دورهٔ دوازدهم، شمارهٔ اول، بهار و تابستان ۱۳۹۹، پیاپی ۷۸/۱، صفحه‌های ۲۴۹- ۲۷۶
(مجلهٔ علوم اجتماعی و انسانی سابق)

پایایی معیارهای پذیرش دانشجوی دکتری مهندسی برق قدرت در ایران: رویکردی بر مبنای تئوری تعمیم‌پذیری[1]

سلیمان ذوالفقارنسب* علی دلاور نورعلی فرخی*** احسان جمالی*****

چکیده

فرایند سنجش و پذیرش افراد توانمند و کارآمد برای تحصیل در دورهٔ دکترای تخصصی در ایران بسیار پر چالش بوده است. اهمیت این مسئله به حدی است که ویژگی‌های روان‌سنجی سنجه‌های پذیرش و ملاک‌های مهم آن، همچنین شیوهٔ آزمون‌گیری به صورت تستی یا تشریحی برای تدوین الگوی بهینهٔ پذیرش داوطلبان همواره در بین مجامع دانشگاهی، سیاست‌گذاران، و متقاضیان ورود به تحصیلات تکمیلی محل بحث و جدال است. هدف اولیهٔ این پژوهش مشخص کردن ضرایب شبه پایایی این معیارها بوده است. روش این پژوهش از نوع توصیفی است، و در چارچوب تئوری تعمیم‌پذیری، پایایی سنجه‌ها و معیارهای-داده‌های ثانویه- مربوط به برنامهٔ سنجش و پذیرش دکتری رشتهٔ مهندسی برق گرایش قدرت در ۳۷ دانشکدهٔ پذیرندهٔ دکتری در دانشگاه‌های ملی مختلف در سال ۱۳۹۷ را بررسی کرده است. داده‌ها با نرم‌افزار mGENOVA بر اساس تحلیل چندمتغیری با یک طرح تک رویه‌ای $p^{\bullet} \times i^{\circ}$ تحلیل شده‌اند. نتایج نشان داد که چهار آزمون تخصصی، و دو آزمون عمومی به دلیل سختی سؤالات و بی‌پاسخ بودن، و به‌کارگیری نمره‌گذاری فرمولی، پایایی و دقت مناسبی ندارند. دو معیار ترکیبی مصاحبه که از سنجه‌های متفاوتی تشکیل شده‌اند، از ضرایب تعمیم‌پذیری و اعتمادپذیری بهتری برخوردار بودند. افزودن معدل کارشناسی و کارشناسی ارشد، به عنوان متغیر پیش‌بینی‌کنندهٔ جداگانه، به دلیل محدودیت دامنه در برآورد نمرات جهانی ترکیبی کارایی چندانی ندارد. در یک برنامهٔ پذیرش دکتری که سازهٔ زیربنایی‌ای بسیار گسترده‌ای تعریف، و سنجه‌های چندگانه‌ای استفاده می‌شود، مؤلفه‌های خطای بیشتری نیز وجود خواهد داشت. بنابراین، نمی‌توان تنها یک مقدار پایایی معین برای سنجه‌های آن مشخص کرد. امّا با تعدیل سطوح دشواری آزمون‌ها، استفاده از طرح‌های نمره‌گذاری سهمی سؤالات بدون جریمه حدس شانسی، نمونه‌گیری کاملتر از سازهٔ زیربنایی و افزایش لایه‌های سطوح می‌توان تصمیم‌گیری‌های آموزشی با ریسک بالا را برای طبقه‌بندی افراد، با پیش‌بینی‌های نادرست کمتری انجام داد.

واژه‌های کلیدی: تئوری تعمیم‌پذیری، پایایی، پذیرش دکتری، نمرهٔ برش، خطای مثبت کاذب، خطای منفی کاذب

---

[1]. این مقاله برگرفته از پایان‌نامهٔ دکتری با عنوان: «ارزیابی جامع معیارهای پذیرش دانشجوی دورهٔ دکتری مهندسی: رویکردی بر مبنای تئوری تعمیم‌پذیری»، است.

* دانشجوی دکتری سنجش و اندازه‌گیری، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی salarnik2001@yahoo.com (نویسندهٔ مسئول)

** استاد تمام گروه سنجش و اندازه‌گیری، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی delavarali@yahoo.com

***دانشیار گروه سنجش و اندازه‌گیری، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی farrokhinoorali@yahoo.com

****استادیار پژوهشی سازمان سنجش آموزش کشور ehsanjamali@gmail.com

تاریخ دریافت مقاله: ۹۸/۷/۱۱ تاریخ دریافت مقاله نهایی: ۹۹/۲/۱۴ تاریخ پذیرش: ۹۹/۴/۱۰

مقدمه

فرایند سنجش و پذیرش افراد توانمند و کارآمد برای تحصیل در دورهٔ دکترای تخصصی در ایران بسیار پرچالش بوده است. اهمیت این مسئله به حدی بوده که ویژگی‌های روان‌سنجی سنجه‌های پذیرش و ملاک های مهم آن، همچنین شیوهٔ آزمون‌گیری به صورت تستی یا تشریحی برای تدوین الگوی بهینهٔ پذیرش داوطلبان، همواره در بین مجامع دانشگاهی، سیاست‌گذاران و متقاضیان ورود به تحصیلات تکمیلی محل بحث و جدال است. در باز تعریفی که از ایجاد دورهٔ دکتری در ایران شده است، به طور خلاصه هدف تربیت تخصصی افرادی است که بتوانند با نوآوری نیازهای کشور را رفع کنند، و در رشتهٔ تخصصی خود مرزهای دانش را در جهان گسترش دهند (برای جزئیات بیشتر به فصل اول آئین‌نامهٔ آموزشی دورهٔ دکتری تخصصی، ۱۳۹۷، نگاه کنید). بدین منظور هر ساله تعداد زیادی داوطلب دورهٔ دکتری خود را امّادهٔ رقابت در آزمون‌های تخصصی و عمومی می‌کنند، پس از به‌دست آوردن نمرات حد نصاب وارد مرحلهٔ مصاحبهٔ علمی می‌شوند، در نهایت افرادی که بیشترین نمرات را کسب کنند پذیرش دورهٔ دکتری می‌گیرند. از گذشته و در طول دورانی که برنامهٔ دکتری اجرا شده است این رویه‌ای عمومی است.

این فرایند سنجش و پذیرش تا پیش از سال ۱۳۹۰ به صورت غیرمتمرکز بود، و مستقیماً به‌وسیلهٔ هر دانشگاه و براساس مقرارت و ضوابط درون دانشگاهی انجام می‌شد؛ برنامهٔ آزمون هر دانشگاه دربرگیرندهٔ دو مرحله آزمون عمومی و تخصصی برای غربال‌گری اولیه، سپس مرحلهٔ مصاحبه برای پذیرش یا رد داوطلبان بود. عمدتاً سؤالات آزمون‌ها باز پاسخ بودند، و سطوح بالای شناختی (وب، ۱۹۹۷) را اندازه‌گیری می‌کردند، و نمره‌گذاری سؤالات باز پاسخ آزمون‌ها و سنجه‌های مصاحبه به صورت سهمی بود. امّا الگوی متفاوت برنامهٔ آزمون دانشگاه‌ها برای رشته‌های یکسان، نامشخص بودن شیوهٔ نمره‌گذاری آزمون‌ها و سنجه‌های مصاحبه، آگاهی نداشتن از میزان خطای معیار و پایایی آزمون‌ها، و اعمال سلیقه‌های مختلف در پذیرش دانشجو، و رعایت نکردن عدالت آموزشی (یونسی، ۱۳۹۵) در برخی از دانشگاه‌ها مشکلات فراوانی را پیشروی داوطلبان ورود به مقطع دکتری قرار داده بود.

به دلیل اعتراضات زیادی که بر نحوهٔ پذیرش بخش‌های دانشگاهی می‌شد تصمیم‌گیرندگان در وزارت علوم به این نتیجه رسیدند که برای کاهش خطا و برقراری عدالت، شیوهٔ پذیرش را مشترکاً به دانشگاه‌ها و سازمان سنجش آموزش کشور بسپارند. این شیوهٔ اجرایی متفاوت که با نام «آزمون‌گیری نیمه متمرکز» شهرت دارد از سال ۱۳۹۰ آغاز شده، و تا هم‌اکنون ادامه دارد. در این فرایند سازمان سنجش مأموریت یافت تا تقریباً در ۲۵۰ رشته، آزمون تخصصی و عمومی به صورت آزمون چهارگزینه‌ای با صرف هزینه‌های بسیار اجرا کند. پس از غربالگری اولیه، چند برابر ظرفیت

هر رشته، داوطلبان به کمیتهٔ پذیرش بخش‌های دانشگاهی معرفی می‌شوند، سپس مصاحبهٔ علمی و آموزشی صورت می‌گیرد. نمرهٔ‌های حاصل از مصاحبه با داوطلبان به سازمان سنجش برگشت داده می‌شوند. نمرهٔ کل حاصل از جمع وزنی نمرات آزمون‌های تخصصی و عمومی چهارگزینه ای –با تصحیح برای شانس – با نمرات سهمی سنجه‌های مصاحبهٔ علمی و آموزشی مشخص می‌کند کدام داوطلبان پذیرش نهایی را برای رشتهٔ مورد نظر دریافت می‌کنند. این فرایند اجرایی سنجش دوگانه با هدف ایجاد یک موقعیت استاندارد برای افزایش عدالت، کاهش خطا و ایجاد روایی و پایایی در حال حاضر اجرا می‌شود. امّا همچنان دو مسئلهٔ جدی سر راه این فرایند اجرایی متفاوت وجود دارد؛ نخست چالش‌های مربوط به تعیین وزن‌های اسمی هر یک از بخش آزمون‌های تخصصی و عمومی با مسئولیت اجرا توسط سازمان سنجش و مصاحبهٔ علمی آموزشی با مسئولیت اجرا توسط دانشگاه در نمرهٔ کل ترکیبی است، دوم این‌که پایایی سنجه‌هایی که در هر بخش به کار می‌روند چه قدر باید باشد؟

در سال‌های گذشته تصمیم‌گیرندگان تغییرات زیادی در سهم و وزن آزمون تخصصی و عمومی سازمان سنجش و مصاحبهٔ علمی آموزشی دانشگاه‌ها اعمال کرده‌اند؛ بدون این‌که از ویژگی‌های روان‌سنجی این فرایند سنجشی دوگانه آگاهی داشته باشند. به عنوان مثال در سال ۱۳۹۰ چند برابر ظرفیت به کمیتهٔ پذیرش دانشگاه‌ها اعلام شد، امّا پذیرش نهایی بر اساس ۱۰۰ درصد نمرهٔ مصاحبه بوده است: به عبارتی آزمون سازمان سنجش تنها به صورت غربال اولیه، و برای معرفی داوطلبان به کمیته‌های پذیرش دانشگاهی در نظر گرفته شده بود. سپس، در سال ۱۳۹۱ پذیرش داوطلبان بر اساس یک سوم سهم آزمون تخصصی و عمومی سازمان سنجش، و دو سوم سهم مصاحبه کمیته دانشگاه‌ها بوده است. در حالی که تغییرات این‌چنینی در سال‌های بعد نیز روی داده است.

جدول۱- تغییرات آزمون دکتری طی سال‌های ۱۳۹۰ تا ۱۳۹۸

سال	تعداد رشته‌ها	تعداد عناوین دروس	وزن اسمی: نحوهٔ پذیرش نهایی داوطلبان
۱۳۹۰	۱۰۷	۳۲۱	۱۰۰درصد نمرهٔ مصاحبه
۱۳۹۱	۲۸۵	۸۵۵	یک سوم نمرهٔ آزمون کتبی و دو سوم نمرهٔ مصاحبه
۱۳۹۲	۲۸٦	۸۵۸	۵۰ درصد نمرهٔ آزمون کتبی و ۵۰ درصد نمرهٔ مصاحبه
۱۳۹۳	۲٦۷	۸۰۱	۳۰ درصد آزمون کتبی و ۷۰ درصد نمرهٔ مصاحبه
۱۳۹٤	۲۵۱	۷۵۳	۳۰ درصد آزمون کتبی و ۷۰ درصد نمرهٔ مصاحبه
۱۳۹۵	۷۷	۲۳۱	در اختیار دانشگاه
۱۳۹٦	۲٤٤	۷۳۲	۵۰ درصد نمرهٔ آزمون کتبی و ۵۰ درصد نمرهٔ مصاحبه
۱۳۹۷	۲٤٦	۷۳۸	۵۰ درصد نمرهٔ آزمون کتبی و ۵۰ درصد نمرهٔ مصاحبه
۱۳۹۸	۲٤۵	۷۳۵	۵۰ درصد نمرهٔ آزمون کتبی و ۵۰ درصد نمرهٔ مصاحبه

این تغییرات ضرورت استانداردسازی فرایند مصاحبه را نیز الزامّاور کرد. به همین منظور یک شیوه نامه برای نمره‌گذاری سنجه‌های مصاحبه تهیه شد تا هر بخش آن، نمره و وزن مشخصی داشته باشد (برای جزئیات بیشتر، به شیوه‌نامهٔ اجرایی آزمون‌های ورودی دورهٔ دکتری، ۱۳۹۵، نگاه کنید). جدول ۱ تغییرات سال‌های مختلف را نشان می‌دهد.

