University of Tehran press

JOURNAL OF FOREIGN LANGUAGE RESEARCH
p-ISSN:2588-4123 e-ISSN:2588-7521
https://jflr.ut.ac.ir **Email**:jflr@ut.ac.ir

An Argument-based Validation of the Vocabulary Subtest of MA Entrance Exam for English Majors

Mahmood Safari✉* **0000-0002-4797-5038**
Department of Foreign Languages, Humanities and Arts Faculty, Hazrat-e Masoumeh University, Qom, Qom, Iran.
Email: m.safari@hmu.ac.ir

ABSTRACT

Despite ample research on university entrance examinations, the vocabulary section of the M.A. entrance exam for English-related majors in Iran, which has long been criticized by both candidates and lecturers, has not been examined comprehensively and exclusively. The present study aimed to evaluate the vocabulary section of the exam in the past five years through an argument-based validation. The participants included 194 English-major undergraduate students, 24 M.A. students, 16 university professors, and six native speakers of English, who responded to the vocabulary section, a vocabulary size test as a criterion measure, and a questionnaire. The lexical items in the vocabulary sections were analyzed against general and specialized corpora as well as major word lists. To examine item characteristics, test validity, and reliability, the researcher employed item analysis procedures, criterion-related validation, and internal consistency (Cronbach's alpha), and the participants' questionnaire responses were analyzed qualitatively. The results indicated that 49.1% of the words were not of appropriate frequency in the BNC and COCA corpora, and 61.1% had no or scant frequency in specialized corpora for English majors. The validity and reliability indices of the test were found to be 0.32 and 0.48, respectively. Many items suffered from problems of item difficulty (70% of the items), item discrimination (38%), and choice distribution (58%), and participants generally deemed the test as unsuitable for admitting M.A. students. Consequently, the use of the test as a criterion for M.A. student admissions was found neither justified nor defensible. Implications, limitations, and suggestions for further research are discussed.

ARTICLE INFO
Article history:
Received: 04 May 2025
Received in revised form: 20 November 2025
Accepted: 24 November 2025
Available online: Winter 2025

Keywords:
Argument-based validation, vocabulary test, frequency, reliability, validity, item difficulty index, item discrimination index.

Safari, M. (2025). An Argument-based Validation of the Vocabulary Subtest of MA Entrance Exam for English Majors . *Journal of Foreign Language Research*, 15(4), 349-375. http//doi.org/ 10.22059/jflr.2025.394038.1207

Publisher: The University of Tehran Press.
DOI: http//doi.org/ 10.22059/jflr.2025.394038.1207

✉* Foreign Languages Department, Humanities and Arts Faculty, Hazrat-e Masoumeh University, Qom, Qom, Iran.

انتشارات دانشگاه تهران

پژوهش‌های زبان‌شناختی در زبان‌های خارجی
شاپای چاپی: ۴۱۲۳-۲۵۸۸ شاپای الکترونیکی:۷۵۲۱-۲۵۸۸
https://jflr.ut.ac.ir **Email**:jflr@ut.ac.ir

ارزیابی کمی و کیفی بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی: یک اعتبارسنجی مبتنی بر استدلال

محمود صفری*✉
گروه زبانهای خارجه، دانشکده علوم انسانی و هنر، دانشگاه حضرت معصومه (س)، قم، قم، ایران.
رایانامه: m.safari@hmu.ac.ir

اطلاعات مقاله
تاریخ ارسال: ۱۴۰۴/۰۲/۱۴
تاریخ بازنگری: ۱۴۰۴/۰۸/۲۹
تاریخ پذیرش: ۱۴۰۴/۰۹/۰۳
تاریخ انتشار: زمستان ۱۴۰۴
نوع مقاله: علمی پژوهشی

کلید واژگان:
اعتبارسنجی مبتنی بر استدلال، آزمون واژگان، بسامد، پایایی، روایی، شاخص دشواری، شاخص تمایز.

