۴–۱ برازش مدل و داده ها

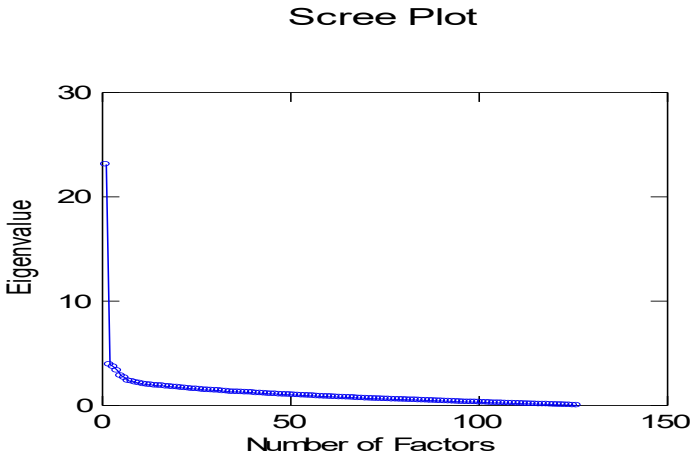
همان‌طور که در پیش گفته شد مدل انتخابی برای تحلیل آزمون و برآورد توانایی‌ها، مدل سه پارامتری بوده است. مقدارهای حاصل از محاسبه مربع خی در اکثر موارد از عدد معیار بزرگ‌تر نبود. با این حال این دو موضوع مورد بررسی قرار گرفت، یکی موضوع تک بعدی یا چند بعدی بودن مدل و دیگری مسأله نامتغیر بودن پارامترها.

۴–۲ ابعاد آزمون

اغلب مدل‌های اندازه گیری این فرض را دستمایه کار خود قرار می‌دهند که مفهومی که داریم اندازه می‌گیریم یک بعدی است، یعنی تنها یک عامل برجسته است که مبنای توضیح آن رفتار خاص دارد. رویکردهای متفاوتی برای واررسی و ارزیابی تک بعدی بودن مدل پیشنهاد شده است. روش‌های رایج عبارت است از تعیین تعداد مقدارهای ویژه‌ای[16] که بزرگ‌تر است از ۱، بررسی نمودار **اسکری**[17]، (شکل ۳) و مقایسه نسبت نخستین مقدار ویژه به دومین مقدار ویژه. در این تحلیل ما از نمودار اسکری و مقایسه نسبت نخستین مقدار ویژه با دومین مقدار ویژه استفاده کرده‌ایم که بر مبنای تکنیک تحلیل عاملی به بررسی ابعاد آزمون می‌پردازد. البته در اینجا برای تعیین تعداد عامل‌های دخیل در مقیاس، به جای روش تحلیل عنصرهای اصلی[18] (PCA)، از روش عامل یابی محور اصلی[19] (PAF) استفاده کرده ایم. زیرا در روش تحلیل عامل یابی محور اصلی فقط واریانس مشترک در محاسبات در نظر گرفته می‌شود در حالی که در روش تحلیل عنصرهای اصلی هم واریانس اشتراکی و هم واریانس اختصاصی در محاسبات دخیل است.

همان‌طور که در پیش گفته شد پرسش‌های آزمون تولیمو دو انتخابی هستند یعنی پاسخ آن‌ها از دو حالت درست یا غلط بیرون نیست. درباره این گونه داده‌های دو انتخابی (درست یا غلط) لازم است بدانیم که تحلیل عاملی روی همبستگی تتراکوریک پاسخ‌ها انجام می‌شود. از بین نرم‌افزارهای معروف برای انجام این محاسبه از نرم‌افزار SYSTAT ۱۰/۲ استفاده شد که در محاسبه هم‌زمان همبستگی‌های تتراکوریک و عامل یابی محور اصلی کارایی چشمگیری دارد. ابتدا مقدارهای نخستین مقدار ویژه (۲۳/۱۰۸) تا بیستمین مقدار ارایه شده بررسی شد. با مشاهده نخستین مقدار ویژه می‌توان دریافت که تفاوت آن با مقدارهای بعدی بسیار فاحش است.

