طراحی و پیاده‌سازی بانک سؤال مدرج شده در آزمون های سراسری[1]

سلیمان ذوالفقارنسب[2]

تاریخ دریافت: ۱۳۸۹/۰۹/۰۸
تاریخ پذیرش: ۱۳۹۱/۰۲/۱۵

چکیده

نوشته حاضر فرایند ایجاد یا توسعه یک بانک سؤال مدرج شده بر اساس تئوری سؤال پاسخ با استفاده از داده های ۳۰ سؤال کنکور آزمایشی حساب دیفرانسیل است که روی یک نمونه ۳۴۰۹ نفری اجرا شده است. تابع آگاهی کل سؤالات برابر با ۴۳/۲۱ است که در فاصله ± 3 پیوستار توانایی نقطه بیشینه آن روی ۰/۹۵ است و در این نقطه، توانایی افراد با حداقل خطا برآورد شده است. یکی از نکات برجسته در راه اندازی چنین بانکی برآورد توانایی افراد بر اساس آزمون‌های مداد-کاغذی استاندارد و کم خطا با اهداف ویژه است تا بتوان در هزینه، تلاش و زمان صرفه جویی کرد و امنیت آزمون ها را بالا برد. بعلاوه فرایند آزمون گیری را که به صورت سنتی و مداد کاغذی اجرا می شود با طی مراحل بعدی به صورت سنجش انطباقی کامپیوتری ارتقاء داد. برای ارائه مثال از بین سؤالات این بانک یک خرده آزمون کوتاه مدادکاغذی با سؤالاتی که قدرت تمیز بالا و ضریب دشواری بیشتری دارند، انتخاب شده است. دامنه دشواری این خرده آزمون کوتاه بین $1 \le b_i \le 2$ و ضریب تمیز $2 \le a_i$ قرار داده شد و در نهایت ۵ سؤال برگزیده شده است که تابع آگاهی این خرده آزمون برابر با ۱۲/۴۸ و نقطه بیشینه آن روی پیوستار توانایی برابر با ۱/۳۵ است. بعلاوه سطح دشواری آزمون۰/۴ واحد انحراف معیار افزایش یافته و به علت کوتاه شدن دامنه آزمون از ۳۰ سؤال به ۵ سؤال میزان

---

[1]. برگرفته از پژوهشی با همین نام در سازمان سنجش آموزش کشور، مرکز تحقیقات، ارزشیابی، اعتبارسنجی و تضمین کیفیت آموزش عالی سال ۱۳۸۹

۲. کارشناس پژوهشی سازمان سنجش آموزش کشور salarnik2001@yahoo.com

آگاهی دهی از توانایی آزمون‌دهندگان نیز ۳/۴۶ برابر کاهش پیدا کرده است. بعلاوه برای اجرای سنجش انطباقی یک شبیه سازی انجام شده است که تنها ۸ تا ۱۰سؤال نیاز بوده تا با ۰/۰۱ احتمال خطا به همان دقت آزمون۳۰ سؤالی توانایی افراد را اندازه‌گیری کرد.

واژگان کلیدی:
بانک سؤال، تئوری سؤال پاسخ، آزمون های کامپیوتری، گنجینه سؤال، سنجش انطباقی، بانک تست

مقدمه

یکی از بزرگ ترین مزیت‌های تئوری سؤال پاسخ[۱](IRT) این بوده است که روش‌های آن این امکان را به راحتی در اختیار سازمان‌ها و مراکز آزمون‌گیری قرار داده تا بتوانند سؤالات خود را در یک بانک سؤال ذخیره کرده و آزمون هایی با پارامترهای دلخواه برای اهداف ویژه اجرا کنند و هم‌زمان بتوانند از یک بانک سؤال انواع آزمون های موازی و معادل تهیه کنند.

به طور معمول یک بانک سؤال مجموعه‌ای از سؤالات مقیاس‌های تک بعدی است که پارامترهای سؤالات آن بر اساس یکی از مدلهایIRT به دست آمده است و از لحاظ حیطه موضوعی، سطح آموزش و شناسنامه توصیفی سؤال تفکیک شده است و هدف از راه اندازی آن علاوه بر بالابردن امنیت برای جلوگیری از لو رفتن سؤالات، ساخت آزمون های مدادکاغذی با ویژگی‌های معین و اجرای کامپیوتری آزمون‌ها و بدست آوردن بهترین برآورد از توانایی آزمون دهندگان با کمترین خطا است(گرون‌لند، ۱۹۹۸).

در یک گنجینه سؤال غنی به راحتی می‌توان آزمونی را طراحی کرد که با احتمال بسیار بالایی روی توزیع صفت مکنون جامعه آزمون دهندگان منطبق باشد و یا در اجرای آزمون های ملاک مرجع سؤالاتی را انتخاب کرد که دقیقاً روی یک نقطه از پیوستار توانایی مورد نظر بتواند آزمون‌دهندگان را برای اعطای گواهی یا درجه به دو نیمه تفکیک کند.

بعلاوه با بکار بردن روزافزون سنجش انطباقی کامپیوتری[۲] (CAT) که در آن موقعیت اجرای آزمون برای هر آزمون دهنده و نوع آزمون یکسان است با این تفاوت که سؤالات مشابه یکدیگر نیستند، داشتن گنجینه ای بزرگ از سؤالات مدرج شده امکان انعطاف بیشتری در برگزاری این آزمون‌های استاندارد شده به سازمان‌ها می‌دهد. تهیه و توسعه گنجینه سؤالات در سنجش انطباقی نیز مستلزم تخصیص و جدا کردن ابعاد داده‌ها قبل از برازندگی مدل هایIRT است. در سنجش انطباقی هدف به حداقل رساندن خطای اندازه گیری در برآورد توانایی افراد با کمترین تعداد سؤالات است.

---

۱. Item Response Theory

۲. Computerized Adaptive Testing

۱-گام های ضروری برای طراحی بانک سؤال برای آزمون های خطی و سنجش انطباقی

۱-۱-تهیه آزمون اولیه و کنترل کیفی

برنامه‌های سنجشی استاندارد در مقیاس بزرگ چه مدادکاغذی و چه کامپیوتری مستلزم فرایند طراحی سؤالات، تهیه و کنترل کلید پاسخ سؤالات، سرهم کردن سؤالات، بازبینی محتوای سؤالات برای اطمینان یابی از اینکه مجموعه سؤالات تمام حیطه مورد نظر را تحت پوشش قرار می دهد و در واقع آزمون از روایی مناسبی برخوردار است و اجرای تجربی سؤالات روی یک گروه با حجم بهینه از آزمون دهندگان می باشد. در واقع هر مرحله باید به عنوان یک مرحله کنترل کیفی مدنظر قرار گیرد. به طور ایده‌آل، گام‌های مربوط به کنترل کیفیت باید در ابتدای اجرا، نمره گذاری و گزارش آزمون آغاز شود. تلاش‌های جدی برای کنترل کیفیت یک نقش حیاتی در کمک به تضمین کارکردهای مناسب سؤال‌ها و آزمون‌ها ایفا می‌کند. کمپیون و مایلر (۲۰۰۶) بیان کرده اند که نیاز به فرایندهای کنترل کیفی موثر به عنوان یاوری برای اطمینان‌یابی از روایی آزمون است.