چکیده

علی‌رغم تحقیقات فراون بر روی آزمون‌های ورودی دانشگاه، بخش واژگان آزمون کارشناسی ارشد رشته‌های زبان انگلیسی، که مورد انتقاد بسیاری از داوطلبین و اساتید می‌باشد، به‌طور جامع و مجزا مورد ارزیابی قرار نگرفته است. تحقیق حاضر تلاش می‌کند بخش واژگان آزمون‌های پنج سال اخیر را از طریق اعتبارسنجی مبتنی بر استدلال مورد ارزیابی قرار دهد. شرکت‌کنندگان ۱۹۴ دانشجوی کارشناسی زبان انگلیسی، ۲۴ دانشجوی کارشناسی ارشد، ۱۶ استاد دانشگاه و شش بومی زبان انگلیسی بودند که به سؤالات بخش واژگان و آزمون اندازه واژگان، به‌عنوان آزمون معیار، و یک پرسش‌نامه پاسخ دادند. کلمات بخش واژگان در پیکره‌های عمومی و تخصصی و فهرستهای مهم واژگان بررسی شدند. برای ارزیابی ویژگی‌های سؤالات و روایی و پایایی آزمون از شیوه‌های تحلیل سؤالات، روایی ملاکی و پایایی درونی (آلفای کرانباخ) استفاده شد و پاسخهای شرکت‌کنندگان به پرسش‌نامه مورد ارزیابی کیفی قرار گرفتند. بررسی‌ها نشان داد ۴۹,۱ درصد کلمات بسامد مناسبی در پیکره‌های BNC و COCA ندارند و ۶۱,۱ درصد کلمات در پیکره‌های تخصصی رشته‌های زبان انگلیسی وجود ندارند یا بسامد ناچیزی دارند. روایی و پایایی آزمونها به ترتیب ۰,۳۲ و ۰,۴۸ به دست آمد. بسیاری از سؤالات مشکل شاخص دشواری (۷۰ درصد سؤالات)، شاخص تمایز (۳۸ درصد سؤالات) و توزیع گزینه (۵۸ درصد سؤالات) داشتند و شرکت‌کنندگان عموماً این آزمون را برای پذیرش دانشجویان کارشناسی ارشد مناسب ندانستند. درنتیجه، استفاده از نتایج این آزمون برای پذیرش دانشجویان کارشناسی ارشد موجه و قابل دفاع تشخیص داده نشد. پیامدها، محدودیتها و پیشنهادات برای

صفری, محمود . (۱۴۰۴). ارزیابی کمی و کیفی بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی: یک اعتبارسنجی مبتنی بر استدلال .*پژوهشهای زبانشناختی در زبانهای خارجی*،۱۵ (۴) ،۳۲۵-۳۴۷.
DOI: http//doi.org/ 10.22059/jflr.2025.394038.1207

Publisher: The University of Tehran Press.
DOI: http//doi.org/ 10.22059/jflr.2025.394038.1207

*✉ گروه زبانهای خارجه، دانشکده علوم انسانی و هنر، دانشگاه حضرت معصومه (س)، قم، قم، ایران.