از عوامل بحث‌برانگیز دیگر تغییراتی است که در چگونگی تأثیردادن نمرهٔ معدل کارشناسی و کارشناسی ارشد در ساخت نمرهٔ کل یک داوطلب بوده است. تا سال ۱۳۸٦، ۸۰درصد نمرهٔ آزمون عمومی و اختصاصی و ۲۰درصد معدل تراز شده محاسبه می‌شد [با وزن مقطع کارشناسی (٦۰درصد) و کارشناسی ارشد (٤۰درصد)] و اثر معدل هر دو مقطع در نمرهٔ آزمون عمومی و اختصاصی از سوی سازمان سنجش اعمال می‌شده است. امّا همان طور که آشکار شده، ارزش یک معدل مثلاً ۱۵ بدست آمده از دانشگاه‌های گوناگون روزانه، با دیگر انواع دانشگاه‌ها (۸ نوع دانشگاه مختلف) تفاوت اساسی دارد. به همین دلیل در سال‌های اخیر در جلسهٔ مصاحبه بر حسب نوع دانشگاه فارغ‌التحصیلی داوطلبان، یک رتبه به معدل آن‌ها داده می‌شود.

بنابراین، اثر معدل در بخش مربوط به نمرات مصاحبه خلاصه می‌شود، و نه به عنوان یک پیش‌بینی‌کنندهٔ مستقل. این روش اخیر اعمال ضریب معدل هم مزایا و هم معایبی دارد؛ شاید مزایای آن این باشد که یک معدل معین حاصل از دانشگاه‌های تراز اول برآمده از تلاش و کوشش بیشتری باشد، تا دانشگاه‌های دیگر (احیاناً اگر این ضریب از سوی کمیتهٔ پذیرش اعمال شود!)، و لازم است به آن ضریب بیشتری داده شود. عیب آن این است که چون به عنوان بخشی از سوابق آموزشی، پژوهشی و فناوری است و با آن ترکیب می‌شود، نقش سهم معدل کارشناسی و کارشناسی‌ارشد به عنوان یک متغیر مستقل در پذیرش نهایی همچنان نامعلوم و پنهان مانده است؛ یعنی به راحتی نمی‌توان برآورد کرد که اگر بتوان نمرات معدل کارشناسی و کارشناسی‌ارشد را به عنوان یکی دیگر از متغیرهای مستقل در معادلهٔ پیش‌بینی پذیرش قرار داد، نرخ خطای پیش‌بینی کاهش پیدا کند؟در جدول (۲) عمده‌ترین معیارهای انتخاب یک داوطلب دورهٔ دکتری و وزن هر کدام آمده است.

جدول۲ معیارها و سنجه‌های پذیرش داوطلبان دکتری سال ۱۳۹۷ و سهم درصدی آن‌ها در نمرهٔ کل ترکیبی

	آزمون متمرکز توسط سازمان سنجش (۵۰درصد)			معدل		مرحلهٔ مصاحبه و بررسی سوابق داوطلبان توسط کمیتهٔ پذیرش دانشگاه (۵۰درصد)		نمرهٔ کل (۱۰۰درصد)
مواد آزمون	آزمون دروس تخصصی (۳تا ۸ درس)	آزمون زبان (۱)	آزمون استعداد (۱)	کارشناسی	کارشناسی ارشد	مصاحبهٔ علمی و سنجش علمی (۳ سنجه)	سوابق آموزشی، پژوهشی و فناوری (٦ سنجه)	
وزن اسمی یا ضرایب	(۰/۳)	(۰/۱)	(۰/۱)	بخشی از سوابق آموزشی و پژوهشی و فناوری است. با توجه به هشت نوع دانشگاه فارغ التحصیلی وزن آن نامشخص است.		(۰/۳)	(۰/۲)	(۱/۰۰)

فرض دانشگاه‌ها این است که افرادی باید وارد دوره‌های دکتری شوند، که سطوح شایستگی خود را بر اساس سنجه‌های پذیرش ترکیبی که در یک برنامهٔ پذیرش دکتری هر ساله اجرا می‌شود نشان بدهند. رویکردی که در این شیوهٔ پذیرش عمومیت دارد این است که، نمرهٔ چند خرده آزمون چهارگزینه‌ای تخصصی و عمومی با حذف اثر حدس شانسی، بر اساس تئوری کلاسیک آزمون‌سازی نمره‌گذاری شوند، و سنجه‌های مصاحبهٔ علمی و آموزشی به صورت سهمی نمره‌گذاری می‌شوند. سپس وزن‌های اسمی به این پیش‌بینی‌کننده‌ها اختصاص می‌یابد، و سرجمع آن‌ها یک نمرهٔ ترکیبی (نگاه کنید به برنامه ،۲۰۰٤) برای هر فرد ساخته می‌شود تا براساس یک رتبه‌بندی در مورد پذیرش نهایی آن‌ها تصمیم‌گیری شود. در این فرایند تهیه آزمون‌ها و پیش‌بینی‌کننده‌ای که هم دقیق و با ثبات باشند (پایایی)، و هم بتوانند در ترکیب با یکدیگر در پیش‌بینی عملکرد موفقیت‌آمیز افراد داوطلب برای خلق نوآوری و گسترش مرزهای دانش در آینده نقش داشته باشند، از اهمیت بالایی برخوردار است (روایی پیش‌بین).

در چنین برنامه‌های آزمون‌گیری دکتری که معمولاً سازهٔ زیربنایی تحت عناوین خلق نوآوری، توان گسترش مرزهای دانش، شایستگی یا صلاحیت (گیل و راگلند[1]، ۲۰۱۸) بسیار گسترده تعریف می‌شود، مستلزم تهیه و توسعهٔ ابزارهای مختلف اندازه‌گیری است که جنبه‌های مختلفی از ایت سازهٔ زیربنایی را اندازه بگیرند (اسمیت، وان در آرک و کانیین[2]، ۲۰۱۸). همچنین، معادلهٔ پیش‌بینی که برای رتبه‌بندی افراد به کار می‌رود، متغیرهای آن با یکدیگر همبستگی بالایی نداشته باشند، امّا با

[1] . Gyll & Ragland
[2] . Smits, van der Ark, & Conijn

متغیرهایی که به عنوان ملاک در نظر گرفته می‌شوند همبستگی بالایی داشته باشند تا روایی پیش‌بین تحقق بیابد. در عین حال این سنجه‌ها و پیش‌بینی‌کننده‌ها باید به اندازهٔ کافی واریانس واقعی تولید کنند تا بدون این‌که همهٔ آن‌ها را در دامنهٔ بالا یا دامنه پایین توزیع نمرات جمع کنند، بتواند بین توانایی داوطلبان تمایز قایل شوند (اتیه[1]، ۱۹۹۹).

در نهایت، هنگامی که یک اندازه‌گیری صورت می‌پذیرد، معمولاً یک کمیت بدست می‌آید. حال این سؤال مطرح است که تاچه حد این اندازه‌گیری معتبر است؟ این امر به مقدار خطای اندازه‌گیری بستگی دارد. هدف اولیهٔ این نوشته تعیین میزان پایایی این سنجه‌ها بر اساس تئوری تعمیم‌پذیری است، چون نمرات آزمون‌های ورودی به‌علاوهٔ معیارهای مصاحبه که برای پذیرش افراد استفاده می‌شود در تعیین سرنوشت افراد یک جامعه نقش اساسی دارند، اطمینان از پارامترهای هر یک از این معیارها، پیش‌بینی‌کننده‌ها و نمرات، از لحاظ تعمیم‌پذیری و اعتمادپذیری باید بخش جدایی‌ناپذیر و حیاتی یک برنامهٔ سنجش و پذیرش، به‌ویژه در دورهٔ دکتری است. در ترکیب این دو نوع فرایند اجرایی متفاوت (آزمون و مصاحبه) اگر میزان خطای پذیرش یا رد افراد در طبقات دو ارزشی ناچیز باشد، به عنوان مثال خطای مثبت کاذب و خطای منفی کاذب کمتر از ۵درصد باشد، می‌توان تنها از لحاظ آماری ادعا کرد، اتخاذ چنین خط مشی نیمه متمرکز در پذیرش دانشجویان تحصیلات تکمیلی مفید بوده است. چرا که رتبه‌ای که در سازمان سنجش تهیه می‌شود، بر اساس تئوری کلاسیک آزمون‌سازی است که از آن مقدار حدس شانسی کم شده است. اگر با یک مدل کارآمدتر مثل تئوری تعمیم‌پذیری «امید نمرات مشاهده شده» برآورد شود، در تعداد خطای مثبت و منفی احتمالاً تغییراتی خواهد بود. بررسی و مقایسه این تغییرات و شناسایی سنجه‌های پرخطا در این رتبه‌بندی‌ها می‌تواند میزان رسیدن به اهداف اولیهٔ سیاست‌گذاری ها را مشخص کند.

در اندازه‌گیری مبتنی بر درجه‌بندی که به وسیلهٔ مصاحبه‌گران در کمیتهٔ پذیرش دانشگاه‌ها صورت می‌گیرد، هم خطای سیستماتیک و هم خطای تصادفی وجود دارد. عواملی که تغییرات سیستماتیک ایجاد می‌کند، می‌تواند تفاوت‌هایی باشد که در سهل‌گیری، یا سخت‌گیری اعضای کمیتهٔ پذیرش در مصاحبه یا تفاوت تفسیر آن‌ها از یک ویژگی شخصیتی، یا تعامل این خطاها، وجود دارد (لین[2]، ۲۰۱٤). در مقابل، تغییرات تصادفی می‌تواند برآمده از تغییرات پیش‌بینی نشده در فرایند نمره‌گذاری باشد. امّا بسیاری از منابع خطا چه سیستماتیک و چه غیرسیستماتیک در اندازه‌گیری در مقیاس بزرگ (ملی) قابل بررسی نیستند، و در اصطلاح‌شناسی تئوری تعمیم‌پذیری تحت عنوان رویه‌ها یا سطوح پنهان نام‌گذاری شده‌اند. پراش ناشی از این سطوح پنهان در واریانس باقیمانده جمع می‌شوند (برنان[3]،

---

[1]. Attiyeh
[2]. Li
[3]. Brennan

۲۰۰۹؛ شیولسون و وب[1]، ۲۰۰٦). با توجه به این دو نوع خطا در فرایند سنجش و پذیرش که متشکل از انواع مختلفی از سنجه‌ها است، چارچوب‌های مفهومی و الگوهای اندازه‌گیری نیرومندتری فراتر از تئوری کلاسیک آزمون‌سازی لازم است، تا بتوان این خطاها را عملیاتی کرد و اندازه‌گیری را بهبود بخشید (برنان[2]، ۲۰۰٤). یک دید روشن از مقدار این خطاهای اندازه‌گیری می‌تواند دو فایده داشته باشد: نخست مشخص کردن این‌که در کجا لازم است فرایندهای اندازه‌گیری بهبود پیدا کند، و دوم کاهش تمایل به تفسیر و عمل بر اساس نمراتی که ممکن است به دلیل خطای اندازه‌گیری در حقیقت بی‌معنی باشند (کان[3]، ۲۰۱۰).

به هر حال، پذیرش دانشجویان دورهٔ دکتری براساس معیارهای صحیح و دقیق تأثیر بسزایی در استانداردسازی شرایط، ایجاد فضای رقابتی یکسان، برقراری عدالت آموزشی خواهد داشت. از این‌رو بررسی گسترده و جامع فرایند پذیرش دانشجو دکتری به‌ویژه مؤلفه‌های مختلف آن بسیار اهمیت دارد. با توجه به این‌که نمرهٔ هر فرد در دو بخش، یعنی آزمون متمرکز سازمان سنجش و مصاحبه دانشگاه‌ها تهیه می‌شود، سپس با تجمیع این دو نمره، یک نمرهٔ کل ترکیبی تشکیل می‌شود؛ سؤالات زیر مطرح است:

۱-ضرایب شبه پایایی نمرات حاصل از معیارهای تهیه شده از سوی سازمان سنجش (مثل آزمون دروس تخصصی، آزمون عمومی استعداد و زبان) و معیارهای تهیه شده از سوی دانشگاه‌ها (مثل نمرات مصاحبهٔ مرکب از مصاحبهٔ علمی، سوابق آموزشی و پژوهشی و نمرات معدل و...) و ضرایب شبه پایایی نمرات کل ترکیبی در رشته مهندسی برق- قدرت چقدر است؟

۲-آیا وزن‌های اسمی یا همان ضرایب پیشین می‌توانند به افزایش ضرایب شبه پایایی کمک کنند؟

۳-آیا افزایش لایه‌های خرده آزمون‌ها و معیارهای مصاحبه می‌تواند به افزایش ضرایب شبه پایایی کمک کند؟

٤-آیا اضافه کردن نمرات معدل کارشناسی و ارشد به جمع آزمون‌های تخصصی و عمومی و مصاحبه به افزایش ضرایب شبه پایایی نمرات کمک می‌کند؟

۵-بر اساس پذیرش در طبقات دو ارزشی میزان برآورد صحیح پیش‌بینی‌ها در رشتهٔ مهندسی برق- قدرت چگونه است؟ یا میزان مطلوب برآورد صحیح چه‌قدر است؟

---

[1] . Shavelson, Webb
[2] . Brennan
[3] . Kane

روش پژوهش

این پژوهش بر اساس «هدف» از نوع کاربردی و بر اساس «روش» از نوع توصیفی است و در چارچوب تئوری تعمیم‌پذیری، پایایی سنجه‌ها و معیارهای-داده های ثانویه-مربوط به برنامهٔ سنجش و پذیرش دکتری رشتهٔ مهندسی برق گرایش قدرت در۳۷ دپارتمان پذیرندهٔ دکتری در دانشگاه‌های ملی مختلف در سال ۱۳۹۷ گزارش شده است.