۱-۲- برآورد پارامتر سؤالات و توانایی: برازندگی مدل به داده ها

تعداد سؤالات و نوع سؤالات نقش بسیار مهمی در فرایند انتخاب مدل آماری مناسب برای بدست آوردن توانایی آزمون دهندگان و برآورد پارامتر سؤالات دارد. عمدتاً در آزمون گیری هایی که مبتنی برIRT هستند برای احتمال پاسخگویی فرد به سؤالاتی که به صورت ۰(پاسخ نادرست) و ۱(پاسخ درست) نمره گذاری می شوند یکی از مدل های یک، دو و یا سه پارامتری برای برآورد توانایی به کار می رود. در آزمون های سراسری به علت چهارگزینه ای بودن سؤالات مدل سه پارامتری بهترین برازش با ماهیت داده ها دارد. در این مدل احتمال پاسخ صحیح یک فرد j به یک سؤال i تابعی است از توانایی او θ_j و سه پارامتر سؤال $a_i\,,\ b_i\,,\ c_i$.

$$\mathrm{P}(u_i=1\mid\theta_j,a_i\,,\ b_i\,,\ c_i)=c_i+\left(1-c_i\right)\ /\ [1+\exp(-1.7a_i(\theta_j-b_i)]$$

معادله(۱)

که پارامتر ها عبارتند از شیب a_i، دشواری b_i، و حدس پذیری c_i سؤال. همچنین در این آزمون ها چون ممکن است تعدادی از آزمون‌دهندگان به تمام سؤالات و یا به هیچ کدام از

سؤالات پاسخ ندهند مدل های بیزین[۱] به علت مقاومتی که در برابر از دست دادن اطلاعات دارند[۲] برای مدرج کردن سؤالات بانک سؤال نسبت به الگوهای آماری دیگر مثل بیشینه درست‌نمایی مشترک[۳](JML) و بیشینه درست‌نمایی کناری[۴](MML) و یا بیشینه درست-نمایی شرطی[۵](CML) برتری دارند (به ایگن۲۰۰۴، بیکر و سوک هو کیم۲۰۰۴، نگاه کنید). شکل کلی معادله بیز در ادامه آمده است.

$$P(\theta_j|u_j,\tau,\xi)=\frac{P(u_j|\theta_j,\xi)g(\theta_j|\tau)}{\int P(u_j|\theta_j,\xi)g(\theta_j|\tau)d\theta_j}$$ معادله(۲)

در این معادله علایم عبارتند از محور احتمال پاسخ فرد u_j، توانایی θ_j، پارامترهای سؤال ξ وتابع توانایی در توزیع جمعیت τ.

آزمون‌های متکی بر کامپیوتر چگونه کار می‌کنند؟

سنجش انطباقی کامپیوتری یکی از روش‌های پیشرفته اندازه‌گیری است که در آن کامپیوتر بر اساس سطح توانایی آزمون‌دهنده سؤالات را از گنجینه سؤالات مدرج شده انتخاب و ارائه می‌کند. در CAT هر فرد به آزمونی منحصر به فرد جواب می‌دهد که با سطح توانایی‌های فردی او همخوان شده است. سؤالاتی با مقدار آگاهی کم دربارهٔ توانایی آزمون‌دهنده به طور نرمال حذف می‌شود و به همین دلیل به چنین آزمون‌هایی انطباقی یا سنجش برازش یافته گفته می‌شود.

برای اجرای CAT چهار چیز نیاز است.

۱- بانک سؤالی که بر اساس یکی از مدل های IRT طراحی شده

۲- معیار مشخص برای انتخاب سؤال

۳- یک الگو برای نمره‌گذاری سؤال

۴- تصمیم‌گیری برای پایان زمان آزمون (گرین، بوک، مامفریز، لین و روکاس، ۱۹۸۴).

---

۱. Bayes

۲. loss of information

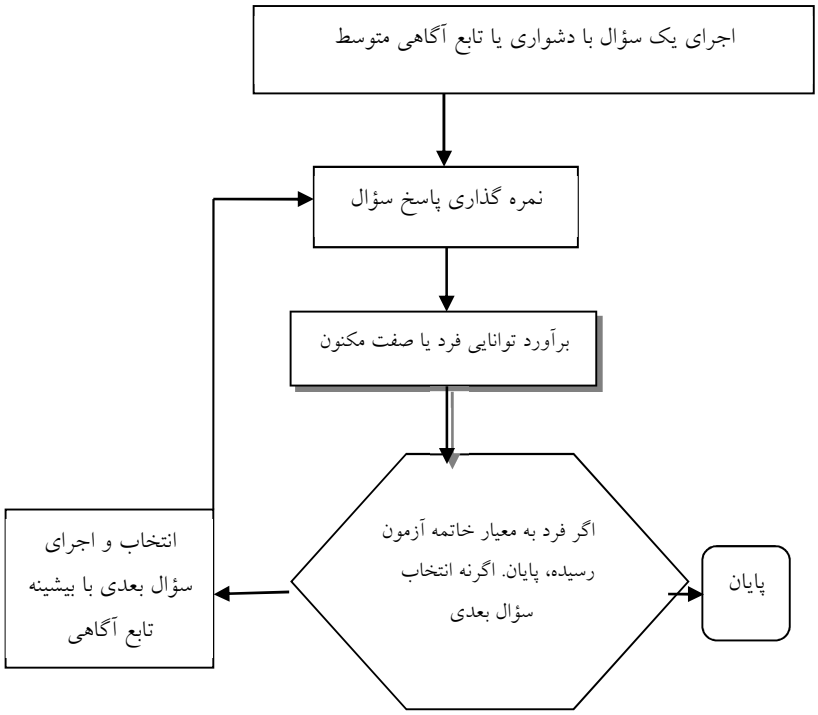
۳. Joint Maximum Likelihood

۴. Marginal Maximum Likelihood

۵. Conditional Maximum Likelihood

در سنجش انطباقی فرایند اجرای آزمون به گونه ای است کـه در آن بـر اسـاس پاسـخ گـویی آزمون دهنده به سؤالات قبل، سؤالات بعدی اجرا می‌شود. این فرایند از طریق بـه کـار بـردن الگوریتمی است که در آن بهترین سؤال بعدی، بـر اسـاس توانـایی بـرآورد شـدهٔ قبلـی فـرد انتخاب می‌شود تا وقتی که زمان آزمون پایان یابد و یا محتـوای مـورد نظـر در آزمـون مـورد پوشش قرار گیرد.

