فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی **Journal of Measurement & Educational Evaluation Studies**
سال سوم، شمارهٔ ٤، زمستان ١٣٩١، صص ١٠٤- ٧٩ **Vol. 3, No. 4, Winter 2014**

وزن‌دهی بهینه به سؤال‌ها و خرده آزمون‌های ورودی برای ساخت نمره کل ترکیبی[1]

سلیمان ذوالفقارنسب *
ابراهیم خدایی **
غلامرضا یادگارزاده ***

تاریخ دریافت مقاله: ۹۲/۰۸/۱۰
تاریخ پذیرش مقاله: ۹۲/۱۲/۰۲

چکیده

این تحقیق به منظور وزن‌دهی بهینه به خرده آزمون‌ها و سؤال‌های آزمون سراسری برای ساخت نمره کل ترکیبی انجام شده است. هدف نهایی این تحقیق پایین آوردن خطای اندازه‌گیری نمره کل ترکیبی بر اساس نظریه کلاسیک آزمون‌سازی بود. وزن‌دهی در سه سطح صورت گرفته است نخست آزمون سی سؤالی چهارگزینه‌ای حساب دیفرانسیل که نمونه آن ٣٤٠٩ نفر بود بر اساس وزن‌دهی در سطح گزینه‌های سؤال (درصد محبوبیت گزینه‌ها، نمره فرمولی) و در سطح سؤال (سرجمع ساده بدون وزن یا وزن مؤثر سؤال، وزن عاملی سؤال و وزن دشواری سؤال) وزن‌دهی شده‌اند. همچنین در سطح خرده آزمون یک مجموعه آزمون سراسری دستیاری پزشکی با پنج خرده آزمون با طول برابر شش سؤال که نمونه آن ٣٥٧٢ نفر بود نیز به روش‌های مختلف (متوسط ضریب همبستگی پیرسون، وزن عاملی و ضرایب رگرسیون β) وزن‌دهی شده‌اند. به علاوه یک مجموعه آزمون دستیاری پزشکی دیگر با طول خرده آزمون‌های نابرابر به ترتیب ٤٥، ٢٦، ٢٤، ٦ و ٦ سؤال که در بین گروه ٣٦٣٨ نفری اجرا شده بر اساس وزن مؤثر خود خرده آزمون‌ها (بدون وزن) مورد بررسی قرار گرفته است. این تحقیق نشان داد که روش نمره فرمولی بیشترین واریانس خطا را نسبت به دیگر روش‌ها تولید می‌کند. تنها وزن‌دهی بر اساس دشواری سؤال می‌تواند رتبه‌بندی افراد را به نفع افراد شایسته‌تر تغییر دهد و دیگر روش‌های وزن‌دهی برای افزایش پایایی رضایت بخش نیستند و ضریب پایایی آزمون در همان ابتدا تحت تأثیر سؤال‌های خوب و خرده آزمون‌های خوش ساخت با طول بهینه است.

واژگان کلیدی: وزن‌دهی، نمره کل ترکیبی، نظریهٔ کلاسیک آزمون‌سازی، ضریب پایایی، نمره واقعی، خطای اندازه‌گیری، نمره فرمولی.

---

۱. این مقاله خلاصه‌ای است از پروژه‌ای با همین نام، که در گروه پژوهش سنجش و اندازه‌گیری مرکز تحقیقات، ارزشیابی، اعتبارسنجی و تضمین کیفیت آموزش عالی سازمان سنجش آموزش کشور در تابستان ١٣٨٩ به پایان رسیده است.

* کارشناس ارشد پژوهشی مرکز تحقیقات، ارزشیابی، اعتبارسنجی و تضمین کیفیت آموزش عالی سازمان سنجش آموزش کشور (مسئول مکاتبات: salarnik2001@yahoo. com)

** معاون وزیر و رییس سازمان سنجش آموزش کشور- دانشیار دانشگاه تهران

*** عضو هیأت علمی سازمان سنجش آموزش کشور

مقدمه

در موقعیت‌هایی که چندین خرده آزمون اجرا می‌شود نمره‌های هر یک از آزمون‌ها ترکیب می‌شوند تا یک نمره کل مرکب را تشکیل دهند. هرگاه یک نمره کل از ترکیب چند سؤال، یا مجموعه‌ای از سؤال‌ها و یا خرده آزمون‌ها ایجاد شده باشد روشی که این سؤال‌ها با یکدیگر ترکیب شده‌اند، ویژگی‌های روانسنجی هر خرده آزمون و مؤلفه‌های تعیین‌کننده آن مثل همبستگی بین آنها، انحراف معیار هر مؤلفه و میزان دشواری هر یک از آنها برای ساخت نمره‌های ترکیبی باید مدنظر قرار گیرد. بنابر این برای ساختن نمره‌های ترکیبی، وزن‌دهی به سؤال‌ها و خرده آزمون‌های یک مجموعه آزمون اجتناب‌ناپذیر است (استانلی[۱]، وانگ[۲] ۱۹٦۸).

بسیاری از طرح‌های وزن‌دهی از نظریهٔ کلاسیک آزمون‌سازی مشتق شده‌اند و تا به حال محققان زیادی تلاش کرده‌اند تا برای افزایش پایایی، منابع مختلف خطای اندازه‌گیری و همچنین رویکردهای مختلف برآورد پایایی یک آزمون را بررسی کنند. مطالعات تجربی مربوط به وزن‌دهی به هر خرده‌آزمون، در یک مجموعه آزمون، در ابتدای قرن بیستم آغاز شد. وزن‌دهی به هر سؤال از سال ۱۹۲۵ و بعدها مطالعاتی در بارهٔ وزن‌دهی به پاسخ‌ها و یا گویه‌های هر سؤال نیز مطرح شدند (ندلسکی[۳]،۱۹۵٤). گالکسن[٤] (۱۹۵۰) مباحث مبسوطی در رابطه با وزن‌دهی بر مبنای مشخصه‌های آزمون مثل انحراف معیار، همبستگی بین خرده آزمون‌ها و پایایی انجام داده است. هدف رویکردهایی که گالکسین توصیف کرده بیشینه‌سازی پایایی نمره کل ترکیبی بوده است (گالکسن، ۱۹۵۰؛ وانگ و استنلی، ۱۹۷۰). گالکسن نشان داد که اگر تعداد زیادی آزمون با همبستگی‌های بالا با یکدیگر ترکیب شوند، نظام‌های متفاوت وزن‌دهی تفاوت خیلی زیادی در نتایج نمره کل ترکیبی حاصل نخواهند کرد؛ اما اگر تعداد کمی آزمون با یکدیگر ترکیب شوند و همبستگی بین آنها کم باشد عمل وزن‌دهی می‌تواند اثرات مهمی بر نمره کل ترکیبی داشته باشد. آلن[۵] و ین[٦] (۱۹۷۹)

---

1. Stanley
۲. Wang
۳. Nedelsky
٤. Gulliksen
5. Allen
٦. Yen

کار مشابهی را با استفاده از متغیرهای پیش‌بینی کننده در چارچوب رگرسیون چند متغیره انجام داد. واینر[۱] و تیزن[۲] (۱۹۹۳) به این نتیجه رسیدند که وقتی به خرده آزمون و مؤلفه‌های آزمون ترکیبی تاریخ اروپا که متشکل از دو بخش آزمون چهارگزینه‌ای با پایایی ۰/۹۰ و آزمون باز پاسخ با پایایی ۰/٤٦ بود وزن‌های بهینه‌ای اختصاص داده شود پایایی نمره کل آن از ۰/۸۰ به ۰/۹۰ افزایش می‌یابد. کولن[۳] (۲۰۰۷)، والکر[٤] (۲۰۰۵) و برنان[٥] (۲۰۰٤) بحث‌ها و تحقیقات زیادی در بارهٔ ترکیب خرده آزمون‌ها و وزن‌دهی به هر خرده آزمون برای ساخت یک نمره کل ترکیبی بهینه ارائه داده‌اند.

در انگلستان نیز آزمون‌های موضوعیِ مدرک عمومی آموزش متوسطه (GCSE)[٦] از چندین خرده آزمون تشکیل شده است که پس از وزن‌دهی به نمره‌های هر فرد در این خرده آزمون ها، برای ارائهٔ نمره کل با یکدیگر ترکیب می‌شوند. در چین که آزمون‌های رقابتی نظیر آزمون سراسری ایران برای پذیرش افراد برای ورود به دانشگاه اجرا می‌شود، تلاش زیادی برای ساخت یک نمره کل ترکیبی وزنی بر اساس ۵ خرده آزمون صلاحیت‌های پایه که در برگیرنده زبان چینی، زبان انگلیسی، ریاضی، علوم طبیعی و مطالعات اجتماعی است انجام می‌گیرد.