شرکت‌کنندگان پژوهش

در این برنامهٔ آزمون ۳٤۱۹۷ داوطلب ثبت نام کردند، که تنها ٤٦۰ داوطلب برای مصاحبه به این ۳۷ دانشگاه معرفی شدند. از بین آن‌ها ۱۷۰ نفر پذیرش نهایی را دریافت کرده‌اند.

ابزارهای پژوهش

آزمون‌ها: برنامهٔ سنجش رشتهٔ مهندسی برق گرایش قدرت در برگیرندهٔ ٤ خرده آزمون تخصصی با وزن اسمی ۰/۳۰ است، که جمعاً ٤٥ سؤال دارد [برخی از رشته‌های مهندسی تا ۸ خرده آزمون تخصصی با٤٥ سؤال]. آزمون‌های عمومی آن‌ها متشکل از ۲ آزمون است؛ یک خرده آزمون ۳۰ سؤالی زبان انگلیسی و یک خرده آزمون ۳۰ سؤالی استعداد تحصیلی. جمعاً وزن اسمی برابر با ۰/۲۰ دارند. نمره‌گذاری سؤالات همهٔ این آزمون‌های اختصاصی و عمومی- با حذف اثر حدس شانسی- بر اساس نمرهٔ فرمولی است. مقیاس نمرات هر خرده آزمون پس از حذف اثر شانس و تبدیل مقیاسی می‌تواند از ۳۰۰۰- تا ۱۰۰۰۰ باشد.

مصاحبه: مصاحبه از داوطلبان بر اساس یک فهرست که ۱۱ سنجه دارد در دانشکده‌ها صورت می‌گیرد. این ۱۱ سنجه تبدیل به دو نمره می‌شوند؛ نمرهٔ بخش نخست که از مصاحبهٔ علمی و سنجش علمی تشکیل شده است (مثل آزمون شفاهی یا کتبی داخلی [احتمالاً]، و مصاحبهٔ تخصصی و نظر استادان کمیتهٔ پذیرش که عمدتاً به طور ذهنی سنجه‌های غیر شناختی و شخصیت فرد را نمره‌گذاری می‌کنند)، وزنی برابر با ۰/۳۰ دارد. مقیاس نمرهٔ تبدیل شدهٔ آن می‌تواند از ۰ تا ۳۰۰۰ باشد. نمرهٔ بخش دوم که از معیارهای مربوط به سوابق آموزشی و پژوهشی و فناوری (مثل مقالات، ثبت اختراعات، شرکت در جشنواره‌های علمی معتبر، تألیف یا ترجمهٔ کتاب، وزن معدل با توجه به کیفیت دانشگاه محل تحصیل، امتیاز پایان‌نامه کارشناسی ارشد، پژوهش‌های انجام شده و نظایر آن) تشکیل شده است. وزنی برابر با ۰/۲۰ دارد. مقیاس نمرهٔ تبدیل شده این بخش می‌تواند از ۰ تا ۲۰۰۰ باشد.

سنجهٔ معدل نیز از میانگین وزنی نمرهٔ دروس کارشناسی و کارشناسی ارشد تشکیل شده و در جلسهٔ مصاحبه در بخش دوم یعنی سوابق آموزشی و پژوهشی و فناوری سهم دهی می شود. امّا

هنگامی که به عنوان یک متغیر مستقل در این پژوهش تجزیه و تحلیل می شود، مقیاس نمرهٔ معدل تبدیل شدهٔ کارشناسی می‌تواند از ۱۲۰۰ تا ۲۰۰۰، و مقیاس نمرهٔ معدل تبدیل شدهٔ کارشناسی ارشد می‌تواند از ۱٤۰۰ تا ۲۰۰۰ باشد.

تحلیل داده‌ها

در این پژوهش بر اساس تئوری تعمیم‌پذیری چندمتغیری، و با یک طرح تک‌رویه‌ای، $p^{\bullet} \times i^{\circ}$، داده‌های ثانویه رشتهٔ مهندسی برق قدرت با نرم‌افزار mGENOVA تحلیل شده است (برای توضیح بیشتر در رابطه با مفاهیم این تئوری به پیوست ۲ نگاه کنید).

یافته‌ها

در این بخش نتایج حاصل از تحلیل داده‌ها آمده است. برای به دست آوردن آماره‌ها و ضرایب شبه پایایی مناسب‌تر از متغیرها، ارزش‌های معتبری که بین ۰ تا ۱۰۰۰۰ بوده‌اند تحلیل شده‌اند. بنابراین، نمره‌های منفی ۱۸٦ نفر که در آزمون‌های تخصصی و عمومی پایین‌تر از صفر بودند (به دلیل تصحیح حدس شانسی) از محاسبات حذف شدند. نمرات منفی باعث می‌شوند پارامترها همگرا نشوند: به عنوان مثال ایجاد همبستگی‌های کاهش نیافته (یعنی ضریب همبستگی بیشتر از ۱) و تشدید کواریانس‌های منفی بین متغیرها می‌کنند (نگاه کنید به برنان، ۲۰۰۹).

جدول ۳- اطلاعات متغیرها، وزن‌های اسمی و آماره‌های توصیفی هر متغیر

اطلاعات مربوط به متغیرها	اندازه نمونه‌ای[3] سطوح و لایه‌ها	میانگین هر لایهi	میانگین بزرگ[2](کل)	وزن‌های پیشین یا وزن اسمی[1]	مقیاس نمرات Min	مقیاس نمرات Max	تعداد N
دروس تخصصی	$n_1 = 4$	۲۱۷٦/٤۲ ۲۲٤۹/۳۱ ۱۷۹٦/۸۳ ۳۸٦۹/۹٤	۲۵۲۳/۱۲	۰/۳۰	-۳۰۰۰	۱۰۰۰۰	٤٦۰
آزمون استعداد و زبان	$n_2 = 2$	۱۷۸٤/٦۹ ۱۱۳۲/۱۹	۱٤۵۸/٤٤	۰/۲۰	-۳۰۰۰	۱۰۰۰۰	٤٦۰
مصاحبه	$n_3 = 2$	۱۷۷۱/۳۱ ٦۳٦/۱٦	۱۲۰۳/۷٤	۰/۵۰	۰ ۰	۳۰۰۰ ۲۰۰۰	٤٦۰
*معدل	$n_4 = 2$	۱۵۱۵/٤۳ ۱٦۷۳/۳۲	۱۵۹٤/۳۸		۱۲۰۰ ۱٤۰۰	۲۰۰۰ ۲۰۰۰	٤٦۰

*در تحلیل‌های اولیه برای حذف اعشار، مقیاس معدل در ۱۰۰ ضرب شده است

1 . priori or nominal weights
2 . grand mean
3 . sample size

بدین ترتیب تعداد کل رکوردهای معتبر برای محاسبه ضرایب تعمیم‌پذیری ۲۷٤ داوطلب است. امّا برای طبقه‌بندی افراد به دو طبقه رد و قبول و محاسبه فراوانی‌های مثبت کاذب و منفی کاذب از همهٔ داده‌های ٤٦۰ نفر استفاده شده است. جدول ۳ خلاصه‌ای از آماره‌های توصیفی نمرات مشاهده شده روی خرده آزمون‌های مربوط به ٤٦۰ نفر از داوطلبان معرفی شده به ۳۷ بخش دانشگاهی در رشته برق گرایش قدرت را در سال ۱۳۹۷ نشان می‌دهد.

در جدول ٤ نتایج مطالعات تعمیم‌پذیری G به همراه تصمیم‌گیری پیش‌گزیده D_0 آمده است. در تصمیم‌گیری پیش گزیده D_0 وزن مؤثر هر متغیر n_i در مجموعه آزمون که انعکاس‌دهنده سهم آماری نسبی لایه‌های متغیر(یعنی خرده آزمون‌ها) در تعیین واحد جهانی یا جهان تعمیم[1] است، بیشترین نقش در برآورد واریانس نمرهٔ جهانی دارد (برنان، ۲۰۰۹؛ ص.۳۰۷) [برای توضیحات بیشتر به پیوست ۲ نگاه کنید]. در جدول ۵ منابع واریانس برای هر متغیر و ضرایب شبه پایایی آن‌ها بر اساس وزن مؤثر هر متغیر در ایجاد نمرهٔ جهانی ترکیبی آمده است.

جدول ٤ نتایج مطالعات تعمیم‌پذیری G و تصمیم‌گیری پیش گزیده D_0 با طرح چند متغیره $i^{\circ} \times p^{\bullet}$ و با همان اندازه نمونه‌ای[2] (n_1=٤, n_2=۲, n_3=۲) و با وزن‌ها مبتنی بر اندازه نمونه‌ای n اولیه (یا وزن مؤثر)

G study				D study			
p	5771539/91	0/05	0/35	P	544360/78	0/00	0/80
	56063/73	1648299/22	0/18		56063/73	−47208/41	0/00
	307726/61	121455/72	1023614/21		307726/61	121455/72	265518/60
i	231675714/63			I	208103/66		
		58327301/25				103256/54	
			176532227/91				321240/23
pi	3594096/77			PI	898524/19		
		1742716/05				871358/02	
			492577/00				246288/50

نکته: مطالعات تعمیم‌پذیری و تصمیم‌گیری با $N = ۲۷٤$ و واریانس برآورد شده(قطری)، کواریانس‌های مشاهده شده(پایین قطری) و همبستگی‌های مشاهده شده(بالای قطری)

در بررسی D_0 مقادیر منفی برآورد واریانس آزمون‌های عمومی(٤۷۲۰۸/٤۱-) نشان می‌دهد تغییرات ناشی از این آزمون پیش از آنکه برآیند واریانس واقعی نمرات داوطلبان باشد، برآیند واریانس خطاهای دو خرده آزمون به دلیل سختی، رها کردن سؤالات، یا نمرهٔ منفی برای تصحیح حدس است. در مقایسه با تحلیل واریانس این اتفاق هنگامی رخ می‌دهد که واریانس درون گروهی SSw بزرگ‌تر از واریانس بین گروهی SSb با شد. همچنین، همبستگی کانونی آزمون‌های عمومی با آزمون‌های اختصاصی ۰/۰۰ و با نمرات مصاحبه ۰/۸۰ است. مقادیر واریانس‌های هر بخش و ضرایب شبه پایایی هر متغیر و سهم درصدی هر یک در جدول ۵ آمده است.

1 . specifying universe of generalization
2 . same sample sizes

جدول ۵ نتایج مطالعات تصمیم‌گیری پیش گزیده D_0 بر اساس طرح چند متغیره $p^{\bullet} \times i^{\circ}$ و با همان اندازه نمونه‌ای ($n_1 = 4, n_2 = 2, n_3 = 2$) و با وزن‌ها مبتنی بر اندازه نمونه‌ای n اولیه (یا وزن مؤثر)

منابع واریانس		آزمون‌های اختصاصی $n_1 = 4$	آزمون‌های عمومی $n_2 = 2$	مصاحبه $n_3 = 2$	ترکیبی
واریانس نمرهٔ جهانی	$\hat{\sigma}^2_{(\tau)}$	٥٤٤٣٦٠/٧٨	−٤٧٢٠٨/٤١	٢٦٥٥١٨/٦٠	٢٥٥٨٦٤/١٣
واریانس خطای نسبی	$\hat{\sigma}^2_{(\delta)}$	٨٩٨٥٢٤/١٩	٨٧١٣٥٨/٠٢	٢٤٦٢٨٨/٥٠	٢٩٤٤٨٣/٩٥
واریانس خطای مطلق	$\hat{\sigma}^2_{(\Delta)}$	١١٠٦٦٢٧/٨٥	٩٧٤٦١٤/٥٧	٥٦٧٥٢٨/٧٤	٣٧٣٠٤٠/٩٢
Gضریب تعمیم‌پذیری	$E\hat{\rho}^2_{\delta}$	٠/٣٧	−٠/٠٥	٠/٥١	٠/٤٦
Phiضریب اعتمادپذیری	$\widehat{\Phi}$	٠/٣٢	−٠/٠٥	٠/٣١	٠/٤٠
سهم ٪ در واریانس نمرهٔ جهانی	$\hat{\sigma}^2_{(\tau)}$	٧٠/٩٦	٤/٥٥	٢٤/٤٩	
سهم ٪ در واریانس خطای نسبی	$\hat{\sigma}^2_{(\delta)}$	٧٦/٢٨	١٨/٤٩	٥/٢٣	
سهم ٪ در واریانس خطای مطلق	$\hat{\sigma}^2_{(\Delta)}$	٧٤/١٦	١٦/٣٣	٩/٥١	

نکته: مطالعات تعمیم‌پذیری و تصمیم‌گیری با N = ٢٧٤

با توجه به سؤال ۱ این پژوهش مبنی بر اندازه ضرایب شبه پایایی معیارهای پذیرش دکتری در سال ۱۳۹۷ برای این رشته می‌توان گفت «بدون احتساب وزن اسمی این سنجه‌ها و تنها بر اساس وزن مؤثر آن‌ها» ضریب تعمیم‌پذیری $E\hat{\rho}^2_{\delta}$ برای آزمون‌های اختصاصی ۰/۳۷، برای آزمون‌های عمومی زبان و استعداد ۰/۰۵−، برای دو معیار مصاحبه ۰/۵۱ و ضریب تعمیم‌پذیری نمرهٔ جهانی ترکیبی برابر با ۰/٤٦ است. همچنین، ضریب اعتمادپذیری $\widehat{\Phi}$ به ترتیب برای آزمون‌های اختصاصی ۰/۳۲، برای خرده آزمون‌های عمومی زبان و استعداد ۰/۰۵−، برای دو معیار مصاحبه ۰/۳۱ و ضریب اعتمادپذیری نمرهٔ جهانی ترکیبی برای این برنامهٔ سنجش و پذیرش –در بهترین شرایط– برابر با ۰/٤٠ است.