نمودار شماره ۱. الگوریتم سنجش انطباقی کامپیوتری



یک نکته قابل توجه در CAT این است که کامپیوتر تقلیدی از یک آزمون‌گیرندهٔ خردمنـد را نمایش می‌دهد. به ویژه زمانی که آزمون گیرنده سؤالی را می‌پرسد که برای آزمون‌دهنده خیلی مشکل است سؤال بعدی به گونه‌ای است که به طور قابل ملاحظه آسان است. این امر از ایـن حقیقت ناشی می‌شود که اگر سؤالاتی ارائه شوند که بسیار سـخت و یـا بسـیار آسـان باشـند اطلاعات خیلی کمی از توانایی تک‌تک آزمون دهنده گان به دست می آیـد. امـا هنگـامی کـه

سؤالات همراه با افزایش سطح توانایی افراد ارائـه شـود آگـاهی بیشـتری مـی‌تـوان از میـزان توانایی افراد به دست آورد(راینز، ۱۹۸۷).

در نمودار۱ چهارچوب کارکردی CAT نشان داده شده است. نتیجهٔ چنین رویکردی افـزایش دقت اندازه‌گیـری در محـدودهٔ گسـترده‌ای از سـطوح توانـایی اسـت (کارسـلون، ۱۹۹۴). در حقیقت CAT به منظورکاهش زمان و کنار زدن آزمون‌های سـنتی و ناکارآمـدی ایجـاد شـده است که سؤالات آسان را به افراد با توانایی بالا و سؤالات سـخت را بـه آزمـون‌دهنـدگان بـا توانایی کم ارائه می‌دهد(دانکل، ۱۹۹۹).

نتایج و داده ها

نوشتن سؤال و تهیهٔ آزمون گام اولیه در ایجاد بانک سؤال است. داده های به کـار بـرده شـده برای این بانک تنها در برگیرنده ۳۰ سؤال چهارگزینـه‌ای حسـاب دیفرانسـیل بـا احتمـال $\frac{۱}{۴}$ پاسخ صحیح و $\frac{۳}{۴}$ پاسخ خطا برای هر سؤال بوده که به صـورت آزمایشـی روی یـک گـروه ۳۴۰۹ نفری اجرا شده است. برای ایجاد بانک سؤال ابتدا باید سؤالات آزمون به صورت پیش تجربی اجرا شوند. سؤالات اجرا نشده در بانک را نیز می توان بر اساس اصول آزمون سـازی بر مبنای سؤالات لنگر مدرج و در بانک ذخیره کرد. بعد از جمع‌آوری داده ها برای شناسـایی ابعاد آزمون و شناسایی صفت مکنون از نرم‌افزارMSP (سیجت سما و مولن آیر،۲۰۰۲) برای مشخص کردن ابعاد آزمون استفاده شده است. نتایج نشان داد که آزمون دو ُبعد برجسـته دارد که در اولین ُبعد با سطح آستانه مقیاس پذیری پیش‌گزیده ۰/۳۰ تعداد ۲۴ سؤال وارد شده اند. ضریب مقیاسیH آنها برابر با ۰/۳۹ بدست آمده که نشان دهنده تک ُبعدی بودن بهینه این ۲۴ سؤال است و سطح معناداری آن برابر با z= ۱۱/۶۰ که نشان‌دهنـده سـطح مقیـاس پـذیریH بیشتر سؤالات و الگوی تکنوا افزایی[۱] آنها است. البته این تنها در سطح تبیین آماری و شهودی می باشد و سازندگان آزمون باید در سطح نظری هم سـؤالات آزمـون را از لحـاظ محتـوایی تفکیک کنند تا بتوانند ابعاد آزمون ها را هم از جنبه شهودی و هم از جنبه محتـوایی و نظـری تفکیک کنند و سپس هر بعد را جداگانه تحلیل و وارد بانک سؤال کنند. بعلاوه ضریب پایایی

۱. Monotone Homogeneity Model

این ۲۴ سؤال برابر با Rho=۰/۹۰ بوده است. مقیاس دوم تنها از دو سؤال تشکیل شده که عبارتند از سؤال ۱۷ و ۲۸. ضریب مقیاسی این دو سؤال برابر است با ۰/۳۲ وضریب پایایی آن برابر باRho=۰/۳۵ است که شاخص خوبی برای یک مقیاس برای اندازه گیری پیشرفت تحصیلی نمی باشد. دلیل آن را هم می توان کم بودن تعداد سؤالات و کوتاه بودن دامنه این مقیاس بیان کرد و سؤالات ۱، ۱۴، ۱۵ و ۲۲ چیزی به غیر از صفت مکنون مورد نظر در سؤالات قبلی این آزمون اندازه می‌گرفتند و به عبارتی در حیطه ابعاد این آزمون جا نمی گیرند به ویژه سؤال ۲۲ که ضرایب کلاسیک آن مثل درجه دشواری بسیار بالا، ضریب تمیز کم و ضرایب IRT آن مثل شیب و تابع آگاهی آن نسبت به دیگر سؤالات پایین بوده است. در هنگام فرایند مدرج کردن باید چنین سؤالاتی از مجموعه سؤالات کنار گذاشته شوند تا پارامترهای برآورد شده بهینه باشند.

جدول شماره ۱. مقادیر ضریب دشواری، مقیاس پذیریH ، ۲۴سؤال پس از کنار گذاشتن سؤالات بدکارکرد.

تعداد سؤال	شماره سؤال	دشواری سؤال	Item H
1	Item2	0.21	0.33
2	Item3	0.26	0.33
3	Item4	0.39	0.36
4	Item5	0.16	0.30*
5	Item6	0.42	0.44
6	Item7	0.23	0.4
7	Item8	0.38	0.39
8	Item9	0.44	0.46
9	Item10	0.57	0.44
10	Item11	0.49	0.4
11	Item12	0.44	0.41
12	Item13	0.33	0.39
13	Item16	0.21	0.32
14	Item18	0.53	0.4
15	Item19	0.26	0.38
16	Item20	0.18	0.31
17	Item21	0.23	0.39
18	Item23	0.09	0.34
19	Item24	0.25	0.46
20	Item25	0.3	0.37
21	Item26	0.13	0.36
22	Item27	0.13	0.49
23	Item29	0.51	0.38
24	Item30	0.2	0.34

همان طور که می بینیم بر اساس یک توزیع نرمال برای پیدا کـردن ارزش هـای H بزرگتـر از ۰/۳۰ تعداد ۲۴ سؤال در این مقیاس باقی مانده‌اند.

