نامهٔ آموزش عالی، دورهٔ جدید
سال سوم، شمارهٔ یازدهم، پاییز 1389، صص 85- 63

تاریخ دریافت مقاله 89/01/21
تاریخ پذیرش 89/06/17

سنجش انطباقی با رایانه

دکتر محمد عسگری *

چکیده

منطبق کردن دشواری سئوال با سطح ویژگی آزمودنی، کارآترین روش اندازه‌گیری یک ویژگی است. برای این منظور به آزمودنی‌ها، سئوال‌هایی ارائه می‌شود که اطلاعات دقیقی در بارهٔ میزان و سطح ویژگی آنها، بدست آید. این موضوع اندیشهٔ زیربنایی سنجش انطباقی است و از یافته‌ها و نتایج مدل‌های نظریهٔ سئوال- پاسخ، در تحلیل سئوال‌های آزمون، متأثر است. سنجش انطباقی از نتایج طبیعی تفکر به شیوهٔ بیز در برآورد و بررسی تابع آگاهی سئوال آزمون است، و دارای دلالت‌های ضمنی برای برای تفسیر و تهیهٔ آزمون است. این شیوه دارای تقریب‌های ساده و متنوعی است. سنجش انطباقی با رایانه (سار)، یک روش پیچیده‌تر سنجش انطباقی است. اجرای برنامهٔ سار تکرار فرآیندی دو مرحله‌ای است که در گام نخست، سئوالی به آزمودنی عرضه می‌شود که دشواری آن با توانایی فعلی برآورد شده برای او برابر باشد. در گام دوم، توانایی آزمودنی براساس اطلاعات حاصل از پاسخ او به سئوال مرحلهٔ اول، برآورد می‌شود. این گام‌ها تکرار می‌شوند تا جایی که ملاک‌های پایان دادن به آزمون محقق شوند. برنامه‌ریزی واقعی، اجرای عملی و حفاظت برنامهٔ سار ساده نیست. چالش‌های عملی و پیچیدگی‌های برنامه سار شامل: خزانه‌های سئوال، حفاظت آزمون، و مسائل مربوط به آزمودنی‌ها.

واژگان کلیدی: سنجش انطباقی، سنجش انطباقی با رایانه، نظریهٔ سئوال- پاسخ

* عضو هیات علمی دانشگاه ملایر (مسئول مکاتبات: drmasgari@yahoo.com)

مقدمه

ایدهٔ اساسی سنجش انطباقی، منطبق کردن سئوال‌های آزمون با سطح صفت (توانایی یا مهارت) آزمودنی، در جریان آزمودن است (اگن[1] ، 2007). به عبارت دیگر، این روش مستلزم انتخاب سئوال‌های آزمون در هنگام اجرای آن است؛ به طوری که سئوال‌هایی که برای آزمودنی اجرا می‌شود از لحاظ دشواری مناسب او باشد. اگر انطباق سئوال‌های آزمون با سطح صفت (توانایی یا مهارت) آزمودنی با استفاده از رایانه انجام شود، به آن «سنجش انطباقی با رایانه[2]» (سار) گفته می‌شود (وای[3]، 2005). در سار، صفحهٔ رایانه در هر زمان، یک سئوال که از خزانه بزرگ سئوال‌های طبقه‌بندی شده بر اساس محتوا و دشواری انتخاب می‌شود، را برای آزمودنی نمایش می‌دهد. اولین سئوال در سنجش انطباقی با رایانه همیشه دارای سطح دشواری متوسط است. انتخاب سئوال بعدی با توجه به عملکرد آزمودنی در سئوال‌های قبل توسط رایانه صورت می‌گیرد. بدین ترتیب فرد به سئوال‌هایی پاسخ می‌دهد که منطبق با صفات زیر بنایی عملکرد یا خصیصهٔ مکنون او باشد. در نهایت مجموع سئوال‌های پاسخ داده شده، آزمونی را تشکیل می‌دهد که منطبق با سطح توانایی آزمودنی است. سنجش انطباقی با رایانه (سار) در عمل دارای مراحل زیر است:

1. داشتن خزانهٔ بزرگی از سئوال‌ها؛ 2. اجرای آزمون یا تعداد زیادی سئوال روی آزمودنی‌ها؛

3. استخراج نتایج آزمون بر اساس مدل‌های نظریهٔ سئوال- پاسخ IRT[4] ؛ 4. گزینش سئوال‌های آزمون منطبق با سطوح توانایی آزمودنی‌ها (اگن،2007؛ وای، 2005؛ هورنکه[5]، 2000).

تأکید اصلی این مقاله بر جنبه‌های روانسنجی سنجش انطباقی با رایانه (سار) است. همچنین کاربردهای عملی آن نیز مورد توجه قرار می‌گیرد. این مقاله با توصیفی کلی از سنجش انطباقی شروع شده، پس از آن دربارهٔ سار، مزیت‌ها، ویژگی‌های روانسنجی، و تلویحات کاربردی آن بحث می‌شود. در نهایت به نتیجه‌گیری از بحث‌ها پرداخته شده است.

1. Eggen
2. Computerized Adaptive Testing (CAT)
3. Way
4. Item Response Theory (IRT)
5. Hornke

سنجش انطباقی

سنجش انطباقی[1]، برازش[2] یک آزمون را در جریان آزمودن سطح یک صفت (توانایی یا مهارت) یک آزمودنی است (اگن، 2007). این سنجش از یافته‌ها و نتایج مدل‌های نظریهٔ سئوال- پاسخ (IRT) در تحلیل سئوال‌های آزمون‌ها متأثر است (ستاری، 1382). درIRT فرض می‌شود که؛ الف) عملکرد آزمودنی در آزمون را می‌توان با مجموعه‌ای از عواملی که صفات[3]، صفات مکنون[4]، و یا توانایی‌ها[5] نامیده می‌شوند، برآورد کرد؛ و ب) ارتباط بین عملکرد آزمودنی در سئوال و مجموعه صفات مذکور با یک تابع تجمعی که ویژگی سئوال نامیده می‌شود، قابل برآورد است (سوتاریدونا[6]، پرنل[7]، و واجو[8]، 2003).

کارآترین روش اندازه‌گیری یک ویژگی یا نمره واقعی، منطبق کردن سطح دشواری سئوال با سطح ویژگی یا نمره واقعی آزمودنی است؛ به این معنی که سئوال‌های دشوار برای آزمودنی‌هایی که دارای نمره واقعی بالا یا سطح بالایی از ویژگی هستند، و سئوال‌های آسان برای آزمودنی‌هایی که دارای نمرهٔ واقعی پایین یا سطح پایینی از ویژگی هستند، اجرا شود. هنگامی که امکان حدس زدن وجود ندارد، کارآترین سئوال، سئوالی است که احتمال پاسخ صحیح دادن آزمودنی به آن، برابر 0/5 باشد. زمانی که گروهی از آزمودنی‌ها با تنوع زیادی از ویژگی‌ها با همان آزمون اندازه‌گیری می‌شوند، امکان ندارد که به طور همزمان، آن آزمون حداکثر کارآیی را برای کلیهٔ آزمودنی‌ها داشته باشد. در این موارد، بهتر است سئوال‌هایی به آزمودنی داده شود که اطلاعات عمده‌تری را در بارهٔ میزان و سطح ویژگی او بدست دهد. برای رسیدن به این هدف، باید سئوال‌های متفاوتی برای آزمودنی‌های متفاوت مطرح شود؛ یعنی اندازه‌گیری انفرادی شود و این همان اندیشهٔ زیربنایی و اساسی سنجش انطباقی است (اگن، 2007).