سپس، در جداول ٦ و ۷ نتایج مربوط به مطالعات تصمیم‌گیری D_1 آمده، جایی که وزن‌های اسمی (یا همان ضرایب) به هر متغیر اختصاص داده شده است. انتظار می‌رود هنگامی که وزن‌های اسمی به متغیرها تعلق می‌گیرد، سهم تعداد لایه‌های هر متغیر بر اساس وزن اسمی آن در تعیین واحد جهانی یا جهان تعمیم تعدیل بیابد؛ در شرایط یکسان، هنگامی که لایه‌های متغیرها برابر با شد، متغیری که وزن اسمی بالاتری به آن داده می شود، سهم بیشتری در واحد جهانی یا جهان تعمیم پیدا می‌کند، و به دنبال آن در واریانس نمرهٔ جهانی ترکیبی نقش بیشتری ایفا می‌کنند و برعکس.

جدول ٦ نتایج مطالعات تصمیم‌گیری D_1 بر اساس طرح چند متغیره $p^{\bullet} \times i^{\circ}$ و
با همان اندازه نمونه‌ای (n_1=٤, n_2=٢, n_3=٢) و با وزن‌های اسمی ($W_1 = 0/3, W_2 = 0/2, W_3 = 0/5$)

D study			
P	544360/78	0/00	0/80
	56063/73	0/00	0/00
	307726/61	121455/72	265518/60
I	208103/66		
		103256/54	
			321240/23
PI	898524/19		
		871358/02	
			246288/50

نکته: مطالعات تعمیم‌پذیری و تصمیم‌گیری با N = ٢٧٤ و واریانس برآورد شده (قطری)،
کواریانس‌های مشاهده شده (پایین قطری) و همبستگی‌های مشاهده شده (بالای قطری)

بنابراین، با توجه به سؤال ۲ این پژوهش مبنی بر این‌که «آیا وزن‌های اسمی یا همان ضرایب می‌توانند به افزایش ضرایب شبه پایایی کمک کنند؟» هنگامی که وزن‌های اسمی به متغیرها یا همان معیارهای پذیرش اختصاص می‌یابد.

جدول ۷ نتایج مطالعات تصمیم‌گیری D_1 بر اساس طرح چند متغیره $p^{\bullet} \times i^{\circ}$ و
با همان اندازه نمونه‌ای (n_1=٤, n_2=٢, n_3=٢) با وزن‌های اسمی ($W_1 = 0/3, W_2 = 0/2, W_3 = 0/5$)

منابع واریانس		آزمون‌های اختصاصی $n_1 = 4$ $W_1 = 0/3$	آزمون‌های عمومی $n_2 = 2$ $W_2 = 0/2$	مصاحبه $n_3 = 2$ $W_3 = 0/5$	ترکیبی
واریانس نمرهٔ جهانی	$\hat{\sigma}^2_{(\tau)}$	٥٤٤٣٦٠/٧٨	۰/۰۰	٢٦٥٥١٨/٦٠	٢٣٨٧٠٨/٩٠
واریانس خطای نسبی	$\hat{\sigma}^2_{(\delta)}$	٨٩٨٥٢٤/١٩	٨٧١٣٥٨/٠٢	٢٤٦٢٨٨/٥٠	١٧٧٢٩٣/٦٣
واریانس خطای مطلق	$\hat{\sigma}^2_{(\Delta)}$	١١٠٦٦٢٧/٨٥	٩٧٤٦١٤/٥٧	٥٦٧٥٢٨/٧٤	٢٨٠٤٦٣/٢٨
Gضریب تعمیم‌پذیری	$E\hat{\rho}^2_{\delta}$	۰/۳۷	۰/۰۰	۰/۵۱	۰/۵۷
Phiضریب اعتمادپذیری	$\hat{\Phi}$	۰/۳۲	۰/۰۰	۰/۳۱	۰/٤٥
سهم در واریانس نمرهٔ جهانی	$\hat{\sigma}^2_{(\tau)}$	٤١/٢٧ درصد	٦/٥٠ درصد	٥٢/٢٣ درصد	
سهم در واریانس خطای نسبی	$\hat{\sigma}^2_{(\delta)}$	٤٥/٦١ درصد	١٩/٦٦ درصد	٣٤/٧٣ درصد	
سهم در واریانس خطای مطلق	$\hat{\sigma}^2_{(\Delta)}$	٣٥/٥١ درصد	١٣/٩٠ درصد	٥٠/٥٩ درصد	

نکته: مطالعات تعمیم‌پذیری و تصمیم‌گیری با N = ٢٧٤

با توجه به جدول ۷، ضریب تعمیم‌پذیری $E\hat{\rho}^2_{\delta}$ برای آزمون‌های اختصاصی ۰/۳۷ (بدون تغییر)، برای آزمون‌های عمومی زبان و استعداد ۰/۰۰ (به دلیل منفی بودن به صورت پیش گزیده صفر تعیین شده)، برای دو معیار مصاحبه ۰/۵۱ (بدون تغییر) و ضریب تعمیم‌پذیری نمرهٔ جهانی ترکیبی از ۰/٤٦ به ۰/۵۷ افزایش پیدا کرده است. همچنین، ضریب اعتمادپذیری $\hat{\Phi}$ به ترتیب برای آزمون‌های اختصاصی ۰/۳۲ (بدون تغییر)، برای آزمون‌های عمومی زبان و استعداد ۰/۰۰، برای دو معیار مصاحبه ۰/۳۱ (بدون تغییر) و ضریب اعتمادپذیری نمرهٔ جهانی ترکیبی برای این برنامهٔ

سنجش و پذیرش –در بهترین شرایط– از ۰/۴۰ به ۰/۴۵ افزایش پیدا کرده است. این افزایش ضریب $E\hat{p}_{\delta}^{2}$ و $\widehat{\Phi}$ نمرهٔ جهانی ترکیبی به دلیل برابر صفر قرار دادن واریانس منفی آزمون عمومی است، و افزایش معتبری نیست.

در جدول ۸ نتایج مطالعات طرح تصمیم‌گیری D_2 آمده است. جایی که همان وزن‌های اسمی به هر متغیر داده شده است، امّا با این تفاوت که لایه‌های متغیرها (به عنوان مثال آزمون‌های اختصاصی به ۵ لایه و مصاحبه به ۳ لایه) افزایش پیدا کرده است؛ چون دانشگاه‌ها حاضر نیستند سهم خود را در انتخاب داوطلبان کاهش دهند، پس طرح تصمیم‌گیری D_2 با افزایش لایه‌ها –بدون کاهش وزن اسمی یک متغیر به نفع متغیر دیگر– بهینه‌ترین و عملی‌ترین راه حل برای افزایش پایایی نمرهٔ جهانی ترکیبی است. البته به این استثناء نیز باید توجه داشت که افزایش لایه‌های هر آزمون به منظور افزایش ضرایب شبه پایایی ممکن است، برای رشته‌های مشابه نتایج یکسان و مناسبی به بار نیاورد. برخی از رشته‌ها به دلیل تعداد زیاد دروس اختصاصی (حتی تا ۸ خرده آزمون) نمرات پرخطایی برای تصمیم به پذیرش در دست دارند. همچنین، این میزان افزایش خرده آزمون‌ها از لحاظ هزینه و زمان نیز باید به صرفه باشد!

جدول ۸ نتایج مطالعات طرح تصمیم‌گیری D_2 بر اساس طرح چند متغیره $p^{\bullet} \times i^{\circ}$ با تغییر (افزایش) در اندازه نمونه‌ای[1] (n_1=۵, n_2=۲, n_3=۳) و همان وزنهای اسمی[2] ($W_1 = 0/3, W_2 = 0/2, W_3 = 0/5$)

منابع واریانس		آزمون‌های اختصاصی $n_1 = 5$ $W_1 = 0/3$	آزمون‌های عمومی $n_2 = 2$ $W_2 = 0/2$	مصاحبه $n_3 = 3$ $W_3 = 0/5$	ترکیبی
واریانس نمرهٔ جهانی	$\hat{\sigma}_{(\tau)}^{2}$	۵۴۴۳۶۰/۷۸	۰/۰۰	۲۶۵۵۱۸/۶۰	۲۳۸۷۰۸/۹۰
واریانس خطای نسبی	$\hat{\sigma}_{(\delta)}^{2}$	۷۱۸۸۱۹/۳۵	۸۷۱۳۵۸/۰۲	۱۶۴۱۹۲/۳۳	۱۴۰۵۹۶/۱۵
واریانس خطای مطلق	$\hat{\sigma}_{(\Delta)}^{2}$	۸۸۵۳۰۲/۲۸	۹۷۴۶۱۴/۵۷	۳۷۸۳۵۲/۴۹	۲۱۳۲۴۹/۹۱
Gضریب تعمیم‌پذیری	$E\hat{p}_{\delta}^{2}$	۰/۴۳	۰/۰۰	۰/۶۱	۰/۶۲
Phiضریب اعتمادپذیری	$\widehat{\Phi}$	۰/۳۸	۰/۰۰	۰/۴۱	۰/۵۲

نکته: مطالعات تعمیم‌پذیری و تصمیم‌گیری با $N = ۲۷۴$ که نتایج بهینه‌ترین و اقتصادی‌ترین آن‌ها در اینجا گزارش می‌شود.

بنابراین با توجه به سؤال پژوهشی ۳ مبنی براین‌که آیا «افزایش لایه‌های خرده آزمون‌های تخصصی و معیارهای مصاحبه» می‌تواند به افزایش ضرایب شبه پایایی کمک کند؟ با توجه به جدول ۸ نتایج نشان می‌دهند که ضریب تعمیم‌پذیری $E\hat{p}_{\delta}^{2}$ برای آزمون‌های اختصاصی از ۰/۳۷ به ۰/۴۳ (افزایش)، برای آزمون‌های عمومی زبان و استعداد ۰/۰۰ (به دلیل منفی بودن به صورت

1 . change in sample size
2 . same nominal weights

پیش گزیده صفر تعیین شده)، برای دو معیار مصاحبه از ۰/۵۱ به ۰/٦۱ (افزایش) و ضریب تعمیم‌پذیری نمرهٔ جهانی ترکیبی از ۰/۵۷ به ۰/٦۲ افزایش پیدا کرده است. همچنین، ضریب اعتمادپذیری $\widehat{\Phi}$ به ترتیب برای آزمون‌های اختصاصی از ۰/۳۲ به ۰/۳۸ (افزایش)، برای آزمون‌های عمومی زبان و استعداد ۰/۰۰، برای دو معیار مصاحبه از ۰/۳۱ به ۰/٤۱ (افزایش) و ضریب اعتمادپذیری نمرهٔ جهانی ترکیبی برای این برنامهٔ سنجش و پذیرش–در بهترین شرایط– از ۰/٤۵ به ۰/۵۲ افزایش پیدا کرده است.

جدول ۹ تفاوت نتایج مربوط به مقایسه تصمیم‌گیری D_1 از D_2 را نشان می‌دهد. به عنوان مثال واریانس خطای نسبی در آزمون‌های اختصاصی به مقدار ۱۷۹۷۰٤/۸٤– کاهش می‌یابد و ضرایب شبه پایایی این آزمون تخصصی به مقدار ۰/۰٦ افزایش پیدا می‌کند.

جدول ۹ تفاوت D_2 از D_1 هنگامی که لایه‌ها n_i آزمون‌های اختصاصی و مصاحبه افزایش می‌یابد و $N = ۲۷٤$

منابع واریانس		آزمون‌های اختصاصی $D_2 - D_1$	آزمون‌های عمومی $D_2 - D_1$	مصاحبه $D_2 - D_1$	ترکیبی $D_2 - D_1$
واریانس خطای نسبی	$\widehat{\sigma}^2_{(\delta)}$	۱۷۹۷۰٤/۸٤–	۰/۰۰	۸۲۰۹٦/۱۷–	۳٦٦۹۷/٤۸–
واریانس خطای مطلق	$\widehat{\sigma}^2_{(\Delta)}$	۲۲۱۳۲۵/۵۷–	۰/۰۰	۱۸۹۱۷٦/۲۵–	٦۷۲۱۳/۳۷–
ضریب تعمیم‌پذیری G	$E\widehat{\rho}^2_{\delta}$	۰/۰٦	۰/۰۰	۰/۱۰	۰/۰۵
ضریب اعتمادپذیری Phi	$\widehat{\Phi}$	۰/۰٦	۰/۰۰	۰/۱۰	۰/۰۷

جدول ۱۰ نتایج به‌دست آمده از طرح تصمیم‌گیری D_3 را نشان می‌دهد، که در آن نمرات معدل کار شناسی و ارشد همهٔ ٤٦۰ داوطلب دعوت شده به مصاحبه به عنوان متغیرهای مستقل وارد معادله شده‌اند. هدف این تحلیل رده‌بندی افراد بر اساس نمرات جهانی ترکیبی و مقایسه آن با رتبه‌های محاسبه شده در سازمان سنجش است، و در نهایت بررسی میزان خطاهای پیش‌بینی در طبقه‌بندی افراد به دو طبقه قبول و رد است.