در اینجا سؤال اساسی این است که وقتی توزیع نمره‌های هر خرده آزمون در یک مجموعه آزمون دارای چولگی، کشیدگی، خطای معیار و واریانس‌های متفاوت باشد پایایی نمره کل ترکیبی په وضعیتی پیدا می‌کند؟ و حدود بهینه پایایی نمره کل ترکیبی در چه دامنه‌ای قرار می‌گیرد؟ یا اگر آزمون از مجموعه‌ای متفاوت از سؤال‌ها تشکیل شده باشد آیا روش نمره فرمولی که برای کاهش اثر حدس شانسی رسیدن به پاسخ درست یک سوم جریمه از نمره کل فرد کم می کنند می‌تواند اندازه دقیق‌تری از نمره واقعی (T) حاصل کند و به نوبه پایایی و روایی آزمون را بالا ببرد؟ چه روش وزن‌دهی بهینه‌ای در سه سطح گزینه‌های سؤال، خود سؤال و خرده آزمون‌ها می‌توان به کار برد تا بتوان نمره دقیق‌تری به دست آورد؟

---

1. Thissen

۲. Wainer

۳. Kolen

٤. Walker

٥. Brennan

٦. General Certificate of Secondary Education (GCSE)

بیان مساله

بحث مربوط به پایایی را می‌توان این گونه خلاصه کرد که نمره به دست آمده از وسیله اندازه‌گیری خاصی با چه قطعیتی شاخص نمره‌های واقعی (T) است. در نتیجه با توجه به مفروضه نظریهٔ کلاسیک یعنی $X = T + E$ می‌توان پایایی را خارج قسمت واریانس نمره واقعی (T) به واریانس نمره مشاهده شده (X) تعریف کرد:

$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2}$. یا به طور جایگزین $\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$. بعلاوه اگر آزمونی بتواند واریانس زیادی تولید کند، چنین آزمونی توانسته تغییرات بیشتری را نشان دهد و چون دامنه پراکندگی زیادی را نشان می‌دهد، توانسته افراد را بهتر تفکیک کند. مقدار واریانس کل هر خرده آزمون تابع واریانس تک تک سؤال‌ها و درجه همبستگی سؤال‌ها با یکدیگر است. بنابر این ایجاد واریانس در هر آزمون، مرهون سؤال‌های خوب است و سؤال خوب چندین معیار و ضابطه دارد که از مهم‌ترین آنها سطح بهینهٔ ضریب دشواری ۰/۵۰ و همبستگی بالای سؤال‌ها با یکدیگر است (مگنسون[۱]، ۱۹۶۶). باید خاطر نشان شود، در آزمون‌های رقابتی مانند کنکور، دشواری سؤال‌ها در دامنه‌ای بین ۰/۳۰ تا ۰/۷۰ بهترین برآوردها را برای تفکیک افراد بدست می‌دهد و واریانس بهینه‌ای را تولید می‌کند.

در چارچوب نظریهٔ کلاسیک براساس فرمول آماری پایه که در برگیرنده ترکیبی خطی از مؤلفه‌ها یا آزمون‌های X_i وزن داده شده هستند، می‌توان پایایی نمره کل ترکیبی L را مورد بررسی قرار داد (فلت[۲] و برنان ۱۹۸۹). برای نمره‌کل ترکیبی L که دربرگیرنده n خرده آزمون یا مؤلفه X_i وزن داده شده است داریم:

$$L = \sum_{i=1}^{n} w_i X_i \qquad (۱)$$

که در آن X_i = نمره مشاهده شده مربوط به خرده آزمون یا مؤلفه i، و w_i = وزن تخصیص داده شده به نمره مشاهده شده در خرده آزمون یا مؤلفه i است.

---

۱. Magnusson

۲. Feldt

فرض بر این است که خطای بین مؤلفه‌ها یا خرده آزمون ها X_i به‌طور خطی مستقل هستند و پایایی r نمره‌کل ترکیبی L را می‌توان بر اساس فرمول زیر به دست آورد (فلت و برنان، ۱۹۸۹؛ تیزن و واینر، ۲۰۰۱؛ وب و همکاران، ۲۰۰۷).

$$r_{CTT,C} = 1 - \frac{\sigma_{c,e}^2}{\sigma_c^2} = 1 - \frac{\sum_{i=1}^{n} w_i^2 \sigma_{e,x_i}^2}{\sum_{i=1}^{n} w_i^2 \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} w_i w_j \sigma_{x_i x_j}}$$

$$= \frac{\sum_{i=1}^{n} w_i^2 r_i \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j(i\neq 1)=1}^{n} w_i w_j r_{i,j} \sigma_{x_i} \sigma_{x_j}}{\sum_{i=1}^{n} w_i^2 \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j(i\neq 1)=1}^{n} w_i w_j r_{i,j} \sigma_{x_i} \sigma_{x_j}} \quad (۲)$$

که در آن r_i = پایایی مؤلفه، سؤال یا خرده آزمون i،

$\sigma_{x_i}^2$ = واریانس مؤلفه، سؤال یا خرده آزمون i،

σ_{e,x_i}^2 = واریانس خطای مؤلفه یا خرده‌آزمون i،

σ_{x_i,x_j} = کواریانس بین مؤلفه، سؤال یا خرده‌آزمون i و خرده‌آزمون j،

$r_{i,j}$ = همبستگی بین مؤلفه، سؤال یا خرده آزمون i و j،

σ_c^2 = واریانس نمره‌کل ترکیبی، $\sigma_{c,e}^2$ = واریانس خطای نمره‌کل ترکیبی.

خطای معیار اندازه‌گیری SEM مربوط به نمره کل ترکیبی برای همه مواردی که در بالا توصیف شدند را می‌توان با استفاده از معادله محاسبه کرد.

$$SEM_{CTT,C} = \sigma_C \sqrt{1 - r_{CTT,C}} \quad (۳)$$

که در آن $r_{CTT,C}$ = پایایی نمره کل ترکیبی است که با فرمول بالا محاسبه شد و σ_C انحراف معیار نمره‌کل ترکیبی L می باشد.

به هر حال این مطالعه به منظور بررسی طرح‌های مختلف وزن‌دهی برای ساخت نمره کل ترکیبی L با استفاده از داده‌های تجربی طراحی شده است. اهداف این تحقیق بررسی روش‌های وزن‌دهی نسبی بهینه به قصد ایجاد بهترین ترکیب ممکن و

همین‌طور فراهم کردن اطلاعات بیشتر در رابطه با طرح‌های مختلف نمره‌گذاری گزینه سؤال‌ها، سؤال‌ها و خرده آزمون‌ها بوده است.

سؤال‌های پژوهشی

۱-آیا وزن‌دهی به گزینه‌های سؤال‌ها بر اساس ضریب تصحیح حدس (یا نمره فرمولی)، محبوبیت گزینه‌ها و وزن عاملی هر سؤال، می‌تواند منجر به افزایش پایایی نمره آزمون شود؟

۲- آیا وزن‌دهی به هر خرده آزمون بر اساس وزن عاملی، متوسط ضریب همبستگی و ضریب رگرسیون β هر خرده آزمون با دیگر آزمون‌ها می‌تواند باعث افزایش پایایی نمره کل ترکیبی شود؟

۳- طول نابرابر خرده آزمون‌ها چه تأثیری بر پایایی نمره کل دارد؟

روش تحقیق از نوع پس رویدادی و با استفاده از داده‌های حاصل از چند خرده آزمون بوده است. داده‌های خرده آزمون سی سؤالی درس حساب دیفرانسیل از نتایج گروه ۳٤۰۹ نفری که برای آمادگی برای کنکور ۱۳۸۸ شرکت کرده بودند جمع‌آوری شده است. سؤال‌های این آزمون پیش از آن در آزمون‌های سراسری استفاده شده یا مشابه آنها بوده است. سؤال‌های خرده آزمون‌های دیگر همگی مربوط به آزمون دستیاری پزشکی سال ۱۳۸۸ بود که ۱۷ خرده آزمون داشت. به این دلیل که خرده آزمون‌ها در تعداد سؤال، تنوع زیادی داشتند (برخلاف خرده آزمون‌های سازمان سنجش که تعداد سؤال‌های برابری دارد) برای تشریح و تبیین وزن‌دهی انتخاب شد. از این خرده آزمون‌ها صرفا برای پیاده سازی روش‌های وزن‌دهی در سطوح مختلف استفاده شد و در تحلیل‌ها نوع آزمون اهمیتی نداشت. نتایج حاصل از این روش‌ها به شرط ثابت بودن شرایط برای تمام نمونه‌ها با درصدی خطا قابل تعمیم است.

در این تحقیق بر اساس نظریهٔ کلاسیک آزمون‌سازی از روش‌های وزن‌دهی زیر استفاده شد:

الف) ضریب تصحیح حدس (یا نمره فرمولی)،

ب) $1-p$ برای وزن‌دهی به گزینه‌های صحیح سؤال (p=ضریب دشواری) و از درصد محبوبیت گزینه‌های غلط q_{abc} برای وزن‌دهی به تک تک گزینه‌های غلط هر سؤال

ج) از وزن عاملی به روش مؤلفه‌های اصلی در سطح سؤال، همچنین از ماتریس‌های همبستگی، ضرایب رگرسیون β و وزن عاملی به روش مؤلفه‌های اصلی به منظور وزن‌دهی به خرده آزمون‌های دستیاری پزشکی استفاده شد. تحلیل داده‌ها نیز با نرم افزار Lertap (نلسون[۱]، ۲۰۰۷) صورت گرفته است.

نتایج و داده‌ها

در بخش نتایج این تحقیق به تفکیک برونداد حاصل از روش‌های مختلف وزن‌دهی و تحلیل‌های حاصل از هر آزمون به ترتیب در بخش‌های الف تا ج آمده است.

الف) آزمون سی سؤالی حساب دیفرانسیل: در جدول (۱) نتایج حاصل از روش‌های مختلف بدون وزن، وزن‌دهی به سؤال (بر اساس پاسخ صحیح $1-p$ و بر اساس وزن عاملی سؤال) و همچنین وزن‌دهی به گزینه‌های سؤال (ضریب تصحیح حدس و وزن‌دهی بر اساس درصد محبوبیت گزینه‌های غلط q_{abc} و گزینه صحیح $1-p$) آمده است.