۱. مقدمه

ارزیابی دانش واژگانی افراد، به‌عنوان بخشی از توانایی زبانی آنان، همواره یکی از اهداف آزمون‌های مهم از قبیل آزمون‌های پیشرفت تحصیلی و آزمون‌های ورودی دانشگاه بوده است. اولین مسئله در طراحی آزمون واژگان انتخاب کلماتی است که مورد سنجش قرار خواهند گرفت. هیوز و هیوز[۱] (۲۰۲۰) انتخاب کلمات برای ارزیابی را چالش ویژه آزمون‌های واژگان بر می‌شمارند. به‌طور سنتی انتخاب واژگان در آزمون‌های زبان دوم عمدتاً بر اساس نظر طراح آزمون در مورد درجه سختی کلمات و سطح توانایی آزمون‌دهندگان بوده است. ولی محققین دریافته‌اند که نظر طراح آزمون در مورد درجه سختی آزمون، به‌ویژه آزمون واژگان، پیش‌بینی مناسبی نمی‌باشد (سیدرنکو[۲]، ۲۰۱۱) . بسیاری از پژوهشگران کلماتی که مکرراً در زبان گفتار و نوشتار به کار می‌روند را مهم‌ترین کلمات هر زبانی می‌دانند و لذا ویژگی اصلی برای انتخاب واژگان آزمون را بسامد کلمات در نظر می‌گیرند (اشمیت و دیگران[۳]، ۲۰۱۹؛ چویی و مون[۴]، ۲۰۱۹؛ میلتون[۵]، ۲۰۰۹). در خصوص آزمون‌های ورودی دانشگاه، یکی از اهداف اصلی طبق نظر متخصصین آزمون‌سازی، ارزیابی توانایی دانشجویان در خواندن و نگارش متون تخصصی رشته‌های خود و امکان ورود و ادامه تحصیل در یک دوره دانشگاهی می‌باشد (بکمن و پالمر[۶]، ۱۹۹۶؛ فولچر و دیویدسون[۷]، ۲۰۰۷). لذا، بخش واژگان این آزمونها نیز باید کلماتی را بسنجند که دانشجویان در خواندن و نگارش متون دانشگاهی خود به آنها نیاز خواهند داشت.

انتخاب واژگان در آزمون‌های مهم مانند آزمون‌های ورودی دانشگاه اهمیت دوچندانی پیدا می‌کند. نتایج برخی از آزمونها تأثیر بسزایی بر زندگی بسیاری از افراد جامعه دارد و می‌تواند آینده آنان را رقم بزند یا تأثیر شگرفی بر آن داشته باشد. این آزمونها را آزمون‌های سرنوشت‌ساز یا آزمون‌های پرخطر[۸] می‌نامند. بکمن و پالمر (۱۹۹۶) آزمون‌های ورود به دانشگاه، اعطای بورسیه و استخدام معلمین را به‌عنوان نمونه هایی از آزمون‌های سرنوشت‌ساز ذکر می‌کنند. محققین معتقدند که این آزمونها به‌دلیل تأثیری که بر روی زندگی افراد دارند باید با دقت و احتیاط بیشتری تولید و اعتبارسنجی شوند و ویژگی‌های یک آزمون خوب از قبیل روایی و پایایی برای این آزمونها باید در حد بسیار بالاتری در نظر گرفته شود (بکمن و پالمر، ۱۹۹۶؛ الدر و الوقلین[۹]، ۲۰۰۳).

آزمون‌های ورودی دانشگاه، به‌عنوان آزمون‌های سرنوشت‌ساز، توجه پژوهشگران فراوانی را در ایران و کشورهای دیگر به خود جلب کرده‌اند. برخی از پژوهشگران بخش زبان عمومی آزمون‌های ورودی دانشگاه در مقاطع کارشناسی، کارشناسی ارشد و دکترا را از جنبه‌های مختلف مانند محتوای آزمون، روایی ملاکی، روایی پیشگو، روایی سازه، تبعیض جنسیتی، تبعیض رشته تحصیلی، بسامد واژگان و جنبه‌های دیگر مورد بررسی و ارزیابی قرار داده‌اند (امیریان و دیگران[۱۰]، ۲۰۲۰؛ پیش قدم و دیگران[۱۱]، ۱۳۹۹؛ خودی و دیگران[۱۲]، ۲۰۲۱؛ خویی[۱۳]، ۱۹۹۸؛ رضوی پور[۱۴]، ۲۰۱۴؛ رفعت بخش و احمدی[۱۵]، ۲۰۲۲، ۲۰۲۴؛ کرمی[۱۶]، ۲۰۱۳