از نرم افزار Bilog-MG (زیموسکی و همکاران،۱۹۹۹) و بر اساس مدل سه پارامتریی IRT برای مدرج کردن این آزمون۳۰ سؤالی استفاده شده است. پارامتر سؤالات بانک وابسته است به مدل انتخاب شدهIRT برای تحلیل داده‌ها و برآورد توانایی افراد. در این تحقیق از روش توزیع پسین مورد انتظار بیز[۱](EAP) برای مدرج کردن سؤالات و همچنین برآورد توانایی آزمون دهندگان استفاده شده است. در این روش یک توزیع نرمال که از روی داده ها برآورد شده است برای برآورد بهینه پارامترها و مدرج کردن سؤالات به عنوان توزیع پیشین در معادله ای که تابع درستنمایی توانایی را محاسبه می کند ضرب می شود تا توزیع پسین مورد انتظار بدست آید. این عمل برای تک‌تک سؤالات و برای تک‌تک آزمون دهندگان صورت می گیرد. نتیجه این عمل برآورد مقادیر ثابتی هم برای پارامترهای سؤالات و هم برای توانایی افراد است (واندر لیندن، ۲۰۱۰: ۱۰-۶). همان طور که اشاره شد، در فرایند مدرج کردن سوالات، بُعدیت آزمون[۲] نقش بسیار حیاتی در برآورد پارامترها دارد و باید فرض استقلال موضعی هم در داده های مربوط به سؤالات و هم در بین آزمودنی‌ها وجود داشته باشد. باید در نظر داشت که مدلی که برای بدست آوردن پارامتر ها به کار می رود با داده ها برازش داشته باشد.در جدول۲ زیر پارامترهای IRT سؤالات آمده است.

جدول شماره ۲. پارامترهای a ،b و c سؤالات همچنین مقدار لوجیت و ضریب همبستگی پیرسون و ضریب دورشته ای

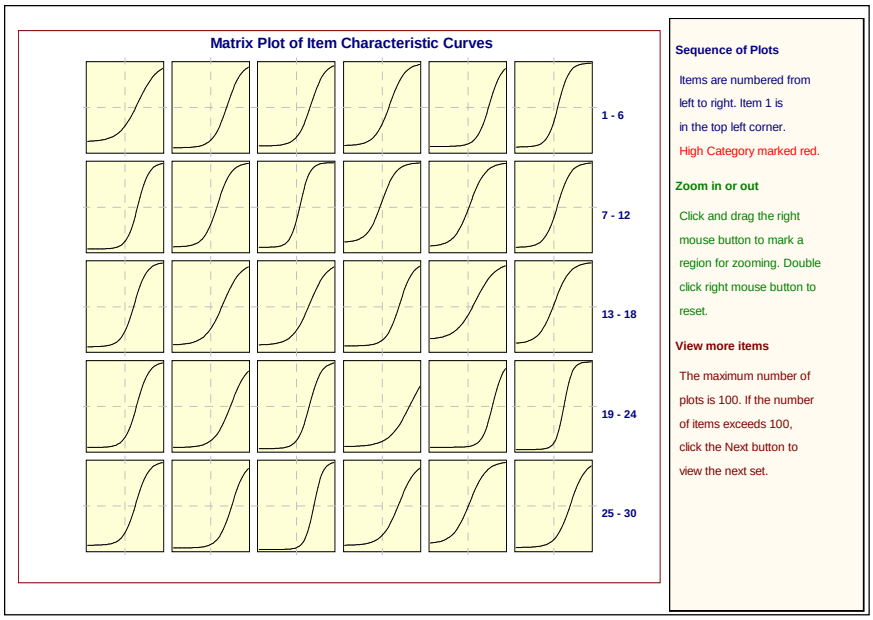
NAME	a	b	c	LOGIT	PEARSON	BISERIAL
Item 1	1.32	1.05	0.12	0.63	0.37	0.477
Item 2	1.755	1.319	0.051	1.3	0.434	0.611
Item 3	1.838	1.155	0.072	1.06	0.454	0.615
Item 4	1.676	0.609	0.075	0.47	0.483	0.615
Item 5	2.028	1.703	0.066	1.69	0.361	0.548
Item 6	2.307	0.366	0.06	0.3	0.572	0.722
Item 7	2.277	1.025	0.033	1.19	0.542	0.749
Item 8	1.859	0.537	0.056	0.47	0.521	0.663
Item 9	2.608	0.289	0.052	0.24	0.603	0.758
Item 10	1.765	-0.048	0.108	-0.27	0.483	0.608

۱. Expectation A Posteriori

۲. test dimensionality

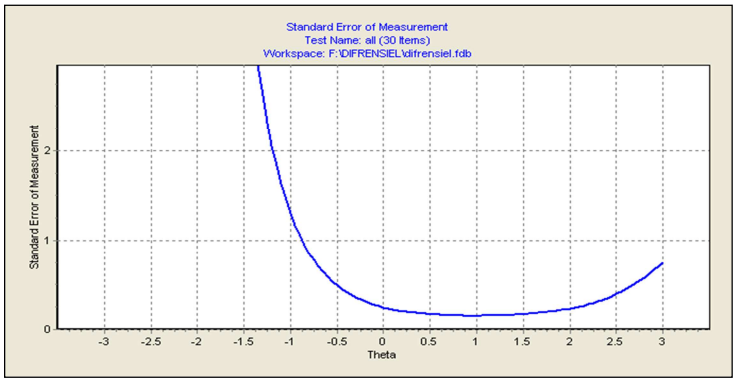
Item 11	1.62	0.152	0.062	0.03	0.482	0.604
Item 12	1.949	0.32	0.051	0.25	0.539	0.679
Item 13	2.017	0.733	0.054	0.72	0.536	0.697
Item 14	1.409	0.992	0.073	0.77	0.42	0.549
Item 15	1.447	1.056	0.073	0.84	0.417	0.549
Item 16	1.881	1.392	0.064	1.36	0.424	0.602
Item 17	1.269	0.611	0.135	0.23	0.387	0.487
Item 18	1.629	0.054	0.088	-0.13	0.48	0.602
Item 19	2.093	0.953	0.041	1.03	0.529	0.714
Item 20	1.512	1.615	0.04	1.54	0.385	0.566
Item 21	2.165	1.038	0.028	1.21	0.534	0.74
Item 22	1.197	2.253	0.051	1.78	0.272	0.421
Item 23	2.38	1.982	0.044	2.27	0.321	0.559
Item 24	3.083	0.818	0.018	1.08	0.638	0.867
Item 25	1.986	0.873	0.062	0.84	0.506	0.667
Item 26	1.853	1.697	0.031	1.87	0.403	0.636
Item 27	2.78	1.389	0.015	1.9	0.528	0.84
Item 28	1.48	1.346	0.061	1.17	0.397	0.547
Item 29	1.514	0.135	0.08	-0.03	0.46	0.576
Item 30	1.721	1.333	0.037	1.37	0.446	0.636