جدول (۱) وزن‌دهی در سطح سؤال و گزینه سؤال و توزیع آماره‌های آزمون ۳۰ سؤالی

شاخص آماری	بدون وزن (۰و۱)	$1-p$	وزن عاملی	ضریب تصحیح حدس	وزن دهی بر اساس محبوبیت گزینه‌های غلط q_{abc} و گزینه صحیح $1-p$
حجم نمونه آزمون دهندگان	۳٤۰۹	۳٤۰۹	۳٤۰۹	۳٤۰۹	۳٤۰۹
تعداد سؤال	۳۰	۳۰	۳۰	۳۰	۳۰
واریانس	٤٦/٤۸	۲۱/۱۲	۱۳/۰۷	٤۳٤/٦٦	۲۱/۲۱
پایایی (آلفا)	۰/۹۰٤۲	۰/۹۰۲۸	۰/۹۰٥٤	۰/۸۷٤۹	۰/۹۱
خطای معیار اندازه‌گیری	۲/۱۱ (٪۷)	۱/٤۳ (٦/۹٪)	۱/۱۱ (۷/۲٪)	۷/۳۷ (۸/۲٪)	۱/۳۷ (٦/٦٪)

۱. Nelson

در ستون دوم جدول (۱) بر اساس روش ۰ و ۱ یا بدون وزن واریانس خرده آزمون ۳۰ سؤالی برابر با ٤٦/٤٨ و پایایی با این روش برابر با ۰/۹۰٤۲ است. خطای معیار اندازه‌گیری نمره‌های مشاهده شده X برابر با ۲/۱۱ می‌باشد که این مقدار خطا تقریباً ۷ درصد واریانس هر نمره مشاهده شده X را به خود اختصاص می‌دهد.

در ستون سوم سؤال‌ها بر اساس یک منهای ضریب دشواری[۱] $1-p$ وزن‌دهی شده‌اند. واریانس برابر با ۲۱/۱۲ می‌باشد و پایایی بر اساس این روش برابر با ۰/۹۰۲۸ است. خطای معیار اندازه‌گیری برابر با ۱/٤۳ می‌باشد که این مقدار خطا تقریباً ٦/۹ درصد واریانس هر نمره مشاهده شده X وزن داده شده را به خود اختصاص می‌دهد.

در ستون چهارم وقتی که سؤال‌ها بر اساس وزن عاملی آنها روی بردار اول وزن‌دهی شده‌اند آمده است. واریانس نمرات برابر با ۱۳/۰۷و پایایی بر اساس این روش برابر با ۰/۹۰٥٤ است. خطای معیار اندازه‌گیری برابر با ۱/۱۱ است که این مقدار خطا تقریباً ۷/۲ درصد واریانس هر نمره مشاهده شده X را به خود اختصاص می‌دهد.

در ستون پنجم وقتی که گزینه‌های سؤال‌ها بر اساس فرمول تصحیح حدس (نمره فرمولی) وزن‌دهی می‌شوند آمده است. برای پاسخ به هر گزینه غلط یک نمرهٔ منفی (۱-) و برای هر پاسخ صحیح سه نمره در نظر گرفته شده است. واریانس نمرات برابر با ٤۳٤/٦٦ و پایایی خرده آزمون با این روش ۰/۸۷٤۹ است. خطای معیار اندازه‌گیری برابر با ۷/۳۷ است که تقریباً ۸/۲ درصد واریانس هر نمره مشاهده شده X را به خود اختصاص می‌دهد. باید توجه داشت که ارزش پایایی به دست آمده با این روش کمتر از روش‌های دیگر است.

در ستون ششم خلاصه آماری توزیع نمره‌های مشاهده شده X هنگامی که وزن‌دهی بر اساس نسبت‌های محبوبیت سه گزینه غلط q_{abc} یعنی درصد پاسخ به گزینه‌های غلط و $1-p$ به گزینه صحیح صورت گرفته آمده است. این روش وزن‌دهی مشابه با نمره فرمولی یا ضریب تصحیح حدس است که به پاسخ صحیح سه نمره و به پاسخ به گزینه‌های غلط یک نمرهٔ منفی تعلق می‌گیرد، با این تفاوت که به

---

۱. در واقع، ضریب دشواری به تفسیر دیگر ضریب آسانی یک سؤال نیز تعریف کرده‌اند. چون نسبت کسانی است که به سؤال پاسخ صحیح داده‌اند به کل گروه. همچنین در متون و ادبیات پیشین تئوری کلاسیک به درصد پاسخگویی به یک سؤال و یا گزینه‌های یک سؤال محبوبیت سؤال یا محبوبیت گزینه‌های سؤال نیز گفته‌اند.

جای۳ امتیاز به گزینه صحیح یک منهای ضریب دشواری $p-1$ آن قرار می‌گیرد و مبنای امتیازدهی به گزینه‌های غلط به جای یک نمره منفی، درصد محبوبیت q_{abc} آنها در بین نمونه آزمون دهندگان است (به پیوست ۱ نگاه کنید). با این روش واریانس نمره‌ها برابر با ۲۱/۲۱ و پایایی برابر با ۰/۹۱است. خطای معیار اندازه‌گیری نمره‌ها ۱/۳۷ است که ٦/٦ درصد واریانس هر نمره مشاهده شده X را به خود اختصاص می‌دهد. پایایی بر اساس این روش وزن‌دهی به بیشینه اندازه خود یعنی ۰/۹۱ افزایش یافته است.

ب) پنج خرده *آزمون با طول نابرابر:* در ادامه جدول (۲) توزیع آماری پنج خرده آزمون با طول‌های مختلف است. در این مرحله هیچ نوع وزن تجربی و یا اسمی برای خرده آزمون‌ها در نظر گرفته نشده و تنها وزن مؤثر آنها که ناشی از کیفیت سؤال (واریانس) و کمیت (تعداد) سؤال‌های هر خرده آزمون است تعیین کننده وزن آنها در مجموعه آزمون بوده است. همان‌طور که قبلا نیز بیان شد وزن مؤثر هر خرده آزمون ناشی از سه عامل است. الف) طول آزمون، ب) واریانس سؤال‌های pq که تحت تأثیر مقدار ضریب دشواری است و مقدار بهینه آن ۰/۵۰ است و به نوبه خود واریانس نمره‌های مشاهده شده X را در هر خرده آزمون متاثر می‌سازد و در حالت ج) همبستگی‌های بین هر خرده آزمون X_i با دیگر خرده آزمون‌ها که تعیین کننده مقدار واریانس کل است. الگوی بیان شده در ج) برآورد واریانس $\rho^{2}_{xx'}$ بر اساس آنچه که در نظریهٔ کلاسیک آزمون‌سازی تحت عنوان همبستگی خطی بین فرم‌های آزمون‌های موازی و همتا بیان می‌شود در این تحقیق مورد توجه قرار گرفته است. رایج‌ترین تفسیر ضریب همبستگی بر اساس مجذور آن انجام می‌شود. مجذور ضریب همبستگی برابر است با نسبت واریانسی از X که با همبستگی خطی با X' برآورد می‌شود با توجه به شش روش بسط و تفسیر ضریب پایایی در نظریهٔ کلاسیک آزمون‌سازی، همبستگی بین نمره آزمون‌های موازی با سمت چپ این فرمول $\rho_{xx'} = \frac{\sigma^{2}_{T}}{\sigma^{2}_{X}} \Rightarrow \rho^{2}_{xx'}$ مشخص می‌شود. برای این که نسبت این معادله یک، یعنی همبستگی کامل باشد، همه واریانس‌ها باید ناشی از تفاوت‌های واقعی بین افراد باشد و نه خطا. یعنی نسبت نمره

واقعی به نمره مشاهده شده $\frac{\sigma_T^2}{\sigma_X^2}$ برابر با یک باشد. به همین دلیل در سطح سؤال باید سؤال‌ها یک ویژگی یا خصیصه را اندازه بگیرند تا همبستگی بین آنها بیشینه شود. این روش در نظریهٔ کلاسیک تحت عنوان پایایی مبتنی بر همسانی درونی[۱] عنوان می‌شود. چنین قاعده‌ای در سطح خرده آزمون‌ها نیز برقرار است. هنگامی واریانس واقعی نمره کل ترکیبی حاصل از خرده آزمون‌ها، بیشینه است (و به دنبال آن پایایی افزایش می‌یابد) که این خرده آزمون‌ها از همبستگی بهینه‌ای با یکدیگر برخوردار باشند و ویژگی مشابهی را اندازه بگیرند. به این ترتیب واریانس خطا به همان شیوه همسانی درونی کاهش می‌یابد (وب، شیولسون و هارتل، ۲۰۰۶).

جدول (۲) توزیع آماره‌های پنج خرده آزمون با طول‌های نابرابر

شاخص آماری	آزمون ۱	آزمون ۲	آزمون ۳	آزمون ٤	آزمون ٥	کل آزمون
حجم نمونه آزمون دهندگان	۳٦۳۸	۳٦۳۸	۳٦۳۸	۳٦۳۸	۳٦۳۸	۳٦۳۸
تعداد سؤال	٤٥	۲٦	۲٤	٦	٦	۱۰۷
واریانس	٥٥/۰٥	۱٤/۱۹	۱٤/۲٤	۲/۸٥	۱/۲٤	۲۳۷/٥۳
پایایی (آلفا)	۰/۸٥	۰/٦٥	۰/۷۰	۰/٥۹	۰/٤٤	
خطای معیار اندازه‌گیری	۲/۸۷ (٦/٤٪)	۲/۲۲ (۸/٥٪)	۲/۰۸ (۸/۷٪)	۱/۰۸ (۱۷/۹٪)	۰/۸۳ (۱۳/۹٪)	

بر اساس جدول (۲) دامنه بلند آزمون ٤٥ سؤالی باعث شده که واریانس آن ٥٥/۰٥ باشد که تقریباً چهار برابر آزمون‌های ۲٦ سؤالی و ۲٤ سؤالی است. بنابر این بیشترین پایایی ۰/۸٥ را به خود اختصاص داده است و به عبارتی می‌توان گفت که ۸٥ درصد واریانس برآورد شده حاصل از این خرده آزمون ناشی از واریانس واقعی است. خرده آزمون پنجم به علت تعداد کم سؤال‌ها و بی‌کیفیت بودن آنها کمینه واریانس ۱/۲٤ را تولید کرده است. ضریب پایایی این خرده آزمون برابر با ۰/٤٤ است. در ستون دوم به بعد در جدول (۲) آماره‌های مربوط به هر خرده آزمون آمده است. در این مجموعه آزمون، مقدار واریانس کل حاصل از نمره کل ترکیبی برابر با ۲۳۷/٥۳ است.