در قسمت بعدی نیز سهم هر یک از عامل‌ها در واریانس همبستگی‌ها دیده شد که سهم عامل نخست نسبت به سایر عامل‌ها بسیار بیشتر است. همچنین نمودار اسکری نیز نشان می‌دهد که عامل‌های دوم به بعد در جایی از نمودار قرار گرفته‌اند که نمودار به حالت تخت نزدیک شده است در حالی که عامل اول با فاصله بسیاری در بالای نمودار قرار گرفته است. بنابراین می‌توان از تک بعدی بودن آزمون و مقیاس مطمئن بود.



شکل ۳ نمودار اسکری

از آنجایی که بررسی ابعاد مدل نشان داد مقیاس ما یک بعدی است ناچار باید بررسی شود که آیا پارامترهای پرسش‌ها و توانایی آزمودنی‌ها نسبت به هم نامتغیر هستند یا نه. برای این کار گروه چهارم از پرسش‌ها که ۳۴ پرسش بودند با توجه به هدف آزمون پرسش‌های مناسبی نبودند حذف شد. یعنی این پرسش‌ها از قدرت تمیز کافی برخوردار بوده و نمی‌توانستد در سطوح متفاوت توانایی آزمودنی‌ها را به صراحت تفکیک کنند. پس از حذف این پرسش‌ها دوباره برای هر یک از آزمودنی‌ها بر آورد توانایی بر اساس روش حداکثر درست نمایی، محاسبه نمره خام و نمره براساس نظریه IRT و رتبه‌بندی آزمودنی‌ها با توجه به ۱۰۶ پرسش باقی‌مانده صورت گرفت. پس از آن نتیجه‌های این مرحله و مرحله پیش مورد مقایسه قرار گرفتند تا مشخص شود حذف پرسش‌ها چه تأثیری در برآورد توانایی آزمودنی‌ها دارد. بر این اساس میانگین تفاوت توانایی‌های برآورد شده با احتمال ۹۹ درصد در فاصله ۰/۱۰۵۷ – تا ۰/۲۷۵۶ قرار گرفته‌اند. بررسی نتیجه‌ها نشان می‌دهد که در اکثریت موارد تفاوت بین توانایی‌های بر آورد شده بسیار ناچیز است و در فاصله برآورد شده قرار گرفته‌اند و میزان جا به جایی افراد در مقیاس توانایی با حذف پرسش‌ها قابل اغماض است. این بیان‌گر نامتغیر بودن توانایی آزمودنی‌ها نسبت به پرسش‌هایی است که برای برآورد توانایی به کار می‌روند.

برای بررسی استقلال پارامترهای پرسش‌ها از نمونه، ۱۰۰ آزمودنی به طور تصادفی از میان ۷۸۴ آزمودنی انتخاب شد. محاسبه پارامترهای پرسش‌ها تفاوت معنی‌داری را نشان نمی‌داد.

تقدیر و تشکر

با تشکر از سازمان سنجش و آموزش کشور که با پشتیبانی و حمایت آن سازمان این پژوهش انجام گرفته است.

۵– پی نوشت

[1] nobservable trait
[2] latent trait
[3] unidimensional
[4] Item-Response Theory
[5] logistic function
[6] A. Birnhaum
[7] fitting
[8] maximum likelihood estimation
[9] Marginal Maximum Likelihood estimation
[10] Maximum Likelihood estimation
[11] dichotomously
[12] goodness-of-fit
[13] chi-square(chi-square goodness-of-fit index)
[14] Van der Wollenberg
[15] similar misfits
[16] Eigenvalues
[17] Scree plot
[18] Principal Components Analysis
[19] Principle Axis Factoring

منابع

[1] Baker F. B.**,** *The basics of item response theory*, **2001**.

[2] Baker F.B., *Item response theory: Parameter estimation techniques*, NewYork: Marcel Dekker, **1992**.

[3] Birnhaum A., *Some latent trait models and their use in inferring an examinee's ability*. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing, **1968**.

[4] Drasgow F., Levine M.V., Tsien S., Williams B.A. and Mead A.D., " *Fitting polytomous item response theory models to multiple-choice tests*", Applied Psychological Measurement, Vol.19, **1995**, pp. 143-165.

[5] Hulin C.L., Drasgow F and Parsons C.K., *Item response theory: Applications to psychological measurement*, Homewood, IL:Dow Jones, **1983**.