نمودارشماره ۲. منحنی ویژگی یا ICC ۳۰سؤال



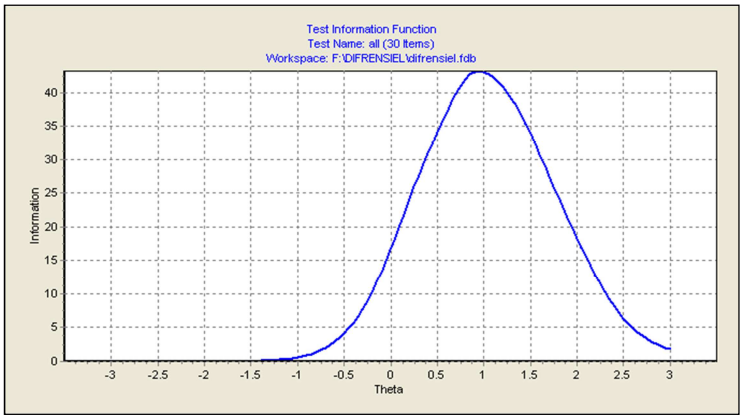
از نرم افزارFastTEST Professional 2.3.0 نیز برای بانک کردن سؤال ها به همراه تمام مشخصه‌های آماری و نوشتاری آنها استفاده شده است. این نرم افزار قابلیت ذخیره کردن تمام اطلاعات سؤالات و آزمون را دارد. در نمودار ۳ می توان مقدار خطای آزمون در بانک سوال کالیبر شده در فاصله‌های مختلف تتا مشاهده کرد.

نمودار شماره ۳. افزایش خطای برآورد تتا در آزمون ۳۰ سؤالی با فاصله گرفتن از نقطه۰/۹۵ صدم تتا



در نمودار۴ تابع آگاهی این آزمون ۳۰ سؤالی در بانک آمده است

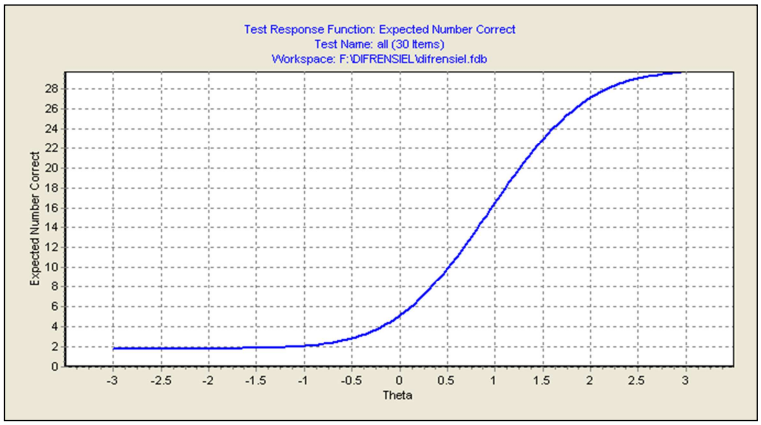
نمودار شماره ۴. تابع آگاهی آزمون ۳۰ سؤالی



همان‌طور که در نمودار۵ می بینیم منحنی ویژگی آزمون ارتباط نزدیکی با تابع آگاهی آزمون دارد. به عبارت دیگر بیشترین شیب آزمون نیز در نقطه ۰/۹۵ تتا است. جایی که قدرت تمیز آزمون در تفکیک افراد به بیشترین مقدار خود می رسد و در این نقطه آزمون با احتمال بالایی افراد قوی و ضعیف را از یکدیگر جدا می کند. خطای برآورد در این نقطه به ۰/۱۵۲۱ می

رسد. انتظار می رود افرادی که توانایی برابر با ۰/۹۵ دارند تقریبا به بیش از ۱۵ سؤال پاسخ دهند: یعنی نصف.

نمودار شماره ۵. ویژگی آزمون ۳۰ سؤالی



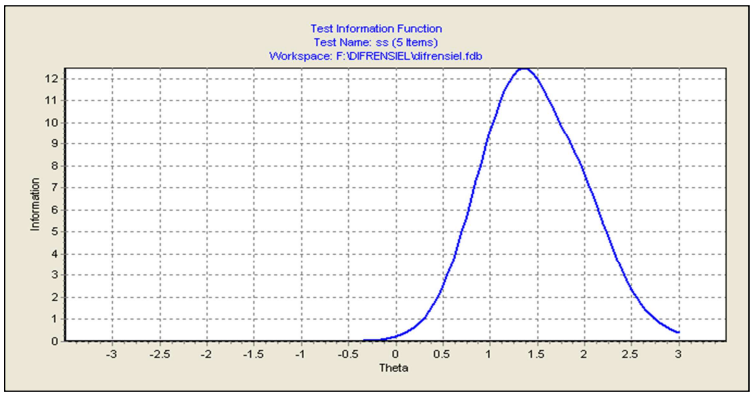
انتخاب سؤال از بانک

معمولاً در یک بانک سؤال مدرج ایده‌آل به اندازه کافی سؤال هست که بتوان فرم‌های چندگانه‌ای از سؤالات ساخت که به کمک آنها بتوان توانایی افراد را اندازه گرفت (راوی و نونیگ، ۲۰۰۲). یک گنجینه خوب و بزرگ کمک می‌کند تا سؤالات بهتری به توانایی افراد برازش داده شود و در نتیجه بتوان توانایی افراد را با دقت بیشتر برآورد کرد. بسته به اهداف آزمون گیری سازمان ها و یا مراکز آموزشی، برای انتخاب یک مجموعه سؤال از بانک رویکردهای متفاوتی وجود دارد. یکی از این روش‌ها که به‌ویژه در سنجش برازش‌یافته کاربرد دارد، انتخاب سؤال بر اساس تابع آگاهی سؤالات است. با توجه به نمودار ۳ و ۴ می‌توان دید که در این گنجینه ۳۰ سؤالی بهترین تابع آگاهی در امتداد پیوستار تتا بین نمره ۰ تا ۲ است که نقطه ماکسیمم آن تقریبا ۰/۹۵ می باشد. این مقدار برابر با ۴۳/۲۱ است. خارج از این محدوده، آگاهی از توانایی آزمون دهندگان بسیار کاهش می‌یابد و نمی‌توان به خوبی سطح توانایی افراد را پیش‌بینی کرد. بنابراین پیوستار تابع آگاهی به ما می گوید که هر سطح از توانایی به چه خوبی اندازه‌گیری می‌شود. به طور کلی یک تابع آگاهی زمانی ایده‌آل است که خط آن در بخش گسترده‌ای از پیوستار توانایی افقی باشد و توانایی را در تمام سطوح بتواند با بیشترین دقت برآورده کند.

باید بدانیم که تابع آگاهی آزمون از مجموع تابع آگاهی تک‌تک سؤالات آزمون به‌دست می آید. این امر کمک می کند تا در آزمون‌گیری‌های متفاوت بسته به اهداف هر آزمون دقیقاً سؤالاتی را از گنجینه سؤالات انتخاب کنیم که بیشترین آگاهی از توانایی آزمون دهنده‌ها را در فاصله دلخواه بدست دهد. صفت مکنون افرادی که در این نقطه از مقیاس توانایی قرار دارند(در نقطه ۰/۹۵ تتا) با کمترین خطا برآورد شده است و هر چه نمرات از این نقطه فاصله داشته باشند احتمال خطای برآورد توانایی افزایش می یابد. نمودار۳ افزایش مقدار خطا را در برآورد توانایی افراد در آزمون ۳۰ سؤالی نشان می دهد.