جدول۱۰ نتایج مطالعات طرح تصمیم‌گیری D_3 بر اساس طرح چند متغیره $p^{\bullet} \times i^{\circ}$

اضافه کردن متغیر معدل با دو سطح به معادله پیش‌بینی و کاهش ده درصدی وزن مصاحبه

منابع واریانس		آزمون‌های اختصاصی $n_1 = 4$ $W_1 = 0/3$	آزمون‌های عمومی $n_2 = 2$ $W_2 = 0/2$	مصاحبه $n_3 = 2$ $W_3 = 0/4$	معدل $n_4 = 2$ $W_4 = 0/1$	ترکیبی
واریانس نمرهٔ جهانی	$\widehat{\sigma}^2_{(\tau)}$	۵٤٤۳٦۰/۷۸	۰/۰۰	۲٦۵۵۱۸/٦۰	۷٦٦٤/۸۰	۱۹۵۱٦۱/۱۳
واریانس خطای نسبی	$\widehat{\sigma}^2_{(\delta)}$	۸۹۸۵۲٤/۱۹	۸۷۱۳۵۸/۰۲	۲٤٦۲۸۸/۵۰	۵۸۱٤/۸۷	۱۵۵۱۸۵/۸۱
واریانس خطای مطلق	$\widehat{\sigma}^2_{(\Delta)}$	۱۱۰٦٦۲۷/۸۵	۹۷٤٦۱٤/۵۷	۵٦۷۵۲۸/۷٤	۱۱۳۱۵/۱۲	۲۲۹٤۹۸/۸۵
ضریب تعمیم‌پذیری G	$E\widehat{\rho}^2_{\delta}$	۰/۳۷	۰/۰۰	۰/۵۱	۰/۵٦	۰/۵۵
ضریب اعتمادپذیری Phi	$\widehat{\Phi}$	۰/۳۲	۰/۰۰	۰/۳۱	۰/٤۰	۰/٤۵

نکته: مطالعات طرح تصمیم‌گیری D_3 با $N = ٤٦۰$

با توجه به جدول ۱۰ و سؤال پژوهشی ٤ مبنی بر این‌که «آیا اضافه کردن نمرات معدل کارشناسی و ارشد به عنوان یک متغیر مستقل با دو لایه به جمع آزمون‌های تخصصی و عمومی و مصاحبه به افزایش ضرایب شبه پایایی نمرات کمک می‌کند؟»

پاسخ این سؤال با توجه به مقایسهٔ نتایج مطالعات تصمیم‌گیری D_3 با نتایج مطالعات تصمیم‌گیری D_1 گزارش شده است. در این راستا، ضریب تعمیم‌پذیری $E\hat{\rho}_\delta^2$ برای آزمون‌های اختصاصی ۰/۳۷ (بدون تغییر)، برای آزمون‌های عمومی زبان و استعداد ۰/۰۰ (به دلیل منفی بودن به صورت پیش گزیده صفر تعیین شده)، برای دو معیار مصاحبه ۰/۵۱ (بدون تغییر) و ضریب تعمیم‌پذیری نمرهٔ جهانی ترکیبی از ۰/٤٦ به ۰/۵۵ افزایش پیدا کرده است. همچنین، ضریب اعتمادپذیری $\widehat{\Phi}$ به ترتیب برای آزمون‌های اختصاصی ۰/۳۲ (بدون تغییر)، برای خرده آزمون‌های عمومی زبان و استعداد ۰/۰۰، برای دو معیار مصاحبه ۰/۳۱ (بدون تغییر) و ضریب اعتمادپذیری نمرهٔ جهانی ترکیبی برای این برنامهٔ سنجش و پذیرش –در بهترین شرایط– از۰/٤۰ به ۰/٤۵ افزایش پیدا کرده است. این افزایش ضریب $E\hat{\rho}_\delta^2$ و $\widehat{\Phi}$ نمرهٔ جهانی ترکیبی به دلیل برابر صفر قرار دادن واریانس منفی آزمون عمومی می باشد و افزایش معتبری نیست. همچنین، نمرات معدل دارای ضریب تعمیم‌پذیری $E\hat{\rho}_\delta^2 = 0/56$ و اعتمادپذیری $\widehat{\Phi} = 0/40$ است (به جدول ۱۰ نگاه کنید). از طرف دیگر، هدف طرح تصمیم‌گیری D_3 با احتساب معدل (جدول۱۰) وD_1 بدون احتساب معدل (جدول ۷) علاوه بر برآورد ضرایب شبه پایایی $E\hat{\rho}_\delta^2$ و $\widehat{\Phi}$ معیارهای پذیرش، رده‌بندی افراد بر اساس نمرات جهانی ترکیبی و مقایسه آن با رتبه‌های محاسبه شده از سوی سازمان سنجش نیز بوده است. بنابراین، با توجه به سؤال پژوهشی ۵، مبنی بر این‌که «بر اساس پذیرش در طبقات دو ارزشی میزان برآورد صحیح پیش‌بینی‌ها در رشتهٔ مهندسی برق– قدرت چقدر است؟»

برای پاسخ به این پرسش نمرات همهٔ داوطلبان (N = ٤٦۰) برآورد شده و بر اساس نمرهٔ جهانی ترکیبی پیش‌بینی شده از نتایج این دو طرح تصمیم‌گیری D_1 و D_3 دوباره داوطلبان رتبه‌بندی شدند. رتبه‌بندی جدید آن‌ها با نتایج واقعی پذیرش در سازمان سنجش مقایسه شدند، و فراوانی‌ها در جداول تصادفی ۲×۲ توزیع شده‌اند.

دو نوع خطایی که در مقایسه رتبه‌بندی افراد بر اساس نتایج واقعی پذیرش با نتایج مبتنی بر نمرهٔ جهانی ترکیبی پیش‌بینی شده ممکن است روی دهد عبارت‌اند از:

الف) خطای مثبت کاذب طبقه‌بندی هنگامی روی می‌دهد که به خطا یک داوطلب ضعیف به عنوان کسی که سطح معینی از پیشرفت دارد، به عنوان موفق و بالای نمرهٔ برش طبقه‌بندی شود.

ب) خطای منفی کاذب هنگامی روی می‌دهد که به خطا یک داوطلب قوی به عنوان کسی که سطح معینی از پیشرفت را ندارد، به عنوان ناموفق و پایین نمرهٔ برش طبقه‌بندی شود.

یک شاخص آماری معقول که می‌تواند دقت نتایج واقعی پذیرش را نشان دهد نرخ ضربه به هدف است که عبارت‌اند از:

ج) پیش‌بینی موفقیت برای فردی که موفق طبقه‌بندی شده و

د) پیش‌بینی عدم موفقیت برای فردی که ناموفق طبقه‌بندی شده

جمع این دو ارزش اخیر تقسیم بر تعداد کل، N = ٤٦٠، درصد پیش‌بینی‌های درست را برای داده‌های واقعی نشان می‌دهد.

همچنین، شاخصی است از دقت نمرات اولیه که به چه خوبی، به طور متوسط، به عنوان یک شاخص طبقه‌بندی‌کننده داوطلبان را انتخاب کرده‌اند. جداول توافقی ۱۱ و ۱۲، توزیع فراوانی نتایج واقعی و پروفایل نمرات پیش‌بینی شده از مطالعات تصمیم‌گیری D_1 و D_3 را نشان می‌دهد. جدول ۱۱ نشان می‌دهد که نرخ ضربه به هدف باید ترکیبی باشد، از کسانی که هم در نتایج واقعی و هم در نتایج پیش‌بینی شده برحسب پروفایل نمرات جهانی باید موفق می‌شده‌اند. به‌علاوه کسانی که هم در نتایج واقعی و هم در نتایج پیش‌بینی شده بر اساس پروفایل نمرات جهانی باید ناموفق طبقه‌بندی می‌شده‌اند تقسیم بر تعداد کل(N =٤٦٠)

جدول توافق ۱۱ مبتنی بر مطالعات تصمیم‌گیری D_1 و با همان اندازه نمونه‌ای n،
و وزن‌های اسمی[1] ($W_1 = 0/3 , W_2 = 0/2, W_3 = 0/5$) و N = ٤٦٠

		نتایج پیش‌بینی شده D_1		جمع
		موفق	ناموفق	
نتایج واقعی	موفق	۱۷۱	۷	۱۷۸
	ناموفق	۷	۲۷۵	۲۸۲
جمع		۱۷۸	۲۸۲	٤٦٠

بنابراین، نرخ ضربه به هدف برای جدول توافق ۱۱ برابر است با:

$$HitRate = \frac{171 + 275}{460} = 0/97$$

و با افزودن متغیر معدل با دو لایه و وزن اسمی ۰/۱ و کاهش همین مقدار از وزن مصاحبه، فراوانی‌ها در جدول توافق ۱۲ خلاصه شده است.

1 . same sample size and nominal sample weights

جدول توافق ۱۲ مبتنی بر مطالعات تصمیم‌گیری D_3 و با افزودن متغیر معدل به معادله پیش‌بینی
و وزن‌های اسمی $W_1 = 0/3\ , W_2 = 0/2, W_3 = 0/4,\ W_4 = 0/1$ و $\mathbf{N = ٤٦٠}$

		نتایج پیش‌بینی شده D_3		جمع
		موفق	ناموفق	
نتایج واقعی	موفق	۱۷۰	۸	۱۷۸
	ناموفق	۸	۲۷٤	۲۸۲
جمع		۱۷۸	۲۸۲	٤٦٠

نرخ ضـــربه به هدف با احتســـاب متغیر معدل در معادله پیش‌بینی برای جدول توافق ۱۲ برابر است با:

$$HitRate = \frac{170 + 274}{460} = 0/965$$

بر ا ساس نتایج واقعی پذیرش و مقایسهٔ آن با نمرات جهانی پیش‌بینی شده مشخص شد، که به‌طور کل اگر با یک مدل قوی‌تر نمرات افراد پیش‌بینی شـــود، تنها در ۳ درصـــد موارد رتبه‌بندی افراد جا به جا خواهد شـــد؛ برای داده‌های بدون معدل تقریبا ۷ جا به جایی و برای داده‌هایی با اضـــافه شـــدن معدل تقریبا ۸ داوطلب جابه جا می‌شـــوند؛ به عبارتی اگر همچنان نمرات معدل در بخش مصـــاحبه نمره‌گذاری شـــوند، میزان برآورد صـــحیح پیش‌بینی‌ها ۰/۹۷ خواهد بود. این ۰/۰۳ درصد جابه‌جایی افراد نیز در وسط توزیع رتبه‌ها رخ داده است. به‌علاوه، این مهم است که بدانیم نمره‌گذاری سؤالات آزمون‌های اختصاصی و عمومی با کم کردن ضریب حدس شانسی تغییرات (خطای) زیادی در این جابه‌جایی‌ها ایجاد می‌کند نسبت به هنگامی که پاسخ سؤالات به صورت ۰ و ۱ نمره‌گذاری می‌شوند.

به طور کل، شـــانس خطای مثبت و خطای منفی تنها در ۰/۰۳ درصـــد فرایند انتخاب این ۳۷ بخش دانشـــگاه‌های ملی که پذیرش دانشـــجوی مهندســـی برق قدرت داشـــته‌اند روی داده اســـت. هنگامی که متغیر معدل اضـــافه می‌شـــود نرخ ضـــربه به هدف تغییری نمی‌کند؛ عملا نمرهٔ معدل کارشناسی و کارشناسی ارشد به عنوان متغیر مستقل، پیش‌بینی‌کننده‌های مناسبی نیستند. این نمرات به دلیل تورم در بالای توزیع و چولگی چپ و محدود بودن دامنهٔ آن در ترکیب با پیش‌بینی‌کننده‌های دیگر، توانایی ایجاد واریانس واقعی بسـیار کمی دارند. امّا به هر حال معدل نقش خود را در غربال اولیه افراد برای ارایه مدرکی دال بر صلاحیت داوطلب شدن ایفا می‌کنند. چون معدل‌های بالای ۱۲ در لیسانس و ۱٤ در فوق لیسانس شرط الزامی برای شرکت در برنامهٔ پذیرش دکتری است.

نتایج آزمون‌های اختصاصی

آزمون‌های تخصصی در رشتهٔ مهندسی برق قدرت دارای ضریب تعمیم‌پذیری ۰/۳۷ و اعتمادپذیری ۰/۳۲ است، که ضریب مناسبی برای آزمون‌های سرنوشت ساز نیست. در تصمیم‌گیری‌های سرنوشت‌ساز بر اساس نمرات یک چنین آزمون‌هایی که ضرایب تعمیم‌پذیری کوچکی دارند باید دقت بیشتری مدنظر داشت. چهار خرده مقیاس این آزمون به ترتیب ۷٦/۲۸ درصد خطای نسبی و ۷٤/۱٦ درصد خطای مطلق نمرهٔ جهانی ترکیبی را به خود اختصاص داده است (به جدول ۵ نگاه کنید). با کاهش وزن مؤثر آن‌ها در مطالعات تصمیم‌گیری D_1 که وزن اسمی برابر با ۰/۳ برای ٤ لایه آزمون اختصاصی آن درنظر گرفته شده، خطای نسبی به ٤۵/٦۱ درصد و خطای مطلق به ۳۵/۵۱ درصد کاهش پیدا کرده است (جدول ۷).