۱. Internal-consistency reliability

جدول (۳) ماتریس همبستگی پنج خرده آزمون با طول نابرابر

همبستگی	آزمون ۱	آزمون ۲	آزمون ۳	آزمون ٤	آزمون ٥	نمره کل ترکیبی L
آزمون ۱	۱	۰/۷۰	۰/۷۳	۰/٦٦	۰/٥۹	۰/۹٥
آزمون ۲	۰/۷۰	۱	۰/٦٤	۰/٥٤	۰/٥۱	۰/۸٤
آزمون ۳	۰/۷۳	۰/٦٤	۱	۰/٦۱	۰/٥٤	۰/۸٦
آزمون ٤	۰/٦٦	۰/٥٤	۰/٦۱	۱	۰/٤۹	۰/۷٤
آزمون ٥	۰/٥۹	۰/٥۱	۰/٥٤	۰/٤۹	۱	۰/٦٦
نمره کل ترکیبی L	۰/۹٥	۰/۸٤	۰/۸٦	۰/۷٤	۰/٦٦	۱
میانگین همبستگی	**۰/۷۳**	**۰/٦٥**	**۰/٦۸**	**۰/٦۱**	**۰/٥٦**	**۰/۸۱**

در جدول (۳) ماتریس همبستگی هر خرده آزمون با دیگر آزمون‌ها، همبستگی یک خرده آزمون X_i با نمره کل ترکیبی L که با محاسبهٔ ضریب همبستگی دو متغیره مربوط به نمره خام خرده آزمون X_i با نمره کل ترکیبی L به دست می‌آید (آخرین ستون) و همچنین میانگین همبستگی هر خرده آزمون X_i با مجموعه[۱] آزمون (آخرین سطر) آمده است. لازم به توضیح است که میانگین وزن هر خرده آزمون از میانگین ضریب همبستگی آن با خرده آزمون‌های دیگر بدست می‌آید. بدین صورت که تحت مفروضه نرمال بودن، ماتریسی از ضرایب همبستگی خرده آزمون‌ها تشکیل می‌شود و با استفاده از سرجمع سادهٔ ضرایب هر ستون ماتریس ضرایب همبستگی (منهای عناصر قطری) در صورت کسر و در مخرج تقسیم بر n یعنی تعداد خرده آزمون‌ها، میانگین وزن هر خرده آزمون در مجموعه آزمون به دست می‌آید (نلسون، ۲۰۰۷). به عنوان مثال برای ستون دوم جدول (۳) داریم (۰/۷۰+۰/۷۳+۰/٦٦+۰/٥۹+۰/۹٥)/٥=۰/۷۳. بر اساس این جدول متوسط همبستگی خرده آزمون ٤٥ سؤالی با دیگر خرده آزمون‌ها ۰/۷۳ می‌باشد و به علت دامنه بلند، این آزمون بیشترین همبستگی را با دیگر خرده آزمون‌ها دارد.

جدول (٤) واریانس خرده آزمون هایی با طول نابرابر و پایایی نمره کل L بدون وزن

خرده آزمون‌ها	وزن	تعداد سؤال	واریانس
آزمون۱	۱	٤٥	٥٥/۰٥
آزمون۲	۱	۲٦	۱٤/۱۹
آزمون۳	۱	۲٤	۱٤/۲٤
آزمون٤	۱	٦	۲/۸٥

۱. Total

آزمون۵	۱	٦	۱/۲٤
واریانس کل	---	---	۲۳۷/۵۳
پایایی نمره کل ترکیبی L= ۰/۷۸۹۱			

بر اساس جدول (٤) ضریب پایایی نمره کل ترکیبی L غیروزنی این پنج خرده آزمون که بر اساس میانگین همبستگی هر خرده آزمون با دیگر آزمون‌ها محاسبه شده برابر با ۰/۷۸۹۱ می‌باشد. می‌توان گفت تقریباً ۷۹ درصد (گرد شده) واریانس نمره کل ترکیبی L این مجموعه آزمون، واریانس واقعی و تقریباً ۲۱ درصد آن واریانس خطا می‌باشد. باتوجه به این که میانگین ضریب همبستگی خرده آزمون‌هایی با دامنه بلندتر بیشتر از خرده آزمون‌هایی با دامنه کوتاه‌تر است، در چنین مواقعی تعداد بیشتر سؤال‌ها نقش وزن مؤثر خرده آزمون را در پایایی نمره کل ترکیبیL بازی می‌کند.

ج) پنج خرده *آزمون با طول برابر:* در جدول‌های زیر خلاصه آماری پنج خرده آزمون با طول برابر (شش سؤال) آمده است. تعداد نمونه برابر با ۳۵۷۲ نفر است. دلیل انتخاب پنج آزمون با طول برابر این بود که بتوانیم با معیاری ثابت (طول آزمون) روش‌های مختلف وزن‌دهی را در آزمون‌های کوتاه با ضرایب پایایی مختلف نشان دهیم و نقش وزن‌دهی پیشین را در میان آنها بررسی کنیم. واریانس، پایایی و مقدار خطای معیار اندازه‌گیری (به همراه درصد این خطا) در هر خرده آزمون آمده است. در این مرحله بر اساس سه روش وزن‌دهی که عبارتند از وزن برابر یک و یا به عبارت بهتر وزن مؤثر هر خرده آزمون، وزن‌دهی بر مبنای میانگین همبستگی هر خرده آزمون با چهار خرده آزمون دیگر، وزن عاملی به روش مؤلفه‌های اصلی و در نهایت ضرایب رگرسیون β بر اساس مدل Inter آمده است[1].

جدول (۵) توزیع آماره‌های پنج خرده آزمون شش سؤالی با طول‌های برابر

شاخص آماری	آزمون ۱	آزمون ۲	آزمون ۳	آزمون ٤	آزمون ۵	کل آزمون
حجم نمونه آزمون دهندگان	۳۵۷۲	۳۵۷۲	۳۵۷۲	۳۵۷۲	۳۵۷۲	۳۵۷۲
تعداد سؤال	٦	٦	٦	٦	٦	۳۰
واریانس	۲/۰٦	۲/۷۲	۱/۱۰	۱/٤۷	۱/۷۱	۱/۸۳

۱. به این دلیل از روش Inter استفاده شد که سهم تمام متغیرها (یعنی خرده آزمون‌ها) در معادله رگرسیون بدون اینکه حذف شوند مشخص شود.

پایایی (آلفا)	۰/۳۵	۰/۵۷	۰/۳۷	۰/۲۳	۰/٤۲	
خطای معیار اندازه گیری	۱/۱٦ (۱۹/۳٪)	۱/۰۹ (۱۸/۱٪)	۰/۸۳ (۱۳/۸٪)	۱/۰٦ (۱۷/۷٪)	۰/۹۹ (۱٦/۵٪)	

همان‌طور که در جدول (۵) می‌بینیم اگر چه طول هر پنج خرده آزمون برابر است اما واریانس (۲/۷۲) خرده آزمون ۲ و ضریب پایایی (۰/۵۷) آن بیشتر از چهار خرده آزمون دیگر است و عملا نقش بیشتری در تفکیک افراد دارد.

جدول (٦) ماتریس همبستگی پنج خرده آزمون با طول برابر

همبستگی	آزمون ۱	آزمون ۲	آزمون ۳	آزمون ٤	آزمون ۵	نمره کل ترکیبی L
آزمون ۱	۱	۰/۱۷	۰/۱۲	۰/۲۰	۰/۰۹	۰/۵۲
آزمون ۲	۰/۱۷	۱	۰/٤٤	۰/۳۳	۰/٤۷	۰/۸۱
آزمون ۳	۰/۱۲	۰/٤٤	۱	۰/۲۵	۰/۳۹	۰/٦۰
آزمون ٤	۰/۲۰	۰/۳۳	۰/۲۵	۱	۰/۲۵	۰/۵۹
آزمون ۵	۰/۰۹	۰/٤۷	۰/۳۹	۰/۲۵	۱	۰/٦۷
نمره کل ترکیبی L	۰/۵۲	۰/۸۱	۰/٦۰	۰/۵۹	۰/٦۷	۱
میانگین همبستگی	**۰/۲۲**	**۰/٤٤**	**۰/۳٦**	**۰/۳۲**	**۰/۳۷**	**۰/٦٤**

در جدول (٦) ارزش همبستگی هر خرده آزمون با دیگر آزمون ها، همبستگی یک خرده آزمون با نمره کل ترکیبی (آخرین ستون) و میانگین همبستگی هر خرده آزمون با مجموعه آزمون (آخرین سطر) آمده است. خرده آزمون (۲) به طور متوسط بیشترین همبستگی (۰/٤٤) با خرده آزمون‌های دیگر و با کل آزمون (۰/۸۱) دارد.