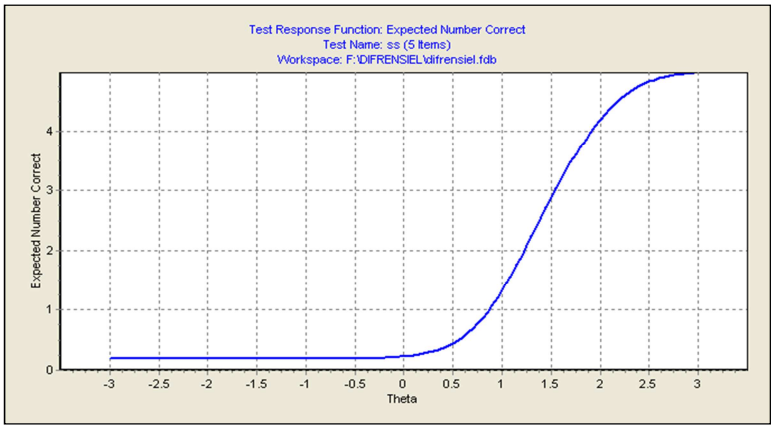
رویکرد دوم انتخاب سؤالاتی است که بهترین شیب را در یک فاصله دلخواه داشته باشند همان طور که می‌دانیم پارامتر شیب سؤال نمایانگر توان سؤال در جداسازی آزمون دهندگان قوی و ضعیف است. رویکرد سوم انتخاب سؤال در یک فاصله معین روی پیوستار صفت مکنون با دامنه‌ای از ضریب دشواری دلخواه است و یا می توان از رویکرد ترکیبی استفاده کرد. به عنوان مثال در این گنجینه ۳۰ سؤالی ما قصد داریم روی پیوستار صفت مکنون که از ۳- تا ۳+ امتداد دارد سؤالاتی را انتخاب کنیم که ضریب دشواری آنها برابر یا بین نقطه ۱ تا ۲ باشد. یعنی $1 \leq b_i \leq 2$و شیب هر سؤال بزرگتر مساوی مقدار دلخواه ۲ باشد به عبارتی $a_i \leq 2$. با این محدودیت های تعریف شده در نهایت پنج سؤال از این۳۰ سؤال انتخاب شدند. سؤال ۵، ۷، ۲۱، ۲۳ و ۲۷. این سؤالات در روی پیوستار توانایی در دامنه ۱ تا ۲ به علت شیب بیشتر از ۲ بهترین قدرت جداسازی افراد را دارند. در نمودار۶ تابع آگاهی این آزمون ۵ سؤالی روی پیوستار تتا نشان داده شده است. در این نمودار بهترین برآورد از توانایی افراد در نقطه ۱/۳۵ تتا صورت خواهد گرفت. تابع آگاهی آزمون در این نقطه برابر است با ۱۲/۴۸۳۵ و مقدار خطای معیار برآورد برابر است با ۰/۷۸۳۰. همانطور که می بینیم مقدار آگاهی از توانایی افراد از ۴۳/۲۱ به ۱۲/۴۸ کاهش یافته است که بزرگ ترین دلیل آن کاهش دامنه آزمون از ۳۰ سؤال به ۵ سؤال است.

نمودار شماره ۶. تابع آگاهی آزمون ۵ سؤالی



و ضریب دشواری آزمون تقریبا ۰/۴ واحد انحراف استاندارد افزایش پیدا کرده و آزمون سخت تر از مجموعه۳۰ سؤالی شده است. انتظار می رود افرادی که تتایی برابر با ۱/۵ دارند حداقل به ۳ سؤال از مجموعه پنج سؤالی پاسخ دهند(به نمودار۷ نگاه کنید).

نمودارشماره ۷. منحنی ویژگی آزمون ۵ سؤالی



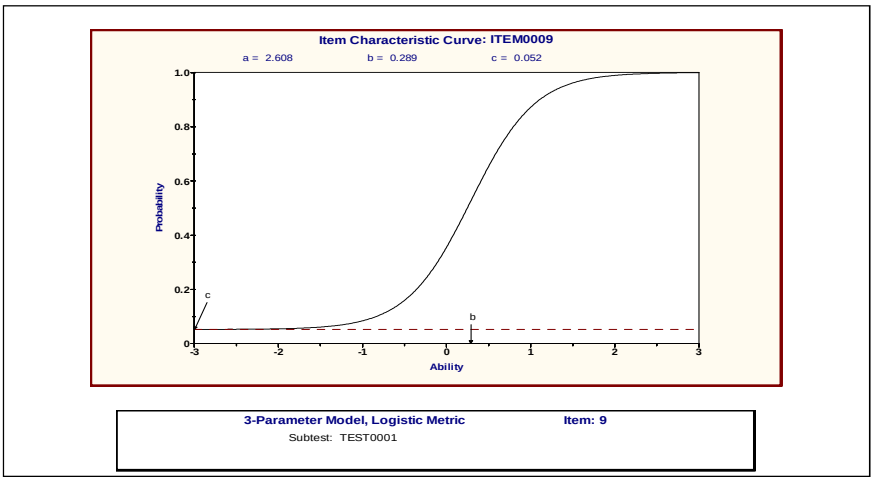
البته باید دوباره یادآوری شود که اگر گنجینه سؤالات خیلی بزرگ باشد برای انتخاب یک مجموعه سؤال دلخواه، طبیعتاً بیش از ۵ سؤال انتخاب خواهد شد و در نتیجه در آزمون های مدادکاغذی که سؤالات آن از بانک مدرج شده تهیه شده است می توان توانایی را با دقت بیشتری برآورد کرد.