1. Adaptive testing
2. Tailoring
3. Traits
4. Latent traits
5. Abilities
6. Sotaridona
7. Pornel
8. Vallejo

سنجش انطباقی، از نتایج طبیعی تفکّر به شیوهٔ بیز دربارهٔ برآورد و بررسی تابع آگاهی سئوال آزمون است (ثرندایک[1]، 1982؛ ترجمهٔ هومن، 1369). هدف اساسی از بکارگیری روش‌های بیز در اندازه‌گیری، افزایش دقت برآورد نمره واقعی آزمودنی است. این مطلب در واقع استفاده مؤثر از روش بیز در سنجش برازش یافته، یعنی تطبیق دقیق دشواری تکلیف آزمون با سطح توانایی فرد است. یعنی کوششی برای طرح‌ریزی روش‌های بخصوصی از سنجش است که در هر مرحلهٔ آن، سئوال بعدی به گونه‌ای انتخاب شود که با توجه به برآورد کنونی جایگاه آزمون شونده در صفت مکنون مورد اندازه‌گیری، بیشترین مقدار اطلاعات ممکن را بدست دهد. تفکّر به شیوه بیز دارای دلالت‌های ضمنی هم برای تفسیر آزمون و هم برای تهیه آزمون است. در مورد تفسیر آزمون، رهنمودی برای تهیه بهترین برآورد ممکن از جایگاه آزمون شونده در یک خصیصه مکنون با توجه به اطلاعاتی که دربارهٔ تاریخچه گذشته فرد و کارکرد فعلی او در یک یا چند آزمون در دست داریم؛ در اختیار ما قرار می‌دهد. در مورد تهیه آزمون، برای گزینش مواد آزمون پایه‌ای منطقی به دست می‌دهد که بیشترین مقدار اطلاعات را برای هر مرحله سنجش فراهم می‌آورد، و در نتیجه درجهٔ دقت برآورد بعدی ما از وضع نسبی فرد، در صفت مکنون با بیشترین سرعت ممکن افزایش می‌دهد. این موضوع در قلب سنجش انطباقی یا سنجش برازش یافته قرار دارد که علاقه و توجه به آن روز افزون است (ثرندایک،1982؛ ترجمهٔ هومن، 1369؛ آلن[2] و ین[3]، ترجمهٔ دلاور،1374؛ افروز و هومن،1375). پیشرفت‌های نظریه‌های اندازه‌گیری، بخصوص نظریهٔ سئوال- پاسخ، تکنولوژی رایانه و فنّاوری اطلاعات و استفاده از آنها در سنجش، ابزارهای لازم برای سنجش انطباقی با رایانه را در سال‌های اخیر تسهیل نموده است (اگن، 2007).

سنجش انطباقی با رایانه

به طور سنتی، ارزشیابی دانش، مهارت‌ها، توانایی‌ها، و سایر ویژگی‌های شخصیتی توسط آزمون‌های کتبی (مداد- کاغذی) انجام می‌گرفته است. پیشرفت و توسعهٔ

1. Thorndike
2. Allen
3. Yen

فنّاوری اطلاعات[1] در دو دهۀ گذشته آزمودن مبتنی بر رایانه[2] را هم در پژوهش‌های آموزشی و هم در عمل تسهیل نموده است (تاو[3]، وا[4]، و چنگ[5]، 2008). در سنجش انطباقی با رایانه که مورد خاصی از سنجش مبتنی بر رایانه است، هر آزمودنی یک آزمون منحصر به فرد که برای سطح توانایی او برازش یافته است، دریافت می‌کند. سار، استفاده از رایانه در انتخاب سئوال‌ها، ضمن پاسخگویی آزمودنی به هر سئوال است. زمانی که آزمودنی به سئوال پاسخ درست می‌دهد، براساس فرمول خاصی، سئوال مشکل‌تری به او ارائه می‌شود. چنانچه به سئوال پاسخ غلط داده شود، سئوال آسان‌تری پیشنهاد می‌گردد. انتخاب سطح دشواری سئوال‌های متوالی، براساس عملکرد آزمودنی در هر مرحله صورت می‌گیرد. بعد از پاسخ آزمودنی، توانایی او به روز شده، و سئوالی از خزانه سئوال برای وی انتخاب می‌شود که بیشترین تناسب را با توانایی جدید او داشته باشد و این روند ادامه پیدا می‌کند تا ملاک‌های پایان دادن به آزمون محقق شوند (ترین تافولو[6]، جرجیادو[7]، و اِکونومیدس[8]، 2006). سنجش انطباقی با رایانه (سار) به صورت روز افزون و در مقیاسی بزرگ در برنامه‌های سنجش و آزمون وارد شده است. مزیت‌های سار هم برای سازندگان و هم برای اجراکنندگان آزمون مفید و امیدوارکننده بوده است. زیرا، به آزمون‌کننده اجازه می‌دهد تا سطح ویژگی آزمودنی را به سرعت بسنجد، بدون اینکه او را وادار کند تا به مجموعه‌ای خسته‌کننده از سئوال‌های بسیار ساده و یا مجموعه‌ای ناخوشایند و ناراحت‌کننده از پرسش‌های خیلی مشکل پاسخ دهد (الیا[9]، رولتا[10]، زیمنز[11]، و آباد[12]، 2000).

1. Information Technology (IT)
2. Computer- Based Testing (CBT)
3. Tao
4. Wa
5. Chang
6. Triantafillou
7. Georgiadou
8. Economides
9. Olea
10. Revuelta
11. Ximenez
12. Abad

انگیزهٔ اصلی استفاده از سار، در کارآیی آن نهفته است. در مقایسه با روش‌های سنجش مداد- کاغذی سنتی، افزایش کارآیی سنجش در نتیجهٔ رایانه‌ای کردن شیوهٔ آزمودن، با تأکید بر کارآیی اندازه‌گیری است. نتایج مطالعات نشان می‌دهد که در سار و با دقت یکسان، به سئوال‌های کمتری در مقایسه با روش‌های غیرانطباقی، نیاز هست. این سنجش به حدود 50 تا 60 درصد سئوال‌های در مقایسه با روش‌های غیرانطباقی نیاز دارد. امتیازهای دیگر سار نسبت به روش‌های سنتی عبارت‌اند از:

1. هر فردی یک آزمون منحصر به فردی را پاسخ می‌دهد که هم محتوا و هم طول آزمون می‌تواند از فردی به فرد دیگر متفاوت باشد؛ 2. سار برای هر فردی بهینه است؛ به گونه‌ای که حداقل دو پیامد مطلوب خواهد داشت: الف) از آنجا که سئوال‌های کمتری برای دستیابی به دقتی مشابه، در سار نیاز است، کارآیی اندازه‌گیری افزایش می‌یابد، ب) هر آزمودنی می‌تواند با سطح توانایی خود در چالش باشد؛ این چالش اثر تحریکی و تجربه شده‌ای به عنوان تقویت بر توانایی آنها دارد؛ 3. چون برای اندازه‌گیری توانایی هر فردی به سئوال‌های کمتری نیاز است، بنابراین در هزینه و وقت آزمودنی‌ها، تهیه‌کنندگان آزمون، و مجریان آن صرفه‌جویی می‌شود؛ 4. از آنجا که سئوال‌های آزمون در چارچوب نظریهٔ IRT تهیه می‌شود، بنابراین آزمون دارای تعدادی ویژگی شناخته شده است؛ بر همین اساس کیفیت احتمالی اندازه‌گیری بهتر است؛ 5. نمره‌گذاری آن عینی و بدون خطاست و امکان نمره‌گذاری و بازخورد مستقیم وجود دارد؛ 6. دستیابی به انواع سئوال‌های جدید امکان‌پذیر است؛ 7. اجرای آزمون کارآیی بیشتری دارد؛ 8. برنامه‌ریزی و مکان‌یابی آزمون انعطاف‌پذیر است؛ 9. ایمنی آزمون بیشتر است؛ 10. آزمودنی‌ها انگیزه بیشتری دارند؛ و 11. امکانات بیشتری برای بکارگیری نظریه‌های جدید آزمون وجود دارد (اگن، 2007).