جدول (۷) وزن‌ها و واریانس‌های مربوط به خرده آزمون هایی با طول برابر شش سؤال و پایایی نمره کل ترکیبی L

خرده آزمون‌ها	وزن: ۱	واریانس	وزن: میانگین همبستگی	واریانس	وزن: بار عاملی	واریانس	وزن: ضریب رگرسیون β	واریانس
آزمون ۱	**۱**	۲/۰٦	**۰/۲۲**	۲/۰٦	**۱/٤۲**	۲/۰٦	**۰/۳۲٦**	۲/۰٦
آزمون ۲	**۱**	۲/۷۲	**۰/٤٤**	۲/۷۲	**۱/۹۲**	۲/۷۲	**۰/۳۷٤**	۲/۷۲
آزمون ۳	**۱**	۱/۱۰	**۰/۳۷**	۱/۱۰	**۱/۵۱**	۱/۱۰	**۰/۲۳۹**	۱/۱۰
آزمون ٤	**۱**	۱/٤۷	**۰/۳۳**	۱/٤۷	**۱/۲۷**	۱/٤۷	**۰/۲۷۸**	۱/٤۷

آزمون ۵	۱	۱/۷۱	۰/۳۸	۱/۷۱	۲/۱۰	۱/۷۱	۰/۲۹۹	۱/۷۱
کل		۱۸/٦۲		۲/٤۸		٥٤/۳٥		۱/۸۳
پایایی نمره کل ترکیبی L	۰/٦٤۱٦		۰/٦٥۲۳		۰/٦۳۸٤		۰/٦۱۷۰	

در جدول (۷) نتایج حاصل از روش‌های مختلف وزن‌دهی به خرده آزمون‌ها با طول برابر آمده است. ستون‌های سمت راست وزن مربوط به خرده آزمون است و ستون سمت چپ آن مقدار واریانس حاصل از روش وزن‌دهی می‌باشد و در سطر پایین جدول، پایایی حاصل از این نوع روش‌های وزن‌دهی آمده است. روش میانگین همبستگی بیشترین مقدار ضریب پایایی (۰/٦٥۲۳) را به دست داده است. باید خاطر نشان کرد که در خرده آزمون‌های با طول برابر اگر دو نفر در دو آزمون مختلف ۱ و ۲ به نسبت مساوی به سؤال‌ها پاسخ دهند هر دو امتیازی مشابه دریافت می‌کنند. اما ارزش نمره‌ای که در خرده آزمون (۲) به دست می‌آید بیشتر از خرده آزمون (۱) است. این درحالی است که بر اساس دو راه مختلف به یک نمره واحد رسیده‌اند. در آزمون‌هایی که خرده مقیاس‌های با طول برابر دارند بهتراست این نابرابری‌ها به حداقل برسد. بدین منظور پیشنهادهایی ارائه شده است.

نتیجه‌گیری

در این تحقیق، روش‌های مختلف وزن‌دهی با سه رویکرد کلی، یعنی در سطح محبوبیت گزینه‌های سؤال، سطح سؤال و سطح خرده آزمون‌ها با طول‌های مختلف، مورد بررسی قرار گرفت. آنچه مسلم است این که وقتی نمره کل افراد را به صورت حاصل جمع ساده نمره‌های آزمون‌های فرعی و یا بر اساس نمره فرمولی که در آن تصحیح برای عامل حدس صورت می‌گیرد حساب می‌کنیم نقش هر یک از سؤال‌ها و یا خرده آزمون‌ها در تعیین نمره کل یکسان نخواهد بود. معمولاً سؤال‌هایی با ضریب دشواری نزدیک به ۰/۵۰ و خرده آزمون‌هایی با تعداد سؤال‌های بیشتر چون واریانس بیشتری تولید می‌کنند و پایایی آنها همراه با افزایش طول آزمون افزایش پیدا می‌کند معیارهای بهتری برای اندازه‌گیری دقیق توانایی افراد دارند. همچنین نسبت به آزمون‌های کوتاه پایایی بهتری دارند.

در پاسخ به سؤال یک این تحقیق که، آیا وزن‌دهی به گزینه‌های سؤال‌ها بر اساس ضریب تصحیح حدس، بر اساس محبوبیت گزینه‌ها و بر اساس وزن عاملی هر سؤال می‌تواند منجر به افزایش پایایی آزمون شود؟ همان طور که از نتایج آزمون سی سؤالی مشاهده می‌شود در سطح گزینه‌های سؤال، بدون شک هیچ نوع روش وزن‌دهی برای افزایش پایایی کارایی چندانی ندارد. اگرچه وزن‌دهی بر اساس درصد محبوبیت گزینه‌های غلط q_{abc} و گزینه صحیح $1-p$ نسبت به دیگر روش‌های وزن‌دهی به گزینه‌ها، پایایی را یک درصد افزایش می‌دهد (طبق ستون ششم جدول (۱) پایایی برابر با ۰/۹۱ می‌باشد) اما نسبت به هزینه و زمان چنین روش وزن‌دهی به صرفه نیست. بررسی پیشینه و ادبیات نتایج تحقیقات مربوط به این روش وزن‌دهی، نشان می‌دهد که این نوع روش وزن‌دهی پیچیده و پرزحمت است و نتیجه‌ای که از آن حاصل می‌شود چشمگیر نیست (استانلی، ۱۹٦۷). در تحقیقات ندلسکی (به نقل از استانلی، ۱۹٦۷) نشان داد که روش‌های وزن‌دهی به گزینه‌های سؤال به روش‌های گوناگون باعث افزایش چشمگیر پایایی نمره‌های ۱۵ درصد افراد در پایین توزیع نمرات می‌شود و پایایی نمره‌های بالای توزیع، تغییر چندانی نمی‌یابد. با این حال وزن‌دهی بر اساس درصد محبوبیت هر گزینه انحرافی به این دلیل ارزشمند است که آزمون دهنده بین پاسخ‌های انحرافی نزدیک به پاسخ صحیح، توانسته است تفاوت قایل شود و به خاطر آن امتیاز دریافت می‌کند چنین فرایندی باعث افزایش پایایی آزمون می‌شود(استانلی، ۱۹٦۷). بیشترین مقدار پایایی در این تحقیق نیز بر اساس این نوع وزن‌دهی به دست آمد. این روش وزن‌دهی به گزینه‌های هر سؤال با مطرح شدن ضریب تصحیح حدس، عملاً کنار زده شد و تحقیقات کمتری بر اساس آن صورت گرفت. هرچند که وزن‌دهی بر اساس تصحیح حدس یا نمره فرمولی باعث کاهش ۳ درصد پایایی نیز می‌شود (طبق ستون پنجم جدول (۱) پایایی برابر با ۰/۸۷٤۹ می‌باشد). در وزن‌دهی به شیوه تصحیح حدس، مهم نیست که فرد به کدام یک از گزینه های غلط سؤال پاسخ دهد در هر حال یک نمره منفی (۱-) دریافت می‌کند و این باعث ساده‌تر شدن نمره‌گذاری گزینه‌های سؤال نسبت به روش وزن دهی بر اساس درصد محبوبیت هر گزینه انحرافی می‌شود.

در سطح سؤال باید گفت که سؤال‌هایی که دشواری آنها نزدیک به یک یا صفر است به علت عدم توانایی در تمایز افراد، کمترین کارایی را دارند. سطح بهینه دشواری برای سؤال چهار گزینه ای در مواقعی که فرمول تصحیح حدس به کار

می‌رود ۰/٦۲۵ است که نه تنها در عمل تفاوت چندانی در واریانس واقعی آزمون ایجاد نمی‌کند، بلکه به کارگیری فرمول تصحیح حدس در بیشتر مواقع باعث افزودن خطای اندازه‌گیری E می‌شود و پایایی نمره مشاهده شده X را نیز کاهش می‌دهد. اگر چه واریانس نمره مشاهده شده بسیار افزایش می‌یابد جدول (۱). اما بالا بودن مقدار واریانس نمره مشاهده شده X فی نفسه مورد توجه نیست بلکه مهم این است که آزمون بتواند نشان دهد که تمایز حاصل از اجرای آزمون در بین افراد دقیق، معنی‌دار و پایاست (مگنسون، ۱۹٦۷). *نتیجه به دست آمده در سطح گزینه‌های سؤال و سؤال به سطح خرده آزمون‌ها نیز قابل تعمیم است. در ادامه به بررسی وزن‌دهی در سطح خرده آزمون می‌پردازیم.*