در مرحله بعد یک شبیه سازی پیش تجربی[۱] برای اجرای واقعی سؤال ها روی یک شبکه کامپیوتری بر مبنای CAT و بر اساس نرم افزار POSTSIM (ویس،۲۰۰۵) صورت گرفته است. برای این کار تمام۳۰ سؤال به عنوان یک آزمون جدید روی یک گروه۵۰ نفری از فایل داده های پاسخ سوال‌ها، اجرای مجازی شده است. به این ترتیب که بهترین سؤال که معمولا باید شیب بالایی داشته باشد ارایه می شود و سپس سؤالات بعدی بر اساس پاسخ گویی فرد به این سؤال ارایه می شوند. اگر پاسخ صحیح بود سؤال بعدی با بهترین تابع آگاهی برای فرد انتخاب می شود و اگر پاسخ غلط بود سؤال آسان‌تری ارائه می شود تا زمانی که بهترین اندازه توانایی افراد، با کمینه خطا و بیشینه تابع آگاهی به دست آید. این الگو از روی رکورد های فایل پاسخ های افراد ایجاد می شود و ترتیب ارایه هر سؤال در فایلی که به همراه خروجی نرم افزار شبیه سازی است برای هر فرد ارایه می گردد. هدف از شبیه سازی این است که آزمون گیرنده از ماهیت سؤالات ویژه ای که از بانک انتخاب کرده و می خواهد در شبکه کامپیوتری بر اساس الگوی CAT اجرا کند مثل تابع آگاهی، پارامترهای سؤالات و نظایر آن، یک چشم انداز اولیه داشته باشد و بتواند سؤالات بدکارکرد را برای CAT خارج کند. در این شبیه سازی سؤال ۹ که دارای شیب زیاد و دشواری مناسبی است برای نقطه شروع اجرای سؤال در CAT انتخاب شده است. نمودار این سؤال، در زیر آمده است. نتایج نشان داد که تنها ۸ تا ۱۰ سؤال برای بدست آوردن توانایی این افراد با احتمال خطای برآورد $p \leq 0/01$ بر اساس الگوی سنجش انطباقی کافی بوده است. در مقایسه با آزمون مدادکاغذی تنها یک سوم سؤالات به همان خوبی آزمون ۳۰ سؤالی کارایی دارند[۲].

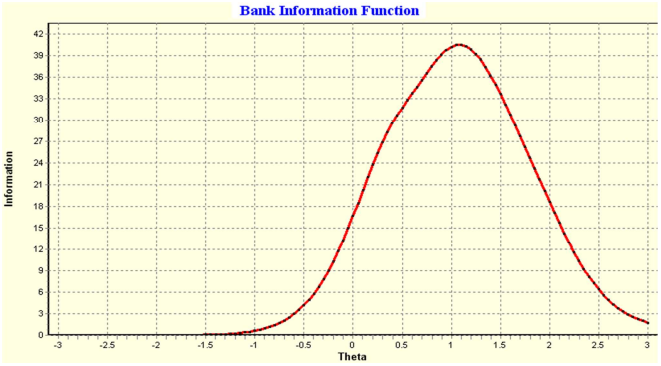
[۱]. Post-Hoc

[۲].اطلاعات بیشتر در این زمینه را می توان در پروژه ای تحت همین عنوان در مرکز مطالعات تحقیقات و ارزشیابی آموزشی سازمان سنجش آموزش کشور بدست آورد.

نمودارشماره ۸. نمودار سؤال ۹، به عنوان نقطه شروع پاسخ گویی به سؤالات در سنجش انطباقی کامپیوتری



نمودارشماره ۹. تابع آگاهی ۳۰ سؤال(برونداد POSTSIM) که برای سنجش انطباقی کامپیوتری روی ۵۰ نفر به طور شبیه سازی شده اجرا شده است



نتیجه گیری

به علت اینکه روش های جدید اندازه گیری هم در تئوری و هم در عمل دارای توان بالایی هستند، در آزمون سازی هایی که بر مبنای این روش ها صورت بگیرد خطای برآورد به نحو چشم گیری کاهش می یابد و کار آزمون سازی نیز با حداقل هزینه و زمان صورت خواهد گرفت. عملا در بعضی از فرایند های آموزشی، اجرای آزمون های ورودی برای دوره های خاص ضروری است. به عنوان مثال، دوره های آموزش زبان انگلیسی نیازمند تعیین سطح افراد در حیطه های چندگانه است. وجود یک بانک سؤال مدرج شده کمک خواهد کرد تا

بدون تلف کردن هزینه، انرژی و زمان، بهترین دوره های آموزشی که دقیقا بـا سـطح ورودی افراد همخوان باشد را اجرا کرد. یا فرض کنیم برای گروههای آموزشی متفاوت نیـاز بـه دوره های ویژه ای باشد می توان به یک بانک یا گنجینه سؤال مدرج رجوع کرد، سـؤالات مـرتبط با موضوع را انتخاب و به راحتی ویژگی های دوره آموزشی و کم و کیف تمرکز بر موضعات خاص دوره را مشخص کرد.

از مزایای دیگر بانک سؤال این است که اجازه می‌دهـد سـؤالات دلخـواه را طراحـی و در آن ذخیره کرد و برای ساخت یک آزمون و یا خرده آزمون دیگـر نیـازی بـه قـرار دادن افـراد در قرنطینه و تهیه مجموعهٔ زیادی از سؤالات آزمایشی و پیش‌تجربی نباشد و به جای آن تنها از یک بانک، سؤالات را بیرون آورد.

یکی دیگر از مزایای بانک سؤال این است که به برنامه‌ریزان درسی کمک می‌کند تـا بـه زبـان مشترکی در رابطه با اهداف برنامه درسـی و موضـوعات آن دسـت یابنـد. همچنـین سـؤالات مدرج شده مشخص می‌کنند که آیا دانشجویان از پس تکالیف خود برمی آیند یا نـه. جایگـاه هر سؤال روی یک پیوستار سلسله مراتبی به افراد درگیر در آموزش این امکان را می‌دهـد تـا دشواری نسبی یک تکلیف درسی خاص را مشخص کنند و این امر کمک می‌کند تا راه هـای آموزش سلسله مراتبی مطالب را از لحـاظ دشـواری مشـخص کـرد و برنامـه درسـی بهتـری سازمان‌دهی کرد.

همانطور که گفتیم در روشهای کلاسیک آزمون سازی نمره افراد تحت تاثیر سختی یـا آسـانی آزمون قرار می گیرد و برآورد توانایی افراد تابع سختی یا آسانی آزمون اسـت و سـؤالات نیـز در مجموعه آزمون معنی پیدا می کنند. عمدتاً دیده شده که در آزمون سازی‌هایی که بصـورت سنتی اجرا می شود، سؤالات یا بسیار دشوار هستند که تعداد کمی مـی تواننـد بـه آن جـواب دهند و یا آسان هستند و تعداد زیادی نمـره مشـابه ای دریافـت مـی کننـد. ایـن امـر باعـث محدودیت در گزینش افراد می شود. اما با استفاده از یک بانک سؤال مـدرج شـده مـی تـوان بهترین سؤالات را که همسان با توانایی آزمون دهندگان است انتخاب کرد. با در اختیار داشتن یک بانک سؤال می توان تاثیر حذف یک سؤال از مجموعه آزمون یـا اضـافه کـردن سـؤالات بیشتر به آزمون را به راحتی کنترل کرد. یادآوری می شود که برآورد توانایی بر اساس آزمون-های حاصل از یک بانک سؤال مدرج شده وابسته به سؤال و گروه آزمودنی‌هـا نیسـت یعنـی

هم ویژگی های سؤالات استقلال دارند و هم توانایی و مقدار صفت مکنون افراد. به این ترتیب ایجاد فرم های موازی از یک آزمون با وزنی دقیقا برابر اما سؤالات متفاوت به راحتی امکان پذیر است.