اجرای برنامهٔ سنجش انطباقی با رایانه، اساساً تکرار یک فرایند دو مرحله‌ای است. در گام اول، سئوالی به آزمودنی ارائه می‌شود که دشواری آن با توانایی فعلی (ذاتی) برآورد شده برای او برابر باشد. در گام دوم، توانایی آزمودنی بر اساس اطلاعات حاصل از پاسخ او به سئوال مرحلهٔ اول برآورد می‌شود. سپس این دو گام تکرار می‌شوند تا جایی که یکی از ملاک‌های پایان دادن به آزمون محقق شوند؛ ملاک‌هایی مثل تعداد معینی سئوال، یا با سطوحی از دقت اندازه‌گیری مورد انتظار. از طریق این

فرآیند الگوریتمی[1] سنجش انطباقی با رایانه توانایی برآورد شده برای آزمودنی، تثبیت[2] تثبیت[2] می‌شود. اگرچه، سنجش انطباقی ایدهٔ نظری نسبتاً ساده‌ای دارد، امّا برنامه‌ریزی برنامه‌ریزی واقعی، اجرای عملی و حفاظت یک برنامهٔ سنجش انطباقی با رایانه بسیار پیچیده است. چالش‌های عملی که این پیچیدگی را به وجود می‌آورند و باید در تهیهٔ برنامه‌های سار مورد توجه قرارگیرند عبارت‌اند از: تهیه و نگهداری خزانه‌های سئوال[3]، حفاظت آزمون[4]، و مسائل مربوط به آزمودنی‌ها[5] (الیا و همکاران، 2000؛ وای، 2005؛ اِگن، 2007).

الف) تهیه و نگهداری خزانهٔ سئوال در سنجش انطباقی با رایانه: در این باره، تعدادی موضوع کاربردی مؤثر بر اندازه‌گیری، باید شناسایی شوند. این شناخت به تهیه خزانه سئوال کمک می‌کند و هنگام اجرا، به اجراکنندهٔ آزمون کمک می‌کند تا نتایج حاصل از آزمون را به صفت بنیادی اندازه‌گیری شده ربط دهد. این موضوعات عبارت‌اند از: 1. اندازهٔ خزانهٔ سئوال و کنترل آن، 2. تعیین ابعاد خزانهٔ سئوال، 3. مدل‌های پاسخ، 4. جابجایی و اصلاح سئوال‌ها 5. اضافه کردن سئوال به خزانهٔ سئوال، 6. حفظ هماهنگی مقیاس‌ها.

1. **اندازه خزانه سئوال برای سنجش انطباقی**: برای سنجش انطباقی با رایانه، وجود خزانه سئوال با تعداد سئوال‌های کافی و برازش یافته با یک مدل IRT ضروری است. همهٔ سئوال‌ها باید به خوبی آماده اندازه‌گیری صفتی باشند که، می‌بایست اندازه‌گیری شود. یک خزانهٔ سئوال خوب، تمام جوانب مرتبط با صفتی را که قرار است با استفاده از سار اندازه‌گیری شود، به خوبی پوشش می‌دهد. در اصل، معیارهای سئوال‌های خوب در سار خیلی متفاوت از سئوال‌های خوب در آزمون‌های کتبی (مداد- کاغذی) نیست (اگن، 2007). قبلاً پیشنهاد شده بود که حداقل به خزانهٔ سئوالی با 100 سئوال، برای سنجش انطباقی با رایانه ضروری است (یوری، 1977، به نقلِ الیا و همکاران، 2000). براین اساس، همهٔ آزمون‌سازان تلاش می‌کردند آزمون‌هاییِ تهیه نمایند که حداقل 100 سئوال داشته باشد. اما سه عامل باعث شد که تعداد سئوال‌های خزانه

1. Algorithm
2. Converge
3. Item Pools
4. Test Security
5. Examinee Issues

سئوال افزایش چشمگیری یابد: نخست این که، در چند دههٔ اخیر آزمون‌های سنتی بهبود چشمگیری داشته‌اند، به گونه‌ای که سطح عملکرد آزمون‌های آموزشی تقریباً دقتی به اندازهٔ سنجش انطباقی دو مرحله‌ای دارند. دوم، الزام‌ها و اجبارهایی که بر شیوه‌های انتخاب سئوال‌های سنجش انطباقی تحمیل شده‌اند (مثل الزام‌های محتوایی و منطقی آزمون) ضرورت ایجاد خزانهٔ سئوال‌های بزرگتری را نسبت به نبود چنین الزام‌هایی، به منظور دستیابی به اطلاعات مشابه، به وجود آورده است. سوم این که، اگر قرار باشد آزمونی که تهیه شده است برای مدتی طولانی مورد استفاده قرار گیرد، باید دارای خزانهٔ سئوال زیادی باشد، تا ایمنی سئوال‌ها به خطر نیافتد. امروزه بسیاری از آزمون‌های انطباقی خزانهٔ سئوال‌هایی با بیشتر از 1000 سئوال در اختیار دارند و با استفاده از چرخش دوره‌ای سئوال‌ها، تعداد سئوال‌های خزانه را به 2000 سئوال می‌رسانند. بدیهی است که، محدود کردن دایرهٔ شمول یک آزمون، نخستین عامل در تعیین این مطلب است که چند سئوال برای آزمون لازم است. در تعیین تعداد سئوال‌های خزانه تمایز بین باید و واقعیت کارآیی سنجش انطباقی ضروری است (الیا و همکاران، 2000؛ وای، 2005).

2. **تعیین ابعاد پاسخ‌های سئوال‌های** : خزانه‌های سئوال یک بعدی نیستند و بنابراین پاسخ‌های سئوال‌ها نیز چنین هستند. بنابراین در بررسی ابعاد پاسخ‌ها، تعامل بین آزمودنی‌ها و سئوال‌های آزمون بررسی می‌شود. لذا ضروری است که در تعریف جامعه مورد نظر روشن و صریح اقدام شود. اگر تحلیل ابعاد با داوطلبان غیرعلاقه مند و یا دانشجویان قبل از آموزش انجام شود، ابعادی که به دست می‌آید شبیه همان ابعادی نخواهد بود که از اجرای آزمون برای گروهی دانشجویان واقعی و با انگیزش فردی بالا پس از آموزش به دست می‌آید. دو نکته ضروری در بررسی ابعاد پاسخ‌ها عبارت‌اند از: **نکتهٔ اول** در زمان تهیهٔ خزانهٔ سئوال باید مد نظر قرار گیرد و آن این ضرورت است که تعیین شود که آیا برای توصیف پاسخ‌های آزمودنی‌ها از مدل یک بعدی و یا چند بعدی سئوال پاسخ باید استفاده نمود. این کار را می‌توا ن با تحلیل ابعاد سئوال‌ها انجام داد. **نکتهٔ دوم** آن است که ضرورت دارد، ابعاد فضای پاسخ‌ها زمانی که محتوای آزمون و یا آزمون شوندگان به صورت منظم تغییر می‌کنند، مورد بررسی قرارگیرند. این موضوع اغلب توسط آزمون‌سازان برای تعیین مقیاس اندازه‌گیری مورد غفلت قرار می‌گیرد. روش‌های زیادی برای تعیین ابعاد فضای پاسخ‌ها وجود دارد که عبارت‌اند از: تحلیل عاملی اکتشافی؛ تحلیل عاملی تأییدی؛

تحلیل عاملی کامل؛ مقیاس پردازی چند بعدی؛ و تکنیک‌های متفاوت وابسته به نظریه‌های سئوال-پاسخ (اگن، 2007؛ الیا وهمکاران، 2000؛ وای، 2005).

اگر تحلیل‌ها نشان داد که رویکرد یک بعدی کافی است، استفاده از مدل سئوال پاسخ یک بعدی مقرون به صرفه و عاقلانه است. اما اگر نتایج تحلیل نشان داد که ساختار چند بعدی برای تبیین پاسخ‌های آزمودنی‌ها ضروری است، منطقی است که با دقت ساختار آزمون را مورد توجه قرار داده و برای ساده‌سازی آن تلاش شود به گونه‌ای که مدل‌های سئوال پاسخ چند بعدی ساده‌تری، پاسخ‌ها را تبیین نماید (الیا و همکاران، 2000).