در پاسخ به سؤال دوم این تحقیق که آیا وزن‌دهی به هر خرده آزمون بر اساس وزن عاملی، میانگین ضریب همبستگی هر خرده آزمون با دیگر آزمون‌ها و بر اساس ضریب رگرسیون β می‌تواند باعث افزایش پایایی نمره کل ترکیبی شود؟ این نتیجه به دست آمد که چه *خرده آزمون‌هایی با طول برابر و چه خرده آزمون هایی با طول نابرابر اگر یک مجموعه آزمون را تشکیل دهند هیچ یک از روش‌های وزن‌دهی باعث افزایش رضایت‌بخش پایایی نمره کل ترکیبی نخواهد شد و نمره کل ترکیبی تحت کنترل واریانس خرده آزمون‌های خوش ساخت و با طول بلندتر قرار می‌گیرد. این نتایج را می‌توان در سطر پایین جدول* (۷) نیز *مشاهده کرد. نتایج تحقیقات هندریکسون*[۱] *و همکاران (۲۰۱۰) نشان می‌دهد که در آزمون جایابی پیشرفته*[۲] *که از ترکیب دو بخش یعنی سؤال‌های چندگزینه‌ای و سؤال‌های کوتاه پاسخ تشکیل شده‌اند وزن‌دهی باعث افزایش پایایی بخش چندگزینه‌ای این آزمون از ۰/۸۵٦ به ۰/۸۸۵ می‌شود. اگر چه این مقدار ۰/۰۲۹ افزایش پایایی نسبت به افزایش پایایی بخش کوتاه پاسخ قابل ملاحظه نیست، اما آنها بیان می‌کنند که همین مقدار افزایش پایایی ۰/۰۲۹ معادل با افزایش ۱۹سؤال به سؤال‌های فعلی آزمون چند گزینه‌ای می‌باشد. نکته قابل توجه این است که وقتی آزمونی از ترکیب دو بخش چند گزینه‌ای و تشریحی یا کوتاه پاسخ است، عمل وزن‌دهی به بخش‌های مختلف خرده آزمون تشریحی و آزمون چندگزینه‌ای باعث افزایش پایایی نمره کل ترکیبی می‌شود و وزن‌دهی در این نوع آزمون‌ها مؤثرتر است (هندریکسون و همکاران ۲۰۱۰و برنان، ۲۰۰٤).*

---

۱. Hendrickson

۲. Advanced Placement Program (AP) Exams

در پاسخ به سؤال سوم این تحقیق که طول نابرابر خرده آزمون‌ها چه تأثیری بر پایایی نمره کل دارد؟

همان طور که از نتایج مندرج در جدول (۲) پیداست، پایایی نمره کل ترکیبی حاصل از ترکیب خرده آزمون‌های نابرابر شدیداً تحت تأثیر آزمون بلندتر قرار می‌گیرد. به عبارت دیگر آزمون ٤٥ سؤالی، واریانسی برابر با ۵۵/۰۵ دارد و پایایی آن ۰/۸۵ است در حالی که آزمون شش سؤالی واریانسی برابر با ۱/۲٤ دارد و پایایی آن ۰/٤٤ است. آزمون‌های بلندتر در ترکیب با آزمون‌های کوتاه‌تر با برآورد واریانس بیشتر در مجموعه آزمون تأثیر خود را روی پایایی اعمال می‌کنند. این کارکرد آزمون‌های بلندتر در ترکیب با آزمون‌های کوتاه‌تر، زمانی مشکل‌زا می‌شود که برای مشخص کردن ضریب درس‌های امتحانی برای انتخاب افراد، طول این خرده آزمون‌ها در نظر گرفته نشود و آزمون کوتاه همان ضریب آزمون بلند را داشته باشد و یا برای ساخت نمره کل ترکیبی اثر حذف یک یا چند سؤال از خرده آزمون‌های با طول برابر نادیده گرفته شود. با این سهل انگاری عملاً وزنه به نفع آزمون بلندتر تغییر می‌یابد. همچنین، چون پایایی در چنین ترکیبی از خرده آزمون‌های بلند و کوتاه، تحت تأثیر وزن مؤثر خرده آزمون‌های با طول بلندتر است، در صورت وزن‌دهی تأثیر زیادی در نتایج پایایی ترکیب آزمون‌ها رخ نخواهد داد. به علاوه در تحقیقات استانلی و وانگ (۱۹٦۸) نشان می‌دهند که در صورت وزن‌دهی به ترکیب خرده آزمون‌های بلند و کوتاه، عملاً پایایی نمره کل ترکیبی به ارزشی بیشتر از پایاترین خرده آزمون در مجموعه آزمون افزایش نمی‌یابد. این نتایج با تحقیق حاضر یکسان است به طوری که در تمام شیوه‌های وزن‌دهی، پایایی نمره کل ترکیبی از کمینه و بیشینه پایایی خرده آزمون‌ها فراتر نمی‌رود و بین این دو قرار می‌گیرد. همچنین چایلدز[1] و همکاران (۲۰۰٤) در تحقیق خود نشان می‌دهند که در پیشینه پژوهشی مبتنی بر وزن‌دهی به خرده آزمون‌های با طول‌های مختلف عملا وزن‌دهی اثری بر کاهش خطای معیار ندارد و یا بسیار اندک است.

در واقع آنچه وزن‌دهی انجام می‌دهد این است که باعث کاهش مشارکت خرده آزمون با پایایی کمتر در نمره کل ترکیبی می‌شود و این نقش را به خرده آزمون‌هایی که بیشترین پایایی را دارند واگذار می‌کند. از طرف دیگر اگر بتوان با تدابیری مثل افزایش طول آزمون، پایایی نامطلوب یک آزمون را افزایش داد، پایایی نمره کل

۱. Childs

ترکیبی و همین طور روایی بدون نیاز به وزن‌دهی به طور خودکار افزایش می‌یابد (استانلی، وانگ، ۱۹٦۸).

همان‌طور که از جدول (۲) پیداست، خرده آزمون ٤٥ سؤالی در مجموعه آزمون بیشترین پایایی (۰/۸۵) را داشت و وزن مؤثر آن باعث کاهش خطای اندازه‌گیری و افزایش پایایی آن شده بود. این وضعیت به ترتیب برای دیگر آزمون‌های ۲۵ و ۲۳ سؤالی نیز حاکم است. تنها در یک موقعیت است که آزمون‌های بلند کارایی کمی دارند و آن در مواقعی است که سؤال‌ها به خوبی تهیه نشده باشند و گزینه‌های انحرافی آنها بیش از گزینه صحیح محبوبیت داشته باشند و به عبارتی گزینه‌های غلط موجب انحراف گروه قوی شده باشد. البته باید اضافه شود که طول آزمون از یک سطح بهینه که عبور می‌کند، افزایش معنی‌داری در واریانس و یا پایایی ایجاد نمی‌شود و به همین خاطر روش‌هایی نیز برای مشخص کردن طول بهینهٔ خرده آزمون‌ها وجود دارد. باید توجه داشت در حالتی که خرده آزمون هایی با طول برابر، یک مجموعه آزمون را تشکیل می‌دهند، اگر بنا به دلایلی یک یا چند سؤال از یکی از خرده آزمون‌ها حذف شود، واریانس آن سؤال نیز از خرده آزمون حذف می‌شود و ممکن است نقش آن خرده آزمون در نمره کل نیز کاهش یابد.

به طور معمول در آزمون‌های شبیه به آزمون‌های ورود به دانشگاه که هنجار مرجع هستند ممکن است با استفاده از نمره‌های استاندارد شده که بر میانگین و انحراف معیار توزیع نمره‌های آزمون‌دهندگان مبتنی هستند، نمره هر فرد در آزمون را با گروه هنجار مورد مقایسه قرار دهند و یا نمره کل ترکیبی L را تهیه کنند. هنگامی که نمره‌ها تبدیل خطی می‌شوند مثل نمره Z، اگر بخواهیم وزن‌دهی مؤثری اعمال کنیم چون واریانس تمام خرده آزمون‌ها یک خواهد بود نقش آنها حذف می‌شود و *بهتر است برای اعمال وزن از میانگین همبستگی خرده آزمون‌ها استفاده شود.*

نکته آخر که باید در مورد توزیع آماری نمره‌ها در نظریهٔ کلاسیک مدنظر قرار گیرد این است که هیچ مفروضه‌ای در بارهٔ توزیع نمره‌ها آنچنان که در نظریهٔ سؤال-پاسخ(IRT)[۱] وجود دارد ساخته نشده است. در نظریهٔ کلاسیک نمره X یک آزمون-دهنده به محتوای آزمون وابسته است. به عبارت دیگر نمره آزمون‌دهنده اندازه‌ای مطلق از خصیصه مورد اندازه‌گیری (مثل توانایی ریاضی یا فیزیک) او نیست. نمره هر فرد مجموع نمره‌های به دست آمده بر اساس سؤال‌های صحیح پاسخ داده شده آزمون

۱. Item Response Theory

است و ضرایب سؤال‌های و آزمون شدیداً تحت تأثیر گروه نمونه آزمون‌دهنده‌ها قرار می‌گیرد و پارامتر توانایی واقعی جامعه آزمون دهندگان همچنان نامشخص است. از این رو باید گفت که آزمون‌سازی بر اساس نظریهٔ کلاسیک رویکرد توزیع آزاد می‌باشد. به عبارت دیگر در این نظریه هیچگاه مقادیر نمره مشاهده شده X با مقادیر نظری T که در این تئوری به عنوان نمره واقعی به آن اشاره شده همگرا نخواهد شد و مقدار خطای اندازه‌گیری E تحت تأثیر اجراهای مجدد یک آزمون تبدیل به چند نوع خطای سیستماتیک می‌شود و ماهیت تصادفی بودن خود را از دست می‌دهد و پیش-بینی یا برآورد مقادیر میانگین و انحراف معیار خطاهای سیستماتیک غیرممکن و یا برای محاسبه به روش‌های پیچیده‌تر آماری مثل تئوری تعمیم پذیری دارند (د کلرک[۱]، ۲۰۰۸؛ برنان،۲۰۰٤).