نتایج آزمون عمومی استعداد و زبان انگلیسی

آزمون‌های عمومی زبان انگلیسی و استعداد ضریب تعمیم‌پذیری ۰/۰۵- و اعتمادپذیری ۰/۰۵- دارند. این دو آزمون به ترتیب ۱۸/٤۹ درصد خطای نسبی و خطای مطلق۱٦/۳۳ درصد نمرهٔ جهانی ترکیبی را به خود اختصاص داده اند (به جدول ۵ نگاه کنید). هنگامی که در مطالعات تصمیم‌گیری D_1 وزن اسمی آن‌ها که در مجموعه آزمون برابر با ۰/۲ است تعیین می‌شود، خطای نسبی و خطای مطلق این دو آزمون به ترتیب ۱۹/٦٦ درصد و ۱۳/۹۰ درصد تغییر می‌کند (به جدول ۷ نگاه کنید). بیشتر آزمون شوندگان سؤالات استعداد را به دلیل سختی و آزمون زبان انگلیسی را به دلیل کم اهمیت شدن آن در مرحله پذیرش بدون پاسخ رها می‌کنند. بدین ترتیب وزن دهی به آزمون‌های پرخطا حتی ممکن است شانس پیش‌بینی‌های درست را کاهش.

نتایج مصاحبه

مصاحبه از یازده سنجه متفاوت کمی و کیفی تشکیل شده که در دو نمرهٔ خلاصه می شود. ضریب تعمیم‌پذیری و اعتمادپذیری آن به ترتیب ۰/۵۱ و ۰/۳۱ است. این متغیر به ترتیب ۵/۲۳ درصد و ۹/۵۱ درصد سهم خطای نسبی و خطای مطلق نمرهٔ جهانی ترکیبی را به خود اختصاص داده‌اند (به جدول ۵ نگاه کنید). در مطالعات تصمیم‌گیری D_1 که وزن اسمی برابر با ۰/۵ برای این متغیر درنظر گرفته شده خطای نسبی و خطای مطلق آن به ۳٤/۷۳ درصد و ۵۰/۵۹ درصد افزایش پیدا کرده است (به جدول ۷ نگاه کنید).

بحث در نتایج آزمون اختصاصی

میانگین پایین هر یک از مقیاس‌های چهارگانه آزمون اختصاصی نشان‌دهندهٔ سختی سؤالات این خرده آزمون‌ها است (به جدول ۳ نگاه کنید). سؤالات خرده آزمون‌هایی که بسیار دشوار هستند و یا در بین متقاضیان خصیصه‌های متفاوتی را اندازه می‌گیرند، سنجه‌هایی با ثبات برای انتخاب

افراد نیستند (سباک و مک میلان[1]، ۲۰۱٤، ص. ۱۱۰). چون نمرهٔ ترکیبی آزمون اختصاصی سرجمع وزنی خرده آزمون‌های تخصصی مختلف است، هنگامی که یک خرده آزمون به دلیل دشواری سؤالات رها می‌شود، نمرات بالا روی خرده آزمون‌های دیگر می‌تواند آن را جبران کند. بدین ترتیب نقاط قوت و ضعف متقاضیان و ویژگی‌های روانسجی هر آزمون پنهان می شود. به هرحال، خرده آزمون‌های اختصاصی منبع بالقوه مهمی برای خطای اندازه‌گیری هستند. هر اندازه پراش بین خرده آزمون‌ها[2] بالا باشد، و ثبات عملکرد[3] افراد داوطلب روی خرده آزمون‌ها پایین تر باشد، به ویژه به دلیل ضریب تصحیح حدس به تعداد سؤال و خرده آزمون بیشتری نیاز هست تا خطا کاهش یابد (جانسون و جانسون[4]، ۲۰۱۰، ص.۷٦). هم‌زمان دشواری زیاد سؤالات اجازه نمی‌دهد لایه‌های آزمون افزایش پیدا کنند. بنابراین، لازم است سطوح دشواری سؤالات خرده آزمون‌های فعلی با سطوح توانای داوطلبان هماهنگ شود. هنگامی که برای پذیرش سؤالاتی به کار می‌روند که با سطوح توانایی آزمون شوندگان به خوبی هماهنگ نیست شانس انتخاب افراد نامناسب نیز افزایش می‌یابد (سباک و مک میلان، ۲۰۱٤).

همچنین، تحلیل‌های کیفی سؤالات آزمون‌های اختصاصی نشان داده‌اند بیشتر سطوح پایین رشد شناختی مثل دانش و فهمیدن و نهایتاً به کاربستن تکالیفی که در دوران تحصیل خوب آموزش داده شده‌اند را اندازه می‌گیرند. سطوح بالای شناختی در آزمون‌های چهارگزینه‌ای اختصاصی و عمومی به خوبی آزمون‌هایی که پیشتر دانشگاه‌ها اجرا می‌کردند، و به صورت سهمی نمره‌گذاری می شدند، قابل بررسی نیستند. برنامهٔ سنجش دانشگاه‌های مهندسی پیشتر مبتنی بر آزمون‌هایی بودند که سؤالات آن فرایندهای حل مسئله[5]، فرایندهای یکپارچه سازی و مهارت‌های فنی(سووان و بورک هارت[6]، ۲۰۱۲) را اندازه می‌گرفتند[7]. حل این مسئله‌ها و تکالیف داوطلبان را ملزم می‌کرد که سطوح بالای توانایی‌های شناختی را به نمایش بگذارند، که ارتباط نزدیکتری با دستاوردهای با ثبات آموزشی مثل تولید علم و نوآوری دارد. با اجرای دوگانه برنامهٔ سنجش نیمه متمرکز فعلی، سؤالات آزمون‌های اندیشایی به صورت چهارگزینه‌ای شدند و جای خود را به محفوظات زودگذر داده‌اند. این امر باعث شده احتمالاً افرادی که خوب تست می‌زنند شانس بیشتری داشته باشند، برای ورود به مرحلهٔ مصاحبه و پذیرش.

---

1 . Sebok, MacMillan
2 . between-test variation
3 . performance consistency
4 . Johnson & Johnson
5 . problem solving ‘processes’
6 . Swan & Burkhardt
7 . assess problem solving, integrating processes and technical skills

بحث در نتایج آزمون عمومی

سؤالات چهارگزینه‌ای *آزمون عمومی استعداد* عمدتاً به دلیل دشواری زیاد و وزن کم در مجموعه آزمون رها می شوند. شکل سؤالات آزمون‌های استعداد در برگیرندهٔ گستره‌ای از انواع سؤالاتی است که با هدف بیرون کشیدن پاسخ‌های آگاهی‌دهنده از داوطلبان است، و عمدتاً پاسخ‌گویی به آن‌ها ربطی به موضوعات آموزشی در دوران تحصیل ندارد (کرایتون[1]، ۲۰۰۳). پاسخ‌گویی به سؤالات آن بسیار وقت‌گیر است. همچنین، نمره‌گذاری هر سؤال با تصحیح حدس، این خرده آزمون را به مقیاسی کاملاً ناکارآمد تبدیل کرده است. دشواری بالای سؤالات و روی آوردن آزمون شوندگان به حدس شانسی باعث شده این مقیاس انعکاسی از خطای اندازه‌گیری باشد، و به دلیل رها کردن و بی‌پاسخ گذاشتن سؤالات آن آماره‌هایی مثل ضریب تمیز، پایایی و دشواری این مقیاس‌ها قابل بررسی نباشند. از طرف دیگر تهیه‌کنندگان این آزمون[2] عمدتاً نمی‌دانند در تهیه سؤالات آن چه خرده مهارت‌ها یا فرایندهای شناختی را هدف‌گذاری کرده‌اند. ساختار چهارگزینه‌ای سؤالات آن در بهترین صورت آزمون شونده را ملزم به یادآوری یا بازشناسی، یا استنباط ضمنی روی چهارگزینه می‌کند.

بر خلاف گذشته در سال‌های اخیر *آزمون عمومی زبان انگلیسی* به عنوان یک معیار الزامی برای پذیرش نیست. داوطبان می‌توانند به سؤالات این آزمون پاسخ ندهند، امّا پس از پذیرفته شدن–در دوران تحصیل– از آزمون‌های مشابهی که مؤسسات مختلف اجرا می‌کنند، و مشابه آزمون تافل یا ایلتس هست یک نمرهٔ قبولی ارایه دهند تا اجازه شرکت در آزمون جامع را داشته باشند. این امر باعث افزایش طول دوران تحصیل پذیرفته‌شدگانی شده است که نتوانسته‌اند به موقع نمرهٔ قبولی زبان انگلیسی دریافت کنند. پیامد آن افزایش هزینه و زمان برای نظام آموزش عالی و هم‌زمان کاهش توان تولید علم پذیرفته شدگان به دلیل ضعف در زبان بین المللی تولید علم است. همچنین، بی‌اهمیت شدن این آزمون باعث شده کلیهٔ داوطلبان دورهٔ دکتری نسبت به گذشته با پیشینهٔ زبان انگلیسی ضعیف‌تری وارد این دوره‌ها بشوند.

نمره‌گذاری دستاوردهای داوطلبان در فرایند مصاحبه

بیشتر کمیته‌های پذیرش ادعا می‌کنند که سنجه‌های مصاحبه که عمدتاً فهرست کار[3] و سنجه‌های غیرشناختی داوطلبان است برای آن‌ها بسیار مهم هستند. امّا عملیاتی کردن این سنجه‌ها برای آن‌ها بسیار مشکل است، و نمی‌دانند چگونه کیفیت سوابق علمی پژوهشی و ویژگی

---

1 . Crighton

2 . test developer

3 . portfolio

غیرشناختی و شخصیتی داوطلبان را نمره‌گذاری کنند. این سنجه‌های کیفی باید بر اساس یک دستورالعمل نمره‌گذاری شوند، امّا عمدهٔ کمیته‌های پذیرش دستاوردهای داوطلبان را به خوبی نمره‌گذاری نمی‌کنند؛ نمونهٔ مقالات داوطلبان در جلسات کوتاه مصاحبه خوانده نمی‌شود، و معنای صلاحیت[1]، استعداد[2] و شایستگی[3] و ابعاد آن برای کمیته‌های پذیرش هنوز مشخص نیست. این مسئله هنوز باقی مانده که پذیرش نهایی را به افراد باصلاحیت بدهند: یعنی کسانی که روی آزمون‌های اختصاصی که مبتنی بر دوره‌های آموزشی استاندارد است، و در آن انتظارات یادگیری مشخص و استاندارد تعریف شده که افراد باید یک نمرهٔ مشخص در آن کسب کنند (گیل و راگلند، ۲۰۱۸)، یا به افرادی بدهند که در سنجه‌های مصاحبه نمرات بهتری کسب کرده‌اند. به هر حال برای افزایش ضریب پایایی معیارهای مصاحبه، بهتر است لایه‌های دوگانه آن به چند لایه بیشتر تفکیک شود. به عنوان مثال، بخشی از معیارهای مصاحبه دربرگیرندهٔ سنجه‌های غیرشناختی داوطلبان است. این سنجه‌های عمدتاً شخصیتی در فرایند مصاحبه می‌تواند به عنوان یک «پیش‌بینی‌کنندهٔ سوم» به طور مجزا نمره‌گذاری شود، به گونه‌ای که ویژگی‌هایی غیرشناختی همچون پشتکار، نوآوری، روحیهٔ همکاری و سخت کوشی را در بر بگیرند. عمدتاً رگه‌هایی از این ویژگی‌ها را می‌توان با دقت بیشتر در کیفیت و کمیت پیشینهٔ علمی و پژوهشی فرد مشاهده کرد.

بحث در نتایج معدل

نمرات معدل کارشناسی و کارشناسی ارشد در دامنه‌ای از ۰ تا ۲۰ معیارهای دیگری هستند که اهمیت کمتری در فرایند پذیرش دکتری به عنوان یک متغیر مستقل پیدا کرده‌اند. اگر چه برای شرکت در آزمون‌های دانشگاه‌های ملی معدل کارشناسی باید بالاتر از ۱۲ و معدل کارشناسی ارشد باید بالاتر از ١٤ باشد. امّا به دلیل همانندی نمرات معدل افراد با اعتبار دانشگاهی متفاوت در فرایند پذیرش به عنوان یک متغیر مستقل در معادله پیش‌بینی اعتماد کمتری به آن‌ها می‌شود.