3. **مدل‌های سئوال پاسخ**: در سار داشتن یک خزانه سئوال با کیفیت بالا ضروری است، اما کافی نیست. سئوال‌ها باید با یک مدل مناسب IRT برازش یافته و مدرج‌سازی[1] شده باشند. بر اساس داده‌های تجربی حاصل از اجرای عملی سئوال‌ها، یک مدل IRT برازش یافته و پارامترهای سئوال و پارامترهای یک یا بیشتر از یک توزیع توانایی (θ) در جامعه برآورد می‌شوند. کار یک مدلIRT ، با مدرج‌سازی سئوال‌های آغاز می‌شود. مدرج‌سازی سئوال‌های برازش یک مدلIRT و برآورد پارامترهای سئوال را در برمی‌گیرد. اگر پارامترهای سئوال‌های برآورد شدند، آنگاه آنها را می‌توان در خزانهٔ سئوال نگهداری و برای سار مورد استفاده قرار داد. در یک سار احتمال عملکرد تتای یک آزمون نقش اساسی در استنباط در بارهٔ آزمودنی دارد. خصوصیات یک مدلIRT خوب، برای استفادهٔ عملی و مناسب در سار عبارت‌اند از:

1. امکان برآورد توانایی (θ تتای) یک آزمودنی با مقیاسی مشترک، و با هر زیر مجموعه‌ای از سئوال‌های مدرج‌سازی شده در خزانهٔ سئوال وجود دارد. بنابراین، ضرورتی ندارد که مجری آزمون، سئوال‌های یکسانی را به آزمودنی بدهد تا برآورد قابل مقایسه‌ای از θ آنها بدست آورد.

2. دشواری سئوال با مقیاس مشترکی مثل توانایی θ آزمودنی‌ها بیان می‌شود. بنابراین، امکان برازش یک آزمون با هر سطحی از توانایی θ یک آزمودنی وجود دارد.

3. آگاهی یک سئوال عملکردی از توانایی θ است، در نتیجه تابع آگاهی می‌تواند به عنوان اساسی برای برازش سئوال انتخابی به کار گرفته شود. با این وجود، ویژگی‌های مطلوب فوق، فقط زمانی درست‌اند که برازش سئوال‌های انتخاب شده برای مدل IRT لحاظ شده باشند (اگن، 2007).

1. Calibration

بسیاری از پژوهشگران انتخاب مدل ویژه سئوال پاسخ را به مجادله تبدیل نموده‌اند. اگرچه، هیچ دلیل نظری و یا عملی وجود ندارد که باید فقط به یک مدل پاسخ محدود شد. برای نمونه، اگر فردی بخواهد یک مقیاس اندازه‌گیری با تعداد محدودی از افراد بسازد، مدل یک پارامتری[1] مطلوب‌ترین مدل احتمالی برای استفاده اوست (لرد[2]، 1983)، حتی اگر انتظار برآزنده بودن سئوال‌های خزانه با مدل یک پارامتری را نداشته باشیم. زمانی که حجم نمونهٔ بزرگتری در دسترس باشد، و فرد قصد افزودن تعدادی سئوال به سئوال‌های خزانه داشته باشد، بدون این که ویژگی‌های مقیاس اندازه‌گیری آسیب ببیند، هیچ چیزی مانع از به کارگیری مدل سه پارامتری[3] برای سئوال‌های جدید نیست. به همان نسبتی که به سئوال‌ها افزوده می‌شود، مقیاس نیز مطلوب‌تر می‌شود (الیا و همکاران، 2000).

برای دستیابی به کیفیتی بالا در مدرج‌سازی سئوال‌های، انتخاب مدل IRT، انتخاب روش برآورد پارامترها، آماره‌های برازش، ماهیت و حجم نمونه آزمودنی‌ها نقش مرتبط به هم، و مهمی دارند (فیشر[4] و مولینر[5]، 1995؛ ون درلیندن[6] وهامبلتون[7]، 1996). در انتخاب مدل IRT، هر چه مدل انتخابی ساده‌تر باشد، به حجم نمونهٔ کمتر نیاز است، و روش‌های آماری بهتری برای برآورد کردن و آزمودن برازش مدل وجود دارد. برای برآورد پارامترها با دقتی معقول، و برای آزمودن مدل با توانی بالا در مدلهای یک پارامتری، دو پارامتری، و سه پارامتری؛ به ترتیب حداقل به 200، 500، و 1000 پاسخ آزمودنی در هر سئوال نیاز است. به عبارت دیگر، دستیابی به برازش خوب برای یک مدل ساده‌تر خیلی مشکل است. این بدین معنی است که برخی سئوال‌های باید از خزانه حذف شوند، که این موضوع روایی خزانهٔ سئوال را تهدید می‌کند (اگن، 2007؛ ون درلیندن و گلاس[8]، 2000).

دو روش احتمالی متداول برای برآورد پارامترهای سئوال وجود دارد. **نخست** روش بیشینهٔ احتمال حاشیه‌ای[9] (MML) است. در این روش فرض می‌شود که

1. One-parameter logistic model (1PL)
2. Lord
3. Three-parameter logistic model (3PL)
4. Fischer
5. Molenaar
6. Van der Linden
7. Hambleton
8. Glas
9. Marginal maximum likelihood (MML)

نمونه‌ای تصادفی از جامعه‌ای با توزیع شناخته شده‌ای از توانایی انتخاب شده است. (به صورت معمول فرض می‌شود توزیع توانایی در جامعه نرمال است). با روش بیشینه احتمال حاشیه‌ای (MML)، پارامترهای سئوال و پارامترهای توزیع توانایی همزمان با هم برآورد می‌شوند. **روش دوم**، روش بیشینه احتمال موقعیتی[1] (CML) است. اگر از این روش استفاده شود، امکان برآورد همزمان پارامترهای سئوال و پارامترهای توانایی وجود ندارد. در روش بیشینه احتمال موقعیتی، به مفروضه‌هایی در بارهٔ توزیع توانایی آزمودنی‌ها نیاز نیست؛ فقط به نمونه‌هایی از جامعه نیاز است. این موضوع در عمل بسیار مطلوب است. زیرا، در آموزش و پرورش دستیابی به نمونه‌های تصادفی واقعی به آسانی امکان پذیر نیست. انتخاب نمونه‌هایی با روش خوشه‌ای روایی برآورد پارامترهای سئوال را با استفاده از روش بیشینه احتمال موقعیتی تهدید نمی کند. با این وجود، روش برآورد (CML)، برای هر مدلی قابل استفاده نیست. از این روش در مدل یک پارامتری و گاهی اوقات در مدل دو پارامتری می‌توان استفاده کرد (اگن، 1990؛ ورهلست[2]، گلاس[3]، و ورسترالن[4] 1995). اگر از (CML) برای برآورد کردن پارامترهای سئوال و برازش مدل آزمون استفاده شود، مدرج‌سازی با برآورد پارامترهای یک توزیع توانایی به صورت جداگانه تکمیل می‌شود (اگن، 2007).

4. **جابه‌جایی، اصلاح و آزمودن مجدد**: در تهیهٔ یک خزانه سئوال و مقیاس اندازه‌گیری برای استفاده در یک آزمون انطباقی، ضروری است شیوه‌هایی را تعیین و به کار گرفته شود که به حذف و کنار گذاشتن سئوال‌هایی که عملکرد خوبی ندارند، کمک نماید. این موضوع ارتباطی به مدل‌های سئوال پاسخ انتخاب شده ندارد. به محض این که سئوال‌های ضعیف شناسایی شدند، باید آنها را از خزانهٔ سئوال کنار گذاشت؛ زیرا سئوال‌های ضعیف علت اصلی خطای برآورد توانایی و تصمیم‌گیری دربارهٔ آزمودنی‌ها هستند. یک روش برای بررسی سئوال‌هایی که برآزنده نیستند، رسم منحنی سئوال- پاسخ[5] تجربی و مقایسه آن با منحنی نظری[6] است. مقایسه منحنی

1. Conditional maximum likelihood (CML)
2. Verhelst
3. Glas
4. Verstralen
5. Item response curve
6. Theoretical curve

نظری با منحنی تجربی در اکثر موارد رویکردی بسیار ساده برای تشخیص بصری سئوال‌های ضعیف است. فقط از طریق عدم پذیرش سئوال‌های ضعیف و شیوهٔ اصلاحی است که می‌توان خزانهٔ سئوالی نگهداری نمود که به هماهنگی اندازه‌گیری کمک نماید. در آزمون انطباقی این مطلب مهمی است که هرچه آزمون کوتاه‌تر باشد، عملکرد غیر عادی سئوال‌های ضعیف، برآن تأثیر بیشتری خواهند داشت (الیا و همکاران، 2000).

5. **افزایش سئوال‌های خزانهٔ سئوال**: زمانی که خزانهٔ سئوالی از سئوال‌های مدرج شده[1] برای یک مقیاس اندازه‌گیری بخصوص، در اختیار است، می‌توان از طریق هر طرح لنگر[2]، سئوال‌هایی را، به خزانهٔ سئوال اضافه نمود. معمولاً گروهی از آزمودنی‌ها به مجموعه‌ای از سئوال‌های مدرج شدهٔ قبلی و تعدادی سئوال جدید مدرج شده، پاسخ می‌دهند. پس از آن یک روش لنگر برای مدرج‌سازی سئوال‌های جدید به کار می‌رود. بکارگیری این روش در یک آزمون انطباقی بسیار منطقی است. زیرا؛ طی آن فرد می‌تواند سئوال‌های جدید را در میان آزمون جدید وارد نماید (الیا و همکاران، 2000).

6. **حفظ هماهنگی مقیاس**: اگر یک مقیاس چند سال مورد استفاده قرار گیرد، میانگین‌ها، درصدها، و سایر ویژگی‌های نمونه در طول این سال‌ها تغییر می‌کنند. طرح لنگر با پارامترهای ثابت می‌تواند این تغییرات محیطی را به گونه‌ای تثبیت کند که مقیاس اندازه‌گیری در پارامترهای برآورد شده، انحراف خیلی زیادی نداشته باشد. اگرچه، منطقی است مطالعه‌ای انجام شود و طی آن برخی از سئوال‌هایی که قبلاً مدرج شده‌اند، مجدداً مدرج شوند تا مشخص شود که انحراف ناچیز است و در محدودهٔ بازگشت‌هایی است که ما از خطای نمونه‌گیری انتظار داریم. به علاوه برای حفظ هماهنگی مقیاس می‌توان شرایط اجرایی آزمون را محدود کرد. موضوعاتی مثل تغییرات اجرایی، محدودیت‌های زمانی، تناسب نمونه و سایر شرایط محیطی که علت ناپایداری در مقیاس اندازه‌گیری می‌شوند، را تعریف و مشخص نمود. اگر می‌خواهیم مقیاسی داشته باشیم که در دوره‌ای طولانی از زمان ثابت و پایدار باقی بماند نیازمند

1. Calibrated
2. Linking design

کنترل یا محاسبهٔ این عوامل هستیم (الیا و همکاران،2000). از آنجا که تمام مدل‌های IRT فرض می‌کنند که سئوال‌های آزمون از نوع سئوال‌هایی هستند که توانایی آزمودنی را می‌سنجند، لذا مهم‌ترین عامل ممکن است سرعت باشد. اگر در آزمون محدودیت زمان پاسخگویی اعمال شود، سئوال‌های آخر آزمون دشواری بالا و سئوال‌های نخست دشواری پایینی خواهند داشت و این فرآیند دوری در تکرار مجدد آزمون‌ها باعث فاصله گرفتن مقیاس اندازه‌گیری آزمون جدید، از آزمون قبلی خواهد شد. ساده‌ترین روش برای حل این مشکل آن است که الف) تغییر مکان هر سئوال در حوزه بخصوصی از آزمون امکان‌پذیر شود. ب) فقط از آزمون‌هایی استفاده شود که محدودیت زمانی برای پاسخگویی به آنها اعمال نشود (سوتاریدونا و همکاران، 2003).

ب) حفاظت آزمون: حفاظت آزمون موضوعی جالب در تمام برنامه‌های آزمودن است. اگر ایمنی آزمون به خطر بیافتد، مهم نیست که چگونه ویژگی‌های روان‌سنجی آزمون تغییر می‌کند، بلکه مهم آن است که روایی تفسیر نمرات آن درست نیست. موفقیت سنجش انطباقی با رایانه (سار) به سلامت خزانه سئوال‌های آن بستگی دارد. به میزانی که سئوالی و یا سئوال‌هایی به وسیلهٔ آزمودنی‌های قبلی شناخته شوند، پارامترهای برآورد شده برای آن سئوال (برآورد شده از طریق یک نمونه مدرج‌سازی)، برای مدتی طولانی قابل استفاده نخواهد بود. به نسبتی که آزمودنی‌های بیشتری محتوای آزمون را شناسایی کنند، پارامتر دشواری سئوال‌های آن آزمون ساده‌تر شده و به قسمت پایین مقیاس توانایی انتقال می‌یابند، پارامتر تمیز سئوال‌های آن آزمون به سمت صفر متمایل شده و پارامتر حدس زدن پاسخ درست سئوال‌های به گونه‌ای غیرمنطقی افزایش می‌یابد (کالتون[1]، 1998؛ الیا و همکاران، 2000). به همین دلیل ضروری است که انتخاب و نمره‌گذاری سار محرمانه باشد. دو موضوع جالب و اساسی دربارهٔ محرمانه بودن سئوال‌ها در سار وجود دارد: افشا شدن سئوال‌ها، سرقت سئوال‌ها.

1) **افشا شدن سئوال‌ها**: در آزمون‌های مداد- کاغذی (کتبی) به طور معمول تمام آزمودنی‌ها را می‌توان به طور همزمان آزمون کرد. دفترچه‌های آزمون به اندازهٔ کافی برای آزمون هر یک از آزمودنی‌ها با قیمت مناسب تهیه، و در اختیار آنها قرار داد.

1. Colton

برعکس، سنجش مبتنی بر رایانه پرهزینه است. این احتمال وجود دارد که تعداد رایانه‌ها کمتر از آزمودنی‌هایی باشد که قرار است آزمون شوند. این مطلب بر این دلالت دارد که برخی از آزمودنی‌ها قبل از دیگران آزمون می‌شوند. بعلاوه یکی از مزیت‌های سنجش مبتنی بر رایانه که قابلیت آن را برای ضرورت آزمودن مشخص می‌نماید، آن است که هر یک از آزمودنی‌ها در زمان‌های متفاوتی آزمون می‌شوند و به تعداد زیادی مجری نیاز نیست. این مطلب گویای آن است که یک تمرین عمومی برای آزمودنی‌ها در طول اجرای آزمون وجود دارد که با خودشان دربارهٔ سئوال‌های آزمون نجوا کنند، بخصوص زمانی که پیامدهای آن برای عملکرد در آزمون بالا باشند. دانش‌آموزانی که سئوال‌ها را مطالعه می‌نمایند و از دوستان خود دربارهٔ آنها مطالبی شنیده‌اند و پس از آن به سئوال‌های آزمون پاسخ می‌دهند، به صورت بالقوه نسبت به سایرین برتری خواهند داشت و توانایی آنها با سوگیری مثبت برآورد خواهد شد. یک راه‌حل این مشکل، استفاده از خزانه سئوال‌های بزرگ است، به گونه‌ای که کم کردن هر کدام از این مجموعه از خزانه، چنین افشاگری‌هایی را تحت تأثیر قرار ندهد.
مشکل مرتبط دیگری که در مدارس رخ می‌دهد، این است که معلمان در بارهٔ سئوال‌های آزمون بخصوصی که از دانش‌آموزان قبلی به عمل آورده‌اند پرس‌وجو می‌کنند و بعد به سرعت محتوای آزمون را یاد می‌دهند، در این صورت عملکرد دانش‌آموزان افزایش می‌یابد (الیا و همکاران، 2000؛ سوتاریدونا و همکاران، 2003؛ وای، 2005) .

2) **سرقت سئوال‌ها**: تمام برنامه‌های آزمون باید در بارهٔ سرقت سئوال‌های آزمون حساس باشند. برخی از طرفداران سار معتقدند که چنین آزمون‌هایی ذاتاً از ایمنی بیشتری برخوردارند؛ زیرا از فرم‌های آزمون هیچ کپی وجود ندارد که بتوان آن را سرقت کرد و یا فتوکپی از آن برداشت. با این حال سار در مقابل سرقت سئوال‌هایش کاملاً آسیب‌پذیر است (کالتون، 1998؛ الیا و همکاران، 2000).

ج) مسائل مربوط به آزمودنی‌ها: اجراکنندگان سار باید راه‌حل‌های مناسبی برای تعدادی از مسائل فنی و عملی پیدا کنند. در یک سار که برای ارزیابی توانایی افراد مورد استفاده قرار می‌گیرد، آزمون‌کنندگان باید نسبت به مسائل بالقوه‌ای که ممکن است در حین اجرای یک سار برای آزمودنی‌ها اتفاق بیافتد، هشیار بوده و در این باره آینده‌نگری نمایند.

در نخستین دید اجمالی، تجربه شرکت در یک سار خیلی متفاوت از یک آزمون سنتی به نظر نمی‌رسد. یک سئوال بر روی صفحه رایانه ظاهر می‌شود، آزمودنی پاسخ سئوال خود را وارد رایانه می‌کند، و پس از آن سئوال بعدی ظاهر می‌شود و این فرآیند تا آزمون تمام شود، ادامه می‌یابد. هر چند، چند جنبهٔ منحصر به فرد وجود دارد که در تجربهٔ سار عملکرد آزمودنی یا آزمودنی‌ها را تحت تأثیر قرار می‌دهد. نخست اینکه، سنجش مبتنی بر رایانه برای بسیاری از آزمودنی‌ها، ناآشناست. سئوال‌هایی که روی صفحه رایانه ظاهر می‌شوند ممکن است برای آزمودنی‌ها از لحاظ خواندن دشوارتر و یا آسان‌تر باشند. لازم است سئوال‌های طولانی‌تر، یعنی سئوال‌هایی که در کل صفحه رایانه گسترده شده‌اند، توسط آزمودنی‌ها بر روی صفحه رایانه بالا یا پایین شوند. ورود پاسخ‌های آزمودنی‌ها با استفاده از صفحه کلید و یا موش‌واره از وارد کردن پاسخ در یک پاسخنامه دفترچه آزمون، یا سیاه کردن آن در پاسخنامه متفاوت است. دوم اینکه، در یک آزمون سنتی آزمودنی‌ها معمولاً به تمام سئوال‌های یک آزمون پاسخ می‌دهند. این موضوع برای آزمودنی‌ها آزادی عمل زیادی در انتخاب سئوال‌ها می‌دهد؛ یعنی می‌توانند از بعضی از سئوال‌ها رد شوند تا آنها را بعداً پاسخ دهند. پاسخ سئوال‌ها را مرور کنند و احتمالاً تغییر دهند. برعکس، آزمودنی‌های سار کنترل بسیار کمی بر این موضوع دارند؛ زیرا سئوال‌ها به صورت نوعی یک بار اجرا می‌شوند و در آن آزمودنی امکان مرور سئوال‌ها و پاسخ‌های خود را ندارند (سوتاریدونا و همکاران، 2003). در این قسمت ضمن بحث در بارهٔ دورنمای آزمودنی‌ها در حین اجرای سار ، سه موضوع مربوط به آنها معرفی می‌شود.

1) **مرور سئوال‌ها**: یکی از سئوال‌های بحث‌انگیز مرتبط با اجرای سار که باید در بارهٔ آن تصمیم‌گیری شود، آن است که آیا به آزمودنی‌ها، اجازهٔ مرور و تغییر پاسخ‌هایی را که قبلا به سئوال‌ها داده‌اند، داده شود (وای، 2005). پژوهشگران قدیمی حوزهٔ سنجش انطباقی معتقد بودند که اجازه دادن به آزمودنی برای برگشت به پاسخ‌های خود و تغییر آنها اثرات منفی بر کارآیی اندازه‌گیری در سار دارد: اگر پاسخی تغییر یافت، سئوال‌های بعدی که در نتیجهٔ این تغییر توسط رایانه انتخاب می‌شوند، طولانی‌تر و بهتر از دیگری نخواهد بود. همچنین برخی از پژوهشگران نگرانند که آزمودنی‌ها برای بهبود نمرات خود در سار از تقلب استفاده کنند. آزمودنی‌ها این کار می‌توانند در ابتدا به سئوال‌ها پاسخ غلط بدهند؛ در نتیجه رایانه به آنها ساده‌ترین سئوال‌های ممکن را عرضه می‌کند. پس از آن در فرصت مرور

سئوال‌ها، پاسخ‌های خود را از غلط به درست تغییر داده و نمرات بالاتری کسب کنند (واینر[1]،2000؛ 1983). به همین دلایل، برخی از برنامه‌های سار، امکان برگشت و مرور پاسخ‌ها وجود ندارد. مرور سئوال‌ها در سار به عنوان روشی برای افزایش کارآیی سنجش انطباقی مورد بررسی قرارگرفته است. مرور سئوال‌ها به زمان اضافی نیاز دارد، پاسخ‌های تغییر داده شده ممکن است خطای استاندارد برآورد توانایی آزمودنی را افزایش دهد. ممکن است آزمودنی‌ها به صورت راهبردی از مرور کردن سئوال‌ها برای افزایش نمرات خودشان استفاده نمایند. اگر چه، واکنش‌های آزمودنی‌ها به سار، مثبت بوده است، اما آنها نارضایتی شدیدی نسبت به فقدان مرورکردن سئوال‌ها ابراز نموده‌اند (باقی[2]، فرارا[3]، گابریز[4]، 1994؛ لگ[5] و بوهر[6]، 1992؛ ویسپول[7]، رکلین[8]، و وانگ[9]، 1994).

اگرچه پژوهش‌های جدید نشان داده‌اند که اگر آزمودنی‌ها اجازه مرور پاسخ‌های خود و تجدید نظر در آنها را داشته باشند، اثر آشکاری بر کارآیی اندازه‌گیری ندارد، و راهبردهای دستیابی به نمرات بالاتر با توسل به تقلب، صرفاً سودمندی ناچیزی برای آزمودنی‌ها در برخواهد داشت (استون[10]و لونز[11]،1994؛ ویسپول[12]، هندریرکسون[13]، و بلایر[14]، 2000). بر اساس نتایج پژوهش‌ها، به نظر می‌رسد دلیل روانسنجی مناسبی وجود ندارد که در سار به آزمودنی‌ها اجازهٔ مرور و تجدید نظر در پاسخ‌هایشان داده نشود. با این وجود، اگر آزمودنی‌ها اجازهٔ مرور و تجدید نظر در پاسخ‌هایشان را داشته باشند، بر پیچیدگی‌های برنامهٔ سار افزوده می‌شود (وای، 2005).

2) **محدودیت‌های زمانی**: در نظر گرفتن محدودیت زمانی معقول و منطقی برای آزمون‌های استاندارد شدهٔ قدیمی چالش برانگیز است. اگر زمان آزمون بسیار طولانی شود، نتیجه آن از دست رفتن کارآیی آزمون است. اگر زمان آزمون بسیار کوتاه باشد،

1. Wainer
2. Baghi
3. Ferrara
4. Gabrys
5. Legg
6. Buhar
7. Vispocklin
8. Roklin
9. Wang
10. Stone
11. Lunz
12. Vispoel
13. Hendrickson
14. Bleirler

برخی از آزمودنی‌ها نمی‌توانند تمام سئوال‌های یک آزمون را پاسخ گویند و در نتیجه عملکرد آنها در آزمون، پایین‌تر از سطح عملکرد استاندارد، برآورد می‌شود. اگرچه، درنظر گرفتن محدودیت زمانی در سار پیچیده‌تر است. یک دلیل آن است که در سار از نمره دقیق به عنوان ملاکی برای پایان دادن به آزمون استفاده می‌شود، و به همان میزان که یک فرد نمی‌داند چه تعدادی سئوال را پاسخ خواهد داد، نمی‌داند که چقدر زمان برای پاسخ دادن به سئوال‌ها دارد. تصمیم‌گیری راجع به محدودیت فرصت‌های زمانی لازم در یک سار موضوعی بسیار مهم است. دلیل این موضوع آن است که تحمیل هر نوع محدودیت زمانی متضاد با برنامهٔ سنجش است که به دانش‌آموزان کمک می‌کند تا سطوح بهینه‌ای از عملکرد خود نشان دهند. محدودیت زمان همچنین می‌تواند دشواری‌هایی در مدرج‌سازی سئوال‌های آزمون به وجود آورد (اگن، 2007؛ وای، 1998؛ الیا و همکاران، 2000).

تحقیقات مرتبط

پژوهش‌هایی که در بارهٔ اثر مرور کردن سئوال‌ها بر عملکرد آزمودنی‌ها در سنجش انطباقی صورت گرفته‌اند، نشانگر آن است که: الف) زمانی که آزمودنی‌ها حق تغییر دادن و مرور کردن سئوال‌ها و پاسخ‌های خود را دارند، عملکرد آنها افزایش می‌یابد؛ (ب) پاداش نمره‌ای که در نتیجهٔ تغییر پاسخ‌ها به دست می‌آید دانش‌آموزان را تشویق می‌کند تا در بارهٔ سئوال‌ها دوباره به تفکّر پرداخته و یا دوباره آنها را بخوانند. بنابراین، از این موضوع می‌توان نتیجه‌گیری کرد که ممانعت از مرور سئوال‌ها، از بین بردن فرصت تغییر پاسخ‌هایی است که عملکرد آزمون را بهبود می‌بخشد. (ج) نبود فرصت برای مرور پاسخ‌ها اضطراب امتحان را افزایش می‌دهد. بدین معنا که نبود فرصت برای مرور سئوال‌ها به معنای فقدان کنترل آزمودنی بر محیط و آزمون است و این مطلب منجر به افزایش اضطراب امتحان می‌شود. همچنین مشخص شده که آزمودنی‌ها زمانی استرس‌های محیطی را بهتر تحمل می‌کنند که، احساس کنند بر محیط کنترل بیشتری دارند. لذا افزایش کنترل بر محیط منجر به کاهش اضطراب و بهبود عملکرد در تکلیف می‌شود (اگن، 2007؛ وای، 1998).

در یک پژوهش دو نسخه از آزمون خزانه لغات انگلیسی برای اسپانیایی زبان‌ها (یک نسخهٔ انطباقی و نسخه دیگر غیرانطباقی) در نمونه‌ای از دانشجویان اسپانیایی سال اول دانشگاه مورد استفاده قرار گرفت. اثرات نوع آزمون (آزمون انطباقی با رایانه

در برابر آزمون غیرانطباقی با رایانه) و شرایط مرور کردن (کسانی که اجازه مرور کردن سئوال و پاسخ را داشتند در مقابل کسانی که این اجازه را نداشتند) بر روی چند متغیر روان‌شناختی آزمایش شد. متغیرهای بین آزمودنی‌ها قبل و پس از مرور کردن به منظور مطالعه اثرات مرور کردن بر روی متغیرهای روان‌شناختی و روان‌سنجی برای شرایط مرور کردن در هر دو آزمون اندازه‌گیری شدند. دو دسته اصلی نتایج پس از مرور کردن بدست آمد:

الف) تعداد پاسخ‌های درست و توانایی برآورد شده افزایش معنی‌داری داشت؛ ب) میزان اضطراب موقعیتی کاهش یافت؛ ج) تفاوت خطاهای اندازه‌گیری معنی‌دار نبود. تعامل اثرات (نوع آزمون با آزادی عمل در مرور کردن) معنی‌دار نبود. پیشنهاد ضمنی این نتایج، آن است که باید در تهیه آزمون‌های انطباقی با رایانه شرایط مرور سئوال‌ها و پاسخ‌ها را برای آزمون شوندگان فراهم نمود (اولیا و همکاران، 2000).

پژوهش‌ها نشان داده‌اند که اگر آزمودنی‌ها اجازه مرور سئوال‌ها و پاسخ داشته باشند، الف) حدود 60 درصد آزمودنی‌ها حداقل یکی از پاسخ‌های خود را تغییر می‌دهند؛ ب) فقط درصد کوچکی (بین 2 تا 5 درصد) از پاسخ‌ها تغییر داده می‌شوند؛ ج) از بین پاسخ‌های تغییر داده شد، حدود 50 درصد از آنها از غلط به درست تغییر می‌یابند؛ د) از بین آزمودنی‌هایی که پاسخ خود را تغییر می‌دهند، بین 42 تا 52 درصد از آنها سطح توانایی خود را در آزمون بهبود می‌دهند. فقط درصد کوچکی از این آزمودنی‌ها (بین 10 تا 15 درصد) سطح توانایی خود را کاهش می‌دهند؛ ه) میزان ناچیزی از دقت از بین می‌رود؛ و) همبستگی بین توانایی برآورد شده قبل و پس از مرور بیشتر از 0/98 است و میانگین تفاوت‌ها بین 0/20 تا 0/07 است؛ ز) همبستگی منفی و معنی‌دار بین سطح اضطراب و توانایی وجود دارد. تعامل بین اضطراب و مرور معنی‌دار نبود؛ ح) اثر معنی‌داری در زمان آزمون دارد (مرور، زمان آزمون را بین10 تا 37 درصد افزایش می‌دهد)؛ ط) زمانی که توانایی به عنوان متغیر مستقل در نظر گرفته شود، مشخص شده است کسانی که از بالاترین سطح توانایی برخوردارند، حداقل تغییرات را در پاسخ‌های خود اعمال می‌کنند و این تغییرات اکثراً از غلط به درست و تعداد کمی از آنها از درست به غلط صورت می‌گیرد. بنابراین کسانی که از بالاترین سطح توانایی برخوردارند از بیشترین مزیت مرور بهره‌مند می‌شوند (اولیا و همکاران، 2000).

در مطالعه‌ای اولیا و همکاران (2000) نشان دادند که 81/52 درصد آزمودنی‌ها از فرصت برای مرور پاسخ‌ها استفاده کردند. در سنجش انطباقی با رایانه 80 درصد آزمودنی‌ها پاسخ‌هایشان را مرور کردند و از بین آنها 66/7 درصد توانایی برآورد شدهٔ خود را افزایش دادند، 25 درصد کاهش دادند، و 8/3 درصد تغییری در توانایی برآورد شدهٔ خود به وجود نیاوردند. در کل، 13/5 درصد پاسخ‌ها تغییر کرد. در سار 12/6 درصد پاسخ‌ها تغییر کرد و از بین آنها 42 درصد از غلط به غلط تغییر یافت، 43 درصد از غلط به درست، و 15 درصد از درست به غلط.

در ارتباط با محدویت زمانی؛ هدف عملی آن است که محدودیت زمانی به گونه‌ای مشخص شود که محدودیت اثر معنی‌دارای بر عملکرد دانش‌آموزان نداشته باشد. این موضوع با یافته‌های برخی گروه‌های اقلیت نژادی که زمان بیشتری برای پاسخگویی قایل شده‌اند، پیچیده‌تر شده است (باقی و همکاران، 1992؛ لگ[1] و بوهر[2]، بوهر[2]، 1992؛ زارا[3]، 1992). اما برخی پژوهش‌ها نشان داده‌اند که دادن زمان پاسخگویی بیشتر به دانش‌آموزان گروه‌های اقلیت در آزمون‌های سنتی عملکرد مرتبط با اکثریت دانش‌آموزان را در آنها افزایش نمی‌دهد. به نظر می‌رسد ارتباط بین محدودیت زمانی و عملکرد در آزمون با اضطراب امتحان بین آزمون‌هایی با محدودیت زمان و بدون محدودیت زمان نشان داده‌اند که این قضاوت‌ها برای آزمودنی‌های با اضطراب بالا بیشتر است (هیلی[4]، 1984؛ سیمن[5] و اونزیگ بوزی[6]، 1995؛ به نقل الیا و همکاران، 2000). این یافته‌ها پیشنهاد می‌کند که طولانی کردن محدودیت زمان در یک سار ممکن است برای برخی آزمودنی‌ها مفیدتر از سایر آزمودنی‌ها باشد. تفاوت‌های فردی پیدا شده در بین آزمودنی‌ها، مشخص می‌کند که احتمالاً مشکل است تنها محدودیت زمانی را به عنوان دفاعی منصفانه تلقی کرد. بنابراین، می‌بایست محدودیت‌های زمانی خیلی آزاد منشانه‌تری در نظر گرفت، یا اصلاً محدودیت زمانی قایل نشویم. این نکته را به خاطر داشته باشید که برنامهٔ سار خیلی کوتاه‌تر از آزمون‌های سنتی همتای آن است؛ می‌بایست در دادن وقت برای بازگشت آزمودنی‌ها دقت نمود. در آن صورت استرس وابسته به امتحان در نتیجهٔ آن،

1. Legg
2. Buher
3. Zara
4. Hilly
5. Seaman
6. Onziuegbuzie

کاهش یافته و روایی آزمون افزایش می‌یابد. با گسترش برنامه‌های آزمون در ایلات متحده آمریکا شواهدی وجود دارد که کودکان گروه‌های اقلیت و فقیر دسترس کمتری به رایانه در منزل و مدرسه دارند (ساتون[1]، 1997). بخاطر اینکه دسترسی کمتر، دلالت بر تجربهٔ کمتر دارد، ارتباط بین تجربه رایانه و عملکرد در سار اهمیت بیشتری پیدا می‌کند.

نتیجه‌گیری

1. برآورد نمرهٔ واقعی آزمودنی‌ها از طریق سنجش انطباقی، دقیق‌تر است. بنابراین، استفاده از این روش هم منطقی است و هم منصفانه. موضوعی که باید در آزمون‌سازی مؤسسات آموزشی کشور مثل دانشگاه‌ها و سازمان سنجش آموزش کشور بیشتر مورد توجه قرارگرفته و دربارهٔ آن پژوهش‌های لازم و ضروری انجام گیرد.

2. سنجش انطباقی به آزمون‌کننده اجازه می‌دهد که سطح ویژگی آزمون‌ها را به سرعت بسنجد، بدون اینکه او را وادار کند تا به مجموعه‌ای خسته‌کننده از سئوال‌های بسیار ساده یا دسته‌ای ناخوشایند و ناراحت‌کننده از پرسش‌های خیلی مشکل پاسخ دهد. بنابراین چنین روشی هم برای آزمودنی‌ها و هم برای آزمون‌کننده و آزمون‌ساز مطلوب است و باید مورد توجه مؤسسات آموزشی، بخصوص آموزش عالی کشور قرار گیرد.

3. سنجش انطباقی با رایانه که تکرار فرآیند دو مرحله‌ای در سنجش است یکی از روش‌های پیچیده‌تر سنجش است که در آن، درگام نخست سوالی به آزمودنی داده می‌شود که دشواری آن با توانایی فعلی برآورد شده برای او برابر باشد. در گام دوم توانایی آزمودنی براساس اطلاعات حاصل از پاسخ او به سوال مرحلهٔ اول برآورد شده؛ این گام‌ها تکرار می‌شوند تا جایی که ملاک‌های پایان دادن به آزمون محقق شود. با لحاظ این مطلب که در دانشگاه‌های کشور امکانات سخت‌افزاری لازم برای اجرای سنجش انطباقی با رایانه ایجاد شده است، این موضوع می‌تواند مورد توجه متخصصان آزمون‌سازی قرار گرفته و از این شیوه که هم کارآیی بیشتری دارد و هم بسیاری از سرخوردگی‌های دانشجویان در ارتباط با مشکلات آزمون‌های سنتی می‌کاهد، در دانشگاه‌ها استفاده شود.

1. Sutton

4. از آنجا که برنامه‌ریزی واقعی، اجرای عملی و حفاظت از برنامهٔ سنجش انطباقی با رایانه بسیار پیچیده است و چالش‌های عملی خاص خود را دارد، و از طرف دیگر ویژگی‌های فرهنگی، اجتماعی، و اقتصادی جامعهٔ ایرانی در این پیچیدگی‌ها و چالش‌ها مؤثرند. باید پژوهش‌های گسترده‌ای در ارتباط با پیچیدگی‌های کاربردی آن صورت گرفته و با لحاظ نتایج و یافته‌های پژوهش‌های مذکور، از این روش در عمل استفاده شود.

5. مطالعاتی به منظور امکان‌سنجی استفاده از سار در سازمان سنجش آموزش کشور و مؤسسات آموزش عالی انجام شود و براساس یافته‌های حاصل از آنها در بارهٔ استفاده از سار و سایر جوانب آن تصمیم‌گیری شود.

منابع

افروز، غلامعلی و هومن، حیدرعلی (1375). *روش تهیهٔ آزمون هوش*. تهران: دانشگاه تهران.

آلن، مری جی و ین، وندی ام (1979). *مقدمه‌ای بر نظریه‌های اندازه‌گیری (روانسنجی)*. ترجمهٔ علی دلاور، 1374، تهران: سمت.

ثرندایک، رابرت ال (1982). *روان‌سنجی کاربردی*. ترجمهٔ حیدر علی هومن، 1369، تهران: دانشگاه تهران.

ستاری، بهزاد (1382). *روان‌سنجی پیشرفته کاربردی*. مشهد: به نشر.

Baghi, H.; Ferrara, S. F., & Gabrys, R. (1992). *Student attitudes toward computer-adaptive test administrations*. Paper presented at the annual meeting of the American Educational Research Assoc, San Francisco, CA.

Colton, G. D. (1998). Exam security and high-tech cheating, *The Bar Examiner*, 67(3), 13-35.

Eggen, T. J. H. M. (1990). Innovative procedures in the calibration of measurements scales. In W.H. Schreiber & K. Ingenkamp (eds). (pp.199- 212). *International developments in large scale assessment.* Windsor, Berkshire: NFER-NELSON

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Hornke L. F. (2000). Item response times in computerized adaptive testing. *Psicologica*, 21, 157-173.

Legg, S. M. & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and practice*, 11(2), 23-27.

Lord, F.M. (1983). Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer- assisted instruction, testing and guidance*. New York: Harper & Row.

Olea J., J., Revuelta J., Ximenez M.C., & Abad F.H. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicologica*, 21, 175-189.

Sotaridonia L. S., Pornel J. B., & Vallejo A. (2003). Some applications of item response theory to testing. *The Philippine Statistician*, 52, 81-92.

Stone, G. E., & lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education*, 7, 211-222.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7,* 211-222.

Sutton, R. E. (1997). Equity and high stakes testing: Implications for computerized testing. *Equity and Excellence in Education*, 30(1), 5-15.

Tao, Y.-H., Wu, Y.-L., & Chang, H.-Y. (2008). A Practical Computer Adaptive Testing Model for Small-Scale Scenarios. *Educational Technology & Society, 11*(3), 259–274.

Triantafillou, E., Georgiadou, E., & Anastasias A. (2006). CAT-MD: Computer Adaptive Test on Mobile Devices. Economides University of Macedonia, Egnatias 156, Thessaloniki 54006, GREECE

Van der Linden, W .J. & Glas, C. A. W. (Eds.) (2000). *Computerized adaptive testing. Theory and practice.* Dordrecht: Kluwer Academic Publishers.

Van der Linden, W. J. & Hambleton, R. K. (Eds.) (1996). *Handbook of modern item response theory.* New-York: Springer-Verlag.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In. G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models: Foundations, recent developments, and applications* (pp.215-237). New York: Springer-Verlag.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computer adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37,* 21-38.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized adaptive, and self-adapted testing. *Applied Measurement in Education,* 7, 53-59.

Wainer, H. (1983). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12,* 15-20.

Wainer, H., (Ed.) (2000). *Computerized adaptive testing: A primer* (2nd Edition). Hillsdale, NJ: Erlbaum.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practices, 17* (4), 17-27.

Way, W. D. (2005). *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments.* Research Report 05-03.

Zara, A. R. (1992). *An investigation of computerized adaptive testing for demographically-diverse candidates on the national registered licensure examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.