محدودیت های بانک سؤال

مهمترین گام اساسی در ایجاد یک بانک سؤال، طراحی آن است که شامل فراهم سازی افراد متخصص در این زمینه، مشخص کردن چیزهایی که برای راه اندازی یک بانک لازم است و مشخص کردن آنچه که امیدواریم با راه اندازی یک بانک سؤال انجام پذیرد.

بی شک ایجاد یک بانک سؤال و تحلیل آزمون ها و سؤالات بر اساس IRT همه مشکلات آزمون سازی را حل نخواهد کرد و همچنان مسئله لو رفتن سؤالات در اجرای آزمایشی و یا اجرای بیش از حد یک سؤال و آشنا شدن افراد با راه حل آن باقی مانده است اما این‌ها مشکلاتی هستند که در آزمون گیری های سنتی نیز وجود دارند. ویژگی های برآورد توانایی در IRT کمک می کند تا تاثیر آشنایی افراد به پاسخ‌گویی به سؤال به حداقل برسد: چون بنابر دلایل قبلی که گفته شد برآورد توانایی با ویژگی های سؤال استقلال موضعی دارند به‌ویژه اگر آزمون تک بعدی باشد. اما بیشترین سعی باید بر این باشد که همیشه بهترین سؤالات در یک بانک سؤال نگهداری شود و بیشترین تلاش نیز در زمینه نوشتن سؤالات باکیفیت شود. بعلاوه یک بانک سؤال باید در حیطه موضوعی کاملاً غنی باشد و آزمونی که از بانک تهیه می شود روایی کاملی داشته باشد تا بتواند صفت مکنون را در آن حیطه موضوعی خاص کاملاً پوشش دهد و برای سؤالات ضعیف سؤالات معادلی با کیفیت بهتر در دسترس باشد. برای این که بتوان به طور صحیح سؤالات آزمون‌ها را مدرج کرد و مقیاس آنها را طراحی کرد سؤالات مستلزم آنند که به آزمون دهنده‌هایی با دامنه‌ای گسترده از توانایی ارائه شوند.

بعلاوه در ساخت آزمون و ایجاد بانک سؤال باید مسئله بعدیت آزمون و گنجینه سؤالات مد نظر قرار گیرد. بر این اساس حتی اگر سؤال‌هایی در دو بعد متفاوت برای یک موقعیت آزمون گیری مناسب باشد، سؤالات باید در دو بعد تحلیل و به عنوان دو آزمون تفسیر شوند. هر بانک سؤال در برگیرنده سؤالات گوناگونی است که ممکن است بعدیت آزمون ساخته شده از بانک را مورد تردید قرار دهد. این امر نیازمند کالیبر کردن مجدد سؤالات، معادل سازی و سازماندهی مجدد آزمون ها و سؤالات است که به نوبه نیازمند حجم بالایی از برنامه ریزی،

طراحی، دانش نظری و عملی است. باید تاکید شود که اگر فرایند ساختن بانک سؤال به‌طـور ناشیانه طراحی شود و از روش های مختلفی که برای تحلیل بعـدیت آزمـون هـا و سـؤالات، تعیین پارامترهای سؤالات، برآورد توانایی افـراد و نظـایر آن وجـود دارد بـه طـور ناآگاهانـه استفاده شود در نهایت چیزی که تهیه می شود انباری از سؤالات به ظاهر تفکیک شده اما بـی نهایت در هم و برهم است که باعث اتلاف وقت و نهایتاً هدر دادن هزینه ها شده است.

این امکان وجود دارد که مؤسسات و سازمان‌ها بتوانند بانک سؤال را با موفقیت ایجاد کننـد و با استفاده از مدل های مختلفIRT آزمون های خود را مدرج کنند اما نیازمند دانش گسـترده ای در موردIRT وCAT و تئوری های مربـوط بـه هـوش و پیشـرفت تحصـیلی هسـتند و حوزه‌های امتحانی یا مؤسسات و سازمان هایی که یک پروژه بانک سؤال را انجـام مـی‌دهنـد باید آگاهی کاملی از جنبه‌های عملی و همین طور تئوریک بانک سؤال داشته باشند. هم‌چنـین این مؤسسات و سازمان‌ها، نیازمند کارکنانی هستند که بتوانند سؤالات آزمون را از لحاظ فنی، سازگاری و همخوانی با برنامه تحصیلی، تک بعدی بودن و توان بالقوه هر سؤال برای انـدازه گیری پیشرفت تحصیلی نقادانه مورد بررسی قراردهند.

فهرست منابع

Baker, F. B., Kim, H.K. (2004), "*Item Response Theory: Parameter Estimation Techniques"*, Second Edition, Revised and Expanded. Marcel Dekker, Inc., 207 Madison Avenue, New York 12701, USA.

Brennan, Robert l. (2006), "*Educational Measurement"*, Fourth edition, American council on education.

Chang, Chih -Hung. (2007), "*Developing tailored instruments: item banking and computerized adaptive assessment"*, Northwestern University Feinberg School of Medicine.

Eggen, Theo J.h.m. (2004*), "Contributions to the theory and practice of computerized adaptive testing.Omslag*", RoelOttow / Harold Kainama.Druk: Print Partners Ipskamp B.V., Enschede © Copyright 2004. ISBN 90–5834-056-2.

Georgiadou, E., Triantafillou, E., Economides, A. (2007), *" A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005",* Journal of Technology, Learning, and Assessment, 5(8). Retrieved [date] from http://www.jtla.org.

Ho Yu, Chong, Angel Jannasch-Pennell, & Samuel DiGangi. (2008), *"A Non-Technical Approach for Illustrating Item Response Theory",* Journal of Applied Testing Technology.

Rudner, Lawrence. (1998), "*Item Banking ERIC/AE Digest*", ERIC Clearinghouse on Assessment and Evaluation Washington DC. ED423310.

Sijtsma, K. &Molenaar I.W. (2002), "*Introduction to Nonparametric Item Response Theory",* Thousand Oaks, CA: Sage Publications.

Sukamolson, Suphat. (1996), "*Computerized test/item banking and computerized adaptive testing for teachers and lecturers*", Language Institute Chulalongkorn University.

Thompson, Nathan A., & Weiss, David A. (2011), *"A Framework for the Development of Computerized Adaptive Tests"*, *Practical Assessment, Research & Evaluation*, 16(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

Tshering, Gembo. (2006), *"IRT in Item Banking, Study of DIF Items and Test Construction (Item Response Theory in Item Banking, Study of Differentially Functioning Items and Test Construction*", Master of Science-Track Educational Evaluation and Assessment Educational Science and Technology. University of Twente the Netherlands.

Van der Linden. J. &Glas, C. A.W. (2010), *"Elements of Adaptive Testing"*, Statistics for Social and Behavioral Sciences. New York: Springer. http://www.springer.com/

Weiss, D. J. (2005), *"Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing"*, Version 2.0. St. Paul MN: Assessment Systems Corporation.