تقریباً هشت نوع دانشگاه مختلف ملی و مستقل در ایران وجود دارد، که از لحاظ کیفیت و شیوهٔ آموزش و پرداخت هزینه‌های تحصیلی با یکدیگر تفاوت دارند. به دلیل نداشتن استانداردهای یادگیری و انتظارات مشخص در دانشگاه‌ها (گیل و راگلند، ۲۰۱۸) از یک طرف، و به دلیل یکسانی مقیاس نمره‌گذاری معدل از طرف دیگر نمی‌توان به خوبی مشخص کرد، که افرادی با معدل مشابه (مثلا با معدل ۱۵) امّا از دانشگاه‌های مختلف تا چه حد می‌توان نوع عملکردی را که در دورهٔ دکتری از آن‌ها انتظار می‌رود، پیش‌بینی کرد. به همین دلیل سهم معدل‌ها را در جلسه مصاحبه

1 . competence
2 . aptitude
3 . merit

مشخص می‌کنند. به هرحال، برتری با معدل افرادی است که از دانشگاه‌های ملی[1] فارغ التحصیل شده اند. بنابراین، نمرات معدل به عنوان یک سنجهٔ مستقل آگاهی کمی از پیش‌بینی عملکرد آینده داوطلبان به دست اندرکاران کمیتهٔ پذیرش می‌دهد.

در این پژوهش فرض بر این بود که با اضافه کردن معدل‌های کارشناسی و کارشناسی ارشد به معادله پیش‌بینی و مقایسه اثربخشی[2] پیش‌بینی‌کنندگی دو معادله، مقادیری تحت عنوان ارزش افزوده با استفاده از معدل (زوویک[3]، ۲۰۰۷، ص.۱۳) به ما ارایه می‌دهد. در این پژوهش عملاً معدل به دلیل اثر سقف در پیش‌بینی پذیرش افراد کارایی کمی داشت. هنگامی که دامنه توزیع نمرات کوتاه می‌شوند احتمال این‌که داوطلبان به اشتباه رتبه‌بندی شوند افزایش می‌یابد، به ویژه هنگامی که خطای اندازه‌گیری بالا است، و فاصله‌های حدود رتبه‌ها با یک نمرهٔ تک رقمی اندازه گرفته شده است (جانسون و جانسون، ۲۰۱۰، ص.۷۵). امّا معنی‌دار نبودن معدل ضرورتاً دلالت بر این ندارد که آن‌ها مشخصه‌ها یا ویژگی‌های عملکرد را اندازه نمی گیرند، بلکه واریانس کوچک آن‌ها به دلیل چولگی منفی و تجمع در بیشترین مقدار ممکن[4] در بالای یک توزیع حاصل از نمرات آزمون‌های ملاک مرجع در دوران تحصیل است (اتیه، ۱۹۹۹، ص.۳۵ و ۳۶).

بحث در نمره‌گذاری فرمولی

در برنامهٔ آزمون دوگانه فعلی با محوریت سازمان سنجش، به دلیل حجم بسیار زیاد شرکت‌کنندگان سؤالات باز پاسخ آزمون‌های پیشین دانشگاه‌ها جای خود را دادند، به سؤالات چهارگزینه ای و نمره‌گذاری آن‌ها به صورت ماشینی با حذف عامل شانس است، که به نمره‌گذاری فرمولی مشهور است. این نوع نمره‌گذاری باعث افزایش واریانس نمرات و تغییرات بیشتر در بین نمرات افراد می‌شود. امّا در نمره‌گذاری فرمولی سؤالات چهارگزینه‌ای آزمون‌های اختصاصی و عمومی با این‌که واریانس نمرات مشاهده شده بسیار افزایش می‌یابد، به خوبی نمی‌تواند واریانس نمرات واقعی را تبیین کند. همان طور که مگنسون[5] (۱۹۶۷) بیان کرده، بالا بودن مقدار واریانس نمرات مشاهده شده، فی نفسه مورد توجه نیست، بلکه مهم این است که آزمون بتواند نشان دهد که تمایز حاصل از اجرای آزمون در بین داوطلبان دقیق، معنی‌دار و پایا است. افزایش واریانس نمرات مشاهده شده برای تضمین پایایی به تنهایی کافی نیست. واریانس خطای اندازه‌گیری نیز مهم است.

---

1 . state college
2 . effectiveness
3 . Zwick
4 . maximum possible value
5 . Magnoson

در برنامهٔ سنجش غیرمتمرکز پیشین با محوریت تک تک دانشگاه‌ها، اعضای کمیته پذیرش سؤالات باز پاسخ آزمون‌های خود را به صورت سهمی نمره‌گذاری می‌کردند. بنابراین، برای سنجش دانش جزئی[1] آزمون شوندگان، به پاسخ درست آن‌ها روی هر بخش از تکلیف امتیازی داده می شد. به عنوان مثال، به پاسخ کاملاً غلط صفر، بخشی از پاسخ نمرهٔ یک، به بخشی دیگر نمرهٔ دو و به پاسخ کاملاً صحیح نمرهٔ سه تعلق می‌گرفت، و به عبارتی نمره‌دهی به صورت سیاه-سفید نبود، و آزمون شونده یک اعتبار سهمی بر اساس سطوح عملکرد موفقیت‌آمیز روی تکلیف دریافت می‌کرد. چنین نمره‌گذاری برآورد دقیقتری از توانایی یک آزمون شونده به دست می‌داد (یو[2]، ۱۹۹۱؛ سووان و بورک هارت، ۲۰۱۲).

در شیوهٔ نیمه متمرکز آنچه که در فرایند نمره‌گذاری سؤال بر اساس نمره‌گذاری فرمولی اتفاق می‌افتد قطعیت پاسخ مطرح است؛ اگر پاسخ آزمون‌دهنده درست باشد نمرهٔ یک دریافت می‌کند، رها کردن سؤال هم باعث نمرهٔ صفر می شود، و هیچ راه میانه‌ای برای بیشینه کردن احتمال پاسخ فرد بر اساس دانش جزئی او و برآورد دقیق‌تر نمرهٔ واقعی وجود ندارد. در حقیقت نمرهٔ منفی برای حذف عامل حدس، نه واریانس واقعی سؤال، و نه واریانس توانایی جامعه آزمون شوندگان است بلکه بیشتر واریانس عدم اطمینان است که در دامنه واریانس خطا قرار می‌گیرد؛ از دیدگاه احتمالات می‌توان ثابت کرد که هنگامی به افراد اطلاع داده می‌شود حدس صحیح سه امتیاز و حدس غلط یک امتیاز منفی دارد، عملاً به دلیل سختی سؤالات و هم‌زمان سرنوشت ساز بودن آن، آزمون شوندگان به اضافه کردن این مقدار واریانس خطا تشویق می شوند، که باعث کاهش پایایی آزمون می شود. به هرحال راه میانه این است که از سرجمع سادهٔ نمرات سؤالات (صفر و یک) برای نمره‌دهی استفاده شود. در نمره‌گذاری سرجمع تعداد پاسخ‌های درست[3]، با این‌که قطعیت پاسخ صحیح مطرح است، و عملاً مثل نمره‌گذاری سهمی هیچ شیوه‌ای برای اندازه‌گیری دانش جزئی وجود ندارد، نسبت به نمره‌گذاری فرمولی واریانس خطای کمتری تولید می‌شود (ذوالفقارنسب، خدایی و یادگارزاده، ۱۳۹۱).

نتیجه‌گیری

برنامه‌های آزمون‌گیری دکتری که پیامدهای معنی‌دار و بازگشت ناپذیری[4] برای متقاضیان دارد درجه بالایی از پایایی برای آزمون‌ها و پیش‌بینی‌کننده‌های آن لازم است (APA, AERA, NCME; 2014, p.42). امّا چون روش‌های متفاوتی برای برآورد پایایی این سنجه‌ها و

1 . partial knowledge
2 . Yu
3 . number-right scoring
4 . significant consequences and irreversible

پیش‌بینی‌کننده‌ها وجود دارد، و هر یک تحت تأثیر منابع مختلفی از خطای اندازه‌گیری است، نمی‌توان به سادگی در یک برنامهٔ آزمون‌گیری تنها یک ضریب پایایی تعیین کرد (APA, AERA, NCME;2014, p.47). بنابراین، بسته به ساختار برنامهٔ سنجش، نوع آزمون‌ها و مقیاس پیش‌بینی‌کننده‌ها و مفرو ضه‌هایی که تو سط محقق ایجاد می شود، در رابطه با این‌که تعامل کدام فاکتورها در واریانس خطا مشارکت دارند و کدام ندارد واریانس خطا می‌تواند پیچیده باشد (جانسون و جانسون،۲۰۱۰).

عموماً ضرایب پایایی آزمون‌ها و پیش‌بینی‌کننده‌ها در فرایند سنجش و پذیرش دکتری تحت تأثیر گستردگی تعریف سازهٔ زیربنایی، میزان دشواری تکالیف (سباک و مک میلان، ۲۰۱٤)، تعداد سؤالات خرده آزمون‌ها، شیوهٔ نمره‌گذاری سؤالات، وزن w خرده آزمون‌ها و ویژگی‌های آماری پیش‌بینی‌کننده‌هاn، به‌علاوه حجم نمونه N آزمون شوندگان، پراکندگی گروه آزمون شونده و ویژگی‌های دموگرافیک (لی و برنان، ۲۰۰۷) آن‌ها قرار دارد.

تحت شرایط یکسان هر چه‌قدر سازهٔ زیربنایی گسترده‌تر تعریف شود–امری که ناگزیر در برنامه‌های آزمون‌گیری دکتری اتفاق می‌افتد– مؤلفه‌های خطای بیشتری در برخواهد گرفت. هرچه خطای کل[1] زیاد باشد، کمتر می‌توان از روی نمرات مشاهده شده استنباط‌های مطمئن دربارهٔ نمرات واقعی آزمون شوندگان کرد. اگر چه تعریف گسترده‌تر از سازهٔ زیربنایی به منظور افزایش روایی بالقوه بهتر از تعریف محدود آن است، امّا برآورد سازه‌هایی که گسترده تعریف شده‌اند و با ابزارهای اندازه‌گیری چندگانه عملیاتی و اندازه‌گیری شده‌اند، در معرض خطای بیشتری قرار دارند تا برآورد سازه‌هایی که محدود تعریف شده‌اند (کان،۲۰۱۰).

همچنین، پایایی پیش‌بینی‌کننده‌ها زمانی اهمیت بیشتری پیدا می‌کند که باید یک نقطه برش برای تقسیم کردن دامنهٔ نمرات انتخاب شود، تا بتوان بر اساس آن افراد را طبقه‌بندی کرد (APA, AERA, NCME; 2014, p.101). فرایند سنجشی که از چندین پیش‌بینی‌کننده و سنجهٔ متفاوت تشکیل شده، و هر یک در نقطه برش درجهٔ متفاوتی از خطای اندازه‌گیری دارند ممکن است منجر به تصمیم‌گیری‌های نادرست برای طبقه‌بندی افراد شود. چون این خطاهای اندازه‌گیری در امتداد مقیاس نمرات، و به ویژه روی نقطه برش نمرات به طور یکسان یا یکنواخت توزیع نشده‌اند (National Research Council, 2002).

یکی از محدودیت های این پژوهش وجود نمرات منفی آزمون شوندگان بر برخی خرده آزمون‌ها بود که باعث می شد کلیهٔ اطلاعات آن‌ها کنار گذاشته شود. اگر چه در نهایت بهترین‌ها با نمرات بالا انتخاب می شوند، امّا از دست دادن اطلاعات بخشی از گروه نمونه به دلیل کارکرد بد

---

1 . overall error

روی خرده آزمون‌ها، پارامترهای گروه نمونه روی کل آزمون را نیز متأثر می‌سازد. از طرف دیگر دشواری بالای سؤالات آزمون‌های عمومی زبان و استعداد و روی آوردن آزمون شوندگان به حدس شانسی باعث شده این آزمون‌ها به جای این‌که مقیاسی برای اندازه‌گیری توانایی باشند، بازتابی از خطای اندازه‌گیری باشند، یا به دلیل بی‌پاسخ گذاشتن و رها کردن سؤالات آن‌ها آماره‌هایی مثل ضریب اعتمادپذیری، تعمیم‌پذیری و دشواری این خرده آزمون‌ها قابل بررسی نباشد. به همین دلیل اثربخشی این آزمون‌ها در تشخیص و تفکیک افراد آن‌چنان که باید قابل بررسی نبود.

از دیگر محدودیت‌ها این بود که یازده سنجه‌ای که مصاحبه را تشکیل می‌دهند، جنبه‌های مختلفی از عملکرد افراد را اندازه می‌گیرند، امّا تنها به دو نمرهٔ علمی و سنجش علمی (با ۳ سنجه) و سوابق آموزشی و پژوهشی و فناوری (با ٦ سنجه) تفکیک شده‌اند. مشخص نیست که نخست، خرده آزمون‌های تخصصی دقیقاً با کدام بخش از سنجه‌های یازده‌گانهٔ مصاحبه واریانس مشترک (همبستگی) دارد، و می‌تواند به خوبی آن‌را پیش‌بینی کند. دوم، این ۱۱ سنجه ممکن است آلوده به متغیرهای دیگری (مثل سخت گیری یا آسان گیری مصاحبه‌گران) نیز باشد. بنابراین، نمره‌ای که یک داوطلب روی دو بخش مصاحبه کسب می‌کند، ممکن است دقیقاً بازتاب توانایی واقعی او روی سازه مورد نظر (موفقیت در آینده، تولید علم و یا نوآوری) نباشد.