نتایج حاصل از روش‌های مختلف ساخت نمره‌های ترکیبی وزنی هم تقریباً مشابه یکدیگر هستند. یکسانی توزیع آماری روش‌های وزن‌دهی نشان می‌دهند که استفاده از روش‌های مختلف وزن‌دهی نمی‌تواند به طور قابل ملاحظه نتایج پایایی را تحت تأثیر قرار دهد (شان ون چانگ[۲]، ۲۰۰۹) و *پایایی را به مقدار رضایت بخشی افزایش دهد. در این روش پایایی از همان ابتدا تحت کنترل وزن مؤثر یا تصادفی هر سؤال و طول بهینه هر خرده آزمون در مجموعه آزمون است. وزن‌دهی می‌تواند بیشترین تأثیر را بر روایی آزمون داشته باشد (برنان، ۲۰۰٤) و قانون ساده این است که مؤلفه، سؤال یا خرده آزمونی که خصیصه موردنظر را بهتر اندازه‌گیری می‌کند وزن بیشتری دریافت کند!*

پیشنهادها

اگر واریانس نمره پاسخ یک سؤال σ_x به عامل‌های اول تجزیه شود (مثل تحلیل واریانس) داریم:

$$\sigma_x = \sigma_i + \sigma_t + \sigma_{it} + \sigma_e \quad (٤)$$

۱. De Klerk

۲. Shun-Wen Chang

σ_i = واریانس سؤال

σ_t = واریانس گروه آزمون دهنده

σ_{it} = تعامل واریانس سؤال i و واریانس گروه t

σ_e = خطای اندازه‌گیری

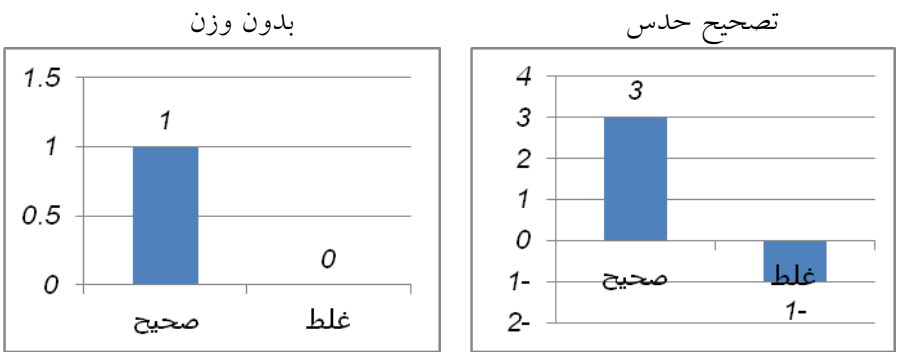
حال سؤال اساسی این است که روش‌های نمره دهی به روش سازمان سنجش (ضریب تصحیح حدس) و یا صفر و یک چه تأثیری بر برآورد نمره واقعی T دارد؟

در روش ۰ و ۱ آنچه که به عنوان پاسخ سؤال در نظر گرفته می‌شود قطعیت صفر یا یک است و هیچ انتخاب میانه یا بیشینه احتمالی وجود ندارد. بنابراین در این نوع نمره دهی هیچ راه حلی برای بیشینه ساختن برآورد توانایی افراد بر اساس پاسخ درست به سؤال وجود ندارد. اما در روش‌های سهمی که به هر بخش از پاسخ صحیح فرد به یک سؤال، سهمی از نمره تعلق می‌گیرد و به عنوان مثال به پاسخ کاملاً غلط صفر، بخشی از پاسخ نمره یک، به بخشی دیگر نمره دو و به پاسخ کاملاً صحیح نمره سه تعلق می‌گیرد و به عبارتی نمره‌دهی به صورت سیاه- سفید نیست، توانایی فرد در مورد یک سؤال بر اساس یک مقدار احتمال، بیشینه می‌شود. به نوعی نمره گذاری گزینه‌های غلط بر اساس محبوبیت آنها نیز از همین الگو پیروی می‌کند. همین امر به کاهش خطای برآورد و افزایش پایایی کمک می‌کند. آنچه که در فرایند نمره گذاری سؤال بر اساس ضریب تصحیح حدس اتفاق می‌افتد نیز شبیه نمره‌دهی به صورت صفر و یک است. در این روش نیز قطعیت پاسخ مطرح است؛ منفی یک (۱-) یا سه (۳) و در نهایت فرد از عدم پاسخگویی به سؤال نمره صفر دریافت می‌کند و هیچ راه میانه‌ای برای بیشینه کردن احتمال پاسخ فرد و برآورد دقیق‌تر نمره واقعی نیست. در حقیقت نمره منفی یک (۱-) نه واریانس σ_i واقعی سؤال و نه واریانس σ_t توانایی حقیقی T جامعه آزمون‌دهنده‌ها، بلکه بیشتر واریانس ترس و عدم اطمینان است که در دامنه واریانس خطا σ_e قرار می‌گیرد. از دیدگاه احتمالات نیز می‌توان ثابت کرد که وقتی به افراد اطلاع داده می‌شود حدس صحیح سه امتیاز و حدس غلط یک امتیاز منفی (۱-) دارد، عملاً آزمون‌دهنده‌ها را به اضافه کردن این مقدار واریانس خطا

تشویق می‌کنیم که این خود باعث کاهش پایایی و افزایش خطای اندازه‌گیری می‌شود. رها کردن سؤال و دریافت نمره صفر نیز هیچ مقداری از توانایی را اندازه نمی‌گیرد.

اولین هدف نمره‌گذاری بر اساس نمره‌دهی سهمی این است که منجر به برآورد دقیق‌تر توانایی آزمون‌دهنده می‌شود تا این که به پاسخ‌های او نمره صفر و یک تعلق بگیرد. بنابر این روش‌های دیگر نمره‌دهی به پاسخ سؤال‌ها از جمله طرح‌های وزن‌دهی متفاوت، یعنی سهمی را می‌توان برای سنجش دانش جزئی آزمون‌دهنده‌ها برای برآورد بیشینه درست‌نمایی[۱] توانایی فرد به کار برد (یو[۲]، ۱۹۹۱).

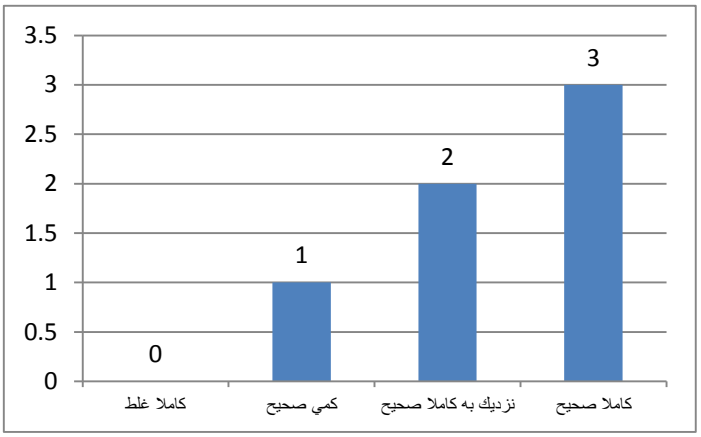
نمودارهای سه روش نمره دهی به سؤال بدون وزن یا صحیح-غلط، وزن‌دهی بر اساس ضریب تصحیح حدس و سهمی



سهمی

۱. maximum likelihood estimation

۲. Yu



اما ساخت سؤال‌هایی که بتوان برای هر مقدار پاسخ صحیح به آن نمره‌ای سهمی در نظر گرفت و همچنین تصحیح آنها در نمونه‌های بزرگ با دشواری‌هایی رو به رو است. یک راه‌حل آسان برای برطرف کردن این مشکلات استفاده از وزن دشواری سؤال $1-p$ برای بیشینه کردن توانایی افراد و همزمان برآورد دقیق‌تر توانایی افراد است. به عبارت دیگر به جای وزن‌دهی سهمی به بخش‌های مختلف پاسخ یک سؤال σ_i، وزن‌دهی بر مبنای سطح دشواری $1-p=q$ یا واریانس گروه σ_t صورت می‌گیرد. بر اساس این روش $1-p=q$ نمره به پاسخ صحیح هر سؤال تعلق می‌گیرد. یعنی افرادی که به سؤال‌های سخت‌تر جواب می‌دهند بر اساس نسبت $1-p=q$ نمره بهتری دریافت می‌کنند. این بهترین و نسبت به بهینه بودن نتایج سریع‌ترین روش وزن‌دهی در نظریهٔ کلاسیک است و در این پژوهش نیز توزیع آماری داده‌ها به خوبی این روش را نشان داده است. در این روش هرگاه دو نفر به تعداد برابر اما با ترکیبی مختلف به سؤال‌ها جواب صحیح داده باشند، جایگاه فردی که به سؤال‌های سخت‌تری جواب می‌دهد در توزیع نمره‌های X گروه آزمون‌دهنده‌ها به نفع فرد توانمندتر رو به بالا جا به جا می‌شود. به عبارت دیگر بر خلاف روش‌های دیگر وزن‌دهی، که در توزیع نمره‌ها تغییر زیادی ایجاد نمی‌کنند، در این روش، کسانی که به سؤال با ضریب دشواری ۰/۲۰ پاسخ داده‌اند ۰/۸۰ امتیاز و به کسانی که به سؤال با ضریب دشواری ۰/٤۰ پاسخ داده‌اند ۰/٦۰ امتیاز تعلق می‌گیرد. در این روش وزن‌دهی بر مبنای σ_t صورت می‌گیرد نه بخش‌های مختلف پاسخ سؤال σ_i که به صورت

سهمی نمره‌گذاری می‌شوند. با یک انتقال خطی نیز می‌توان نمره‌های به دست آمده را به هر مقیاس و میانگین عددی تبدیل کرد.