به هرحال نیمه متمرکز شدن آزمون‌های دکتری اثربخش نبوده است. به این معنی که نیمه مستقل شدن دانشگاه‌ها در فرایند سنجش و پذیرش باعث شده آزمون‌های تشریحی آن‌ها به آزمون‌های چهارگزینه ای تبدیل شوند. آزمون‌های چهارگزینه ای نمی‌توانند به خوبی فرایندهای شناختی سطح بالا را اندازه‌گیری کنند. نمره‌گذاری سؤال‌ها با حذف اثر شانس باعث شده نمرات این آزمون‌ها واریانس خطای بیشتری داشته باشند. آزمون‌های پذیرش تا آنجا که امکان دارد باید تشریحی و با نمره‌گذاری سهمی پاسخ‌ها باشند. نمرهٔ آزمون زبان برای همهٔ رشته‌ها باید پیش از پذیرش نهایی اهمیت پیدا کند، و سطوح دشواری آزمون‌های استعداد باید تعدیل شود، و طراحان این نوع آزمون‌ها مشخص کنند هر سؤال کدام بخش از مهارت‌های شناختی (مثلاً انواع استدلال عددی، استدلال تحلیلی، قضاوت و یا نظایر آن) را اندازه می‌گیرند. این پژوهش نشان داد اگر چه وجود آزمون زبان و استعداد ضروری است، امّا ضریب پایایی آن‌ها صفر است، و واریانسی از توانایی داوطلبان برآورد نمی‌کنند و کلاً برآیندی از واریانس خطا بوده‌اند.

«نتایج کلی» این پژوهش برای دیگر رشته‌های دانشگاهی نیز استفاده شدنی هست. به عنوان مثال پیشنهاد می‌شود با بهبود بخشیدن به فرایند سنجش و پذیرش، نظیر برگشتن به شیوهٔ غیرمتمرکز و استقلال دانشگاه‌ها در اجرای برنامهٔ سنجش و پذیرش همانند سال‌های پیش از ۱۳۹۰، آموزش اعضای کمیته‌های پذیرش در نمره‌گذاری سهمی تکالیف و دستاوردهای داوطلبان، تهیه آزمون‌هایی با

مجموعه تکالیف مؤثر[1] و بر اســـاس یک چارچوب تئوریک قوی مثل الگوی ژرفای دانش وب و رده‌بندی بلوم و با نمونه‌گیری کامل‌تری از سازه‌های اندازه‌گیری شده شیوه‌های سنجش و پذیرش را اثربخش‌تر کرد، و هم‌زمان خطای اندازه‌گیری را کاهش داد (کان، ۲۰۱۰، ص.۱۹ و ۲۰). همچنین، با اهمیت‌تر شـدن نمرات آزمون‌های عمومی اسـتعداد و به‌ویژه زبان انگلیسـی در پذیرش به عنوان آزمون‌های اولیه در یک سنجش دو مرحله‌ای (اسمیت، وان در آرک و کانیین، ۲۰۱۸) و نمره‌گذاری سهمی سؤالات باز پاسخ آزمون‌ها، می‌تواند این اطمینان را به وجود آورد که نمرات به اندازه کافی پایا هسـتند، و اگر فرایند سـنجش برای یک گروه داوطلب دوباره تکرار شـود نتایج ردی یا قبولی افراد تغییر زیـادی نمی‌کنـد (National Research Council, 2002؛ APA, AERA, NCME; 2014, p.101).

اگر چه «نتایج عددی و آماری» را دقیقاً نمی‌توان به آزمون‌های دیگر رشـــته‌های مختلف مثلاً علوم انسانی، علوم پایه دامپزشکی و... تعمیم داد. امّا نتایج و نتیجه‌گیری‌های کلی همهٔ رشته‌ها شبیه به هم هستند. به عبارتی، وضعیت آزمون‌های استعداد و زبان در رشته‌های علوم انسانی، علوم پایه، دامپزشـکی بهتر نیسـت. نتایج اولیهٔ تحلیل بر روی چند رشـته از جمله مهندسـی برق الکترونیک، هوش مصـــنوعی، عمران، فناوری اطلاعات و بســیاری دیگر از رشـــته‌ها نشـــان می‌دهد آن‌ها نیز وضعیت مشابه‌ای دارند؛ اگر که بدتر نباشند! در رشته مهندسی برق قدرت و الکترونیک از بهترین رشـــته‌ها بوده‌اند. تحلیل‌های اولیه داده‌های آن‌ها نشـــان داد اســـتادان آن‌ها آزمون مناســـب‌تری تهیه می‌کنند، و در جلسات مصاحبه دستاوردهای داوطلبان را با دقت بیشتر و واقع‌بینانه‌تر نمره‌گذاری می‌کنند: نمرات دروس اختصاصی، دروس عمومی و سنجه‌های مصاحبه رشته‌های مهندسی برق و الکترونیک نسبت به دیگر رشته‌ها ضرایب همبستگی کانونی بالاتری با یکدیگر دارند، و نمرات روی یک پیوستار خطی کاهش یا افزایش پیدا می‌کنند (به پیوست ۱ نگاه کنید). اگر برنامهٔ سنجش و پذیرش آن‌ها به همین شـکل ادامه پیدا کند این «نتایج عددی» مربوط به ضـرایب تعمیم‌پذیری و اعتمادپذیری آزمون‌هایی که آن‌ها تهیه کرده‌اند را می‌توان به سال‌های گذشته و آینده هم تعمیم داد. در نهایت، گزارش نهایی این پژوهش برای کمک به تصـــمیم‌گیران و دســـت‌اندرکاران برنامه‌های سنجش و پذیرش دکتری به ویژه کمیته‌های پذیرش دانشکده‌ها بوده است، و هدف غایی آن کمک به دانشـگاه‌ها در اسـتقلال مجدد برای سـنجش و پذیرش دانشـجوی دکتری، متناسـب با نیازهای خودشان و رفع نیازهای جامعه در حال توسعه ایران بوده است.

---

1 . developing effective task sets

منابع

الف. فارسی

آلن، مری جی و ین، وندی ام (۱۹۷۹). *مقدمه‌ای بر نظریه‌های اندازه‌گیری (روان‌سنجی)*. ترجمهٔ دلاور، علی (۱۳۷٤) انتشارات سازمان مطالعه و تدوین کتب علوم انسانی دانش‌گاه‌ها (سمت). چاپ اول.

ذوالفقارنسب، سلیمان، خدایی، ابراهیم و یادگارزاده، غلامرضا (۱۳۹۱). وزن‌دهی بهینه به سؤال‌ها و خرده آزمون‌های ورودی برای ساخت نمرهٔ کل ترکیبی. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*. ۳ (٤)، ۱۰٤-۷۹.

سازمان سنجش آموزش کشور (۱۳۹۵). *شیوه‌نامهٔ اجرایی آزمون ورودی دورهٔ دکتری (Ph.D.)*. سازمان سنجش آموزش کشور:
http://www.sanjesh.org/FullStory.aspx?gid=13&id=2316

وزارت بهداشت، درمان، و آموزش پزشکی (۱۳۹۷). *آیین‌نامهٔ آموزشی دورهٔ دکتری تخصصی (Ph.D.)*. مصوب شصت و نهمین جلسهٔ شورای عالی برنامه‌ریزی علوم پزشکی:
http://satim.tums.ac.ir/app/webroot/upload/files/PhD(2).pdf

یونسی، جلیل (۱۳۹٤). کاوش نگرش داوطلبان ورود به دورهٔ دکتری و اعضای هیئت علمی نسبت به شیوه‌های متفاوت گزینش متقاضیان ورود به دوره‌های دکتری. *فصلنامه اندازه‌گیری تربیتی*، ٦ (۲۲)، ۲۲۷ تا ۲٦۰.

ب. انگلیسی

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Attiyeh, G. M. (1999). Determinants of persistence of graduate students in Ph. D. programs. *ETS Research Report Series*, Volume 1999, Issue 1, i-43.

American Educational Research Association. American Psychological Association & National Council for Measurement in Education [AERA, APA & NCME] (2014). *The Standards for Educational and Psychological Testing.* Washington.

Brennan, R. L. (2004). Some perspectives on inconsistencies among measurement models. Center for Advanced Studies in Measurement and Assessment (CASMA *Research Report, Number* 8. Iowa City, IO: University of Iowa. (Available on https://education.uiowa.edu/sites/education.uiowa.edu/files)

Brennan, R. L. (2009). Notes about nominal weights in multivariate generalizability theory. CASMA Technical Note, No. 4 Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, the University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Crighton, J. V. (2003). *Standardized Tests and Educational Policy*. Encyclopedia of Education.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 161–188. New York: John Wiley

Gyll, S., & Ragland, S. (2018). Improving the validity of objective assessment in higher education: Steps for building a best-in-class competency-based assessment program. *The Journal of Competency-Based Education*, 3(1), 1-8.

Johnson, S., & Johnson, R. (2010). *Component reliability in GCSE and GCE*. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578865/Component_reliability_in_GCSE_and_GCE.pdf)

Kane, M. (2010). *Errors of Measurement, Theory, and Public Policy*. William H. Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Service. (Available on https://www.ets.org/Media/Research/pdf/PICANG12.pdf)

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315(5815), 1080-1081.

Li, D., & Brennan, R. (2007). A multi-group generalizability analysis of a large-scale reading comprehension test. *In annual meeting of the National Council on Measurement in Education*. Chicago, IL.

Lin, C. K. (2014). Issues and challenges in current generalizability theory applications in rated measurement (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Magnoson, D. (1991). *Theoretical Basis of Psychological Tests*.-Trans. by Baraheni MN. Tehran: Institute University Publisher.

National Research Council. (2002). Performance assessments for adult education: Exploring the measurement issues: Report of a workshop. National Academies Press.

Nussbaum, A. (1984). Multivariate generalizability theory in educational measurement: An empirical study.-*Applied Psychological Measurement*,-8(2), 219-230.

Sebok, S. S., & MacMillan, P. D. (2014). Assessment of a Master of Education Counselling Application Selection Process Using Rasch Analysis and Generalizability Theory. *Canadian Journal of Counselling and Psychotherapy*, *48*(2). Retrieved from https://cjc-rcc.ucalgary.ca/article/view/60970

Shavelson, R. J., & Webb, N. M. (2006). *Generalizability Theory*. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (p. 309–322). Lawrence Erlbaum Associates Publishers.

Smits, N., van der Ark, L. A., & Conijn, J. M. (2018). Measurement versus prediction in the construction of patient-reported outcome questionnaires: can we have our cake and eat it? *Quality of Life Research*, 27(7), 1673-1682.

Swan, M., & Burkhardt, H. (2012). A designer speaks: Designing assessment of performance in mathematics. Educational Designer: *Journal of the International Society for Design and Development in Education*, 2(5), 1-41.

Yu, M. (1991). The assessment of partial knowledge. *Journal of National Chengchi University*, 63, 401-428.

Webb, N. M., & Shavelson, R. J. (2005). *Generalizability theory: overview*. Encyclopedia of statistics in behavioral science.

Webb, N. L. (1997). Determining Alignment of Expectations and Assessments in Mathematics and Science Education. *Nise Brief*, 1(2), n2.

Zwick, R. (2007). College admission testing. *National Association for College Admission Counseling*, 1-44. (https://offices.depaul.edu/enrollment-management/test-optional/Documents/ZwickStandardizedTesting.pdf)

Studies in Learning & Instruction
Vol.12, No.1, 2020, Ser 78/1

English Abstract

Reliability of Electrical Power Engineering PhD Admission Criteria's in Iran: Based on Generalizability Theory

Soleyman Zolfagharnasab[1] **Ali Delavar[2]**
Norali Farokhi[3] **Ehsan Jamali[4]**

The assessment process for the admission of qualified candidates applying for doctoral programs in Iran has always been a challenging issue, leading to debates among academic communities about the psychometric properties of the admission criteria. Using the generalizability theory as the framework, this study investigated the reliability-like coefficients of the secondary data measures and the criteria related to the Ph.D. admission program at 37 departments admitting Electrical Power Engineering candidates at the Ph.D. level in different state universities in 2018. Data were analyzed using mGENOVA software with one-facet $p^{\bullet} \times i^{\circ}$ design. The results suggested that due to the high difficulty of items, omitted responses, and using scoring formula, four Subject Test Batteries (STB) and two General Test Batteries (GTB) did not have adequate generalizability and dependability coefficients. Two interview scores, which included different measures, showed better coefficients. However, little predictive value was found for the candidates' undergraduate grade-point average (UGPA) and graduate grade-point average (GGPA) due to their restricted range. Indeed, broadly defining the underlying construct and including multiple measures in the assessment design for doctoral programs leads to more error components in the results. Therefore, it is not possible to determine an exclusive reliability index for the predictive measures involved in the test. However, high-stakes educational decisions for the classification of applicants may be accomplished with fewer false predictions by adjusting test difficulty, using partial scoring schemes without correction for guessing, specifying the underlying construct more thoroughly, and increasing the levels of the facets.

Keywords: generalizability theory, reliability, PhD admission, cut score, false positive error, and false negative error

---

1. PhD candidate of Assessment & Measurement, Allameh Tabataba'i University, Iran. Email: salarnik2001@yahoo.com
2. Emeritus Professor Assessment & Measurement Department Psychology & Education Faculty Allameh Tabataba'i University, Iran.,Email: delavarali@yahoo.com (corresponding author)
3. Associate Professor Assessment & Measurement Department Psychology & Education Faculty Allameh Tabataba'i University, Iran. Email: farrokhinoorali@yahoo.com
4. Assistant Professor National Organization of Educational Testing (NOET), Iran. Email: ehsanjamali@gmail.com