بر اساس روشی که در ادامه می‌آید و محبوبیت زیادی داشته (به نقل از استانلی و وانگ، ۱۹٦۸) برای پاسخگویی غلط افرادی که با حدس به سؤال‌ها پاسخ می‌دهند، می‌توان جریمه‌ای نیز برابر با نسبت $p-$ در نظر گرفت. به عبارت دیگر بر اساس q نمره مثبت و بر اساس $p-$ جریمه دریافت می‌کنند. در این روش میانگین آزمون برابر با صفر می‌شود. در کل نمونه آزمون‌دهنده‌ها، میانگین هر سؤال برابر است با $qp+(-pq)\equiv ۰$ و بنابر این میانگین نمره آزمون برای همه آزمون‌دهنده‌ها نیز برابر با صفر می‌شود (استانلی و وانگ، ۱۹٦۸).

پیشنهاد دوم دربارهٔ روش‌های انتخاب بهترین متغیرها یا خرده آزمون هایی (X_i) است که نشان دهد این متغیرها می‌توانند فرد را به بهترین صورت در یک رشته انتخابی (Y) جای گذاری کنند (البته با فرض حذف آزمون کنکور و افزایش روایی نمره‌های پیشینه تحصیلی). اگر سنجه‌های متغیر ملاک (Y) در دسترس باشد با بسط معادله $wX_i+b=Y$ به روش رگرسیون چند متغیره می‌توان مجموعه‌ای از وزن‌ها را بدست آورد که در آن با کمینه کردن خطای پیش‌بینی از روی نمونه‌ای که براساس آن وزن‌ها به دست آمده‌اند-تحت مفروضه‌های نرمال بودن و خطی بودن- پیش‌بینی هایی را فراهم کرد. هنگامی که هیچ ملاک (Y) خارجی وجود ندارد یعنی حالت $wX_i+b=۰$، می‌توان با توجه به مفروضه‌های معینی که در ارتباط‌اند با ماهیت متغیری که فرض می‌شود نمره کل ترکیبی L آن را اندازه می‌گیرد، وزن‌های بهینه‌ای را به متغیرها و یا مؤلفه‌ها (X_i) اختصاص داد که مقدار خطای برآورد را کمینه سازد یعنی: $X_i=\frac{-b}{w}$. متغیرها (X_i) و وزن‌ها (w) را می‌توان با بیشینه ساختن یک ملاک درونی معین، مثل افزایش پایایی نمره کل ترکیبی L انتخاب کرد. هرگاه پایایی نمره کل حاصل از ترکیب وزنی و غیروزنی متغیرها (X_i) افزایش بیشینه داشت آن ترکیب بهترین پیش بینی کننده یک (Y) فرضی یعنی عملکرد فرد خواهد بود و یا می‌توان با یک ماتریس، واریانس-کواریانس متغیرهایی[1] را وارد معادله کرد که بیشینه کواریانس را به دست دهد.

صرف نظر از اینکه کدام یک از روش‌ها برای به دست آوردن وزن‌ها انتخاب شوند، فلسفه وزن‌دهی بر اساس این حقیقت است که وزن بزرگ‌تر به متغیری تعلق

۱. این کار را می‌توان بر اساس تهیه سؤالات مناسب برای هر خرده آزمون و افزایش روایی آنها تحقق بخشید

می‌گیرد که ملاک (Y) انتخاب شده را بهتر اندازه بگیرد یعنی بتواند عملکرد موفقیت آمیز بعدی فرد را پیش بینی کند و وزن کمتر و شاید منفی به متغیری تعلق می‌گیرد که ملاک (Y) انتخاب شده را بدتر اندازه بگیرد (استانلی و وانگ، ۱۹٦۸؛ هی، ۲۰۰۹؛ برنان، ۲۰۰٤).

منابع

- پرند، کوروش و ذوالفقارنسب، سلیمان (۱۳۸۹). گزارش تحلیلی آزمون دستیاری پزشکی۱۳۸۸/۱۲/۱۳. وزارت بهداشت، درمان و آموزش پزشکی. انتشار نیافته.
- آلن، مری جی. و ین، وندی ام. (۱۹۷۹). *مقدمه‌ای بر نظریه‌های اندازه‌گیری (روانسنجی)*. ترجمه علی دلاور (۱۳۷٤). سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها.

- Brennan, Robert L. (2004). Some Perspectives on Inconsistencies among Measurement Models. *Measurement and Assessment (CASMA) College of Education*. University of Iowa City, IA 52242. Tel: 319-335-5439. Fax: 319-384-0505. Web: www. uiowa. edu/˜casma.
- Childs, Ruth A., Susan Elgie, Tahany Gadalla, Ross Traub, Andrew P. Jaciw (2004). IRT-Linked Standard Errors of Weighted Composites. *Practical Assessment, Research & valuation*, 9 (13). Retrieved May 17, 2010 from http: //PAREonline. net/getvn. asp?v=9&n=13
- De Klerk, G. (2008). Classical test theory (CTT). In M. Born, C. D. Foxcroft & R. Butter (Eds.), *Online Readings in Testing and Assessment*, International Test Commission, http: //www. intestcom. org/Publications/ORTA. php.
- He, Qing ping (2009). Estimating the Reliability of Composite Scores. Office of Qualifications and Examinations Regulation Spring Place Coventry Business Park Herald Avenue Coventry CV5 6UB. www. ofqual. gov. uk
- Hendrickson, A., Patterson, B. & Ewing, M. (2010). Developing Form Assembly Specifications for Exams with Multiple Choice and Constructed Response Items: Balancing reliability and validity concerns. *Paper presented at the Annual Conference of the National Council for Measurement in Education*, May 1, 2010, Denver.
- Nedelsky, L. (1954). Absolute Grading Standards for Objective Tests.Educational and Psychological Measurement. 1954, v. 14 no. 1.
- Nelson, L. R. (2001). Item Analysis for Tests and Surveys Using Lertap 5. University of Technology Perth, Western Australia. www. lertap. curtin. edu. au/HTMLHelp/LRTP5HHelp. pdf
- Roid, G. H., & Carson, A. D. (2003). *Special Composite Scores for the SB5*. (Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 4). Itasca, IL: Riverside Publishing.

- Shun-Wen Chang (2008). Effects of Gaps-Minimizing Approaches on the Raw-to-Scale Score Conversions When Forms Vary in Difficulty. Bulletin of Educational Psychology, 2008, 39, Special Issue on Test and Measurement, 151-174 National Taiwan Normal University, Taipei, Taiwan, R. O. C.
- Shun-Wen Chang (2009). Choice of Weighting Scheme in Forming the Composite. Department of Educational Psychology and Counseling. National Taiwan Normal University. *Bulletin of Educational Psychology*, 2009, 40 (3), 489-510
- Stanley, J. C., & Wang, M. D. (1968). Differential weighting a survey of methods and empirical studies. New York/10027
- Webb, N. M., Shavelson, R. J., & Haertel E. H. (2006). Reliability Coefficients and Generalizability Theory. Handbook of Statistics, Vol. 26. ISSN: 0169-7161. © 2006 Elsevier B. V. DOI: 10. 1016/S0169-7161 (06) 26004-8.
- Yu, Min-Ning (1991). The Assessment of partial Knowledge. *The Journal of National Chengchi University*, Vol. 63, 1991.

پیوست ۱

وزن‌های اختصاص یافته به گزینه‌ها و به گزینه صحیح $1-p$				سؤال	محبوبیت گزینه‌ها و ضریب دشواری سؤال				
الف	ب	ج	د		بدون پاسخ	الف	ب	ج	د
0/65	0/06	0/26	0/11	1	23%	35%	6%	26%	11%
0/06	0/05	0/04	0/79	2	64%	6%	5%	4%	21%
0/06	0/15	0/16	0/74	3	37%	6%	15%	16%	26%
0/61	0/02	0/06	0/09	4	44%	39%	2%	6%	9%
0/17	0/84	0/11	0/08	5	49%	17%	16%	11%	8%
0/03	0/58	0/02	0/01	6	51%	3%	42%	2%	1%
0/02	0/05	0/03	0/77	7	68%	2%	5%	3%	23%
0/12	0/62	0/01	0/03	8	45%	12%	38%	1%	3%
0/08	0/56	0/18	0/02	9	28%	8%	44%	18%	2%
0/43	0/05	0/03	0/03	10	32%	57%	5%	3%	3%
0/08	0/04	0/06	0/51	11	33%	8%	4%	6%	49%
0/56	0/10	0/04	0/14	12	28%	44%	10%	4%	14%
0/03	0/01	0/67	0/04	13	59%	3%	1%	33%	4%
0/68	0/05	0/27	0/00	14	36%	32%	5%	27%	0%
0/02	0/70	0/04	0/09	15	55%	2%	30%	4%	9%
0/03	0/79	0/22	0/12	16	43%	3%	21%	22%	12%
0/05	0/01	0/56	0/03	17	46%	5%	1%	44%	3%
0/01	0/07	0/01	0/47	18	38%	1%	7%	1%	53%
0/01	0/03	0/04	0/74	19	66%	1%	3%	4%	26%
0/09	0/09	0/82	0/09	20	56%	9%	9%	18%	9%
0/77	0/06	0/06	0/04	21	61%	23%	6%	6%	4%
0/11	0/08	0/08	0/86	22	58%	11%	8%	8%	14%
0/03	0/02	0/91	0/02	23	84%	3%	2%	9%	2%
0/75	0/01	0/01	0/01	24	72%	25%	1%	1%	1%
0/05	0/03	0/70	0/03	25	59%	5%	3%	30%	3%
0/03	0/01	0/20	0/87	26	63%	3%	1%	20%	13%
0/06	0/87	0/01	0/02	27	78%	6%	13%	1%	2%
0/01	0/02	0/14	0/76	28	59%	1%	2%	14%	24%
0/05	0/49	0/02	0/15	29	27%	5%	51%	2%	15%
0/03	0/02	0/80	0/06	30	55%	3%	16%	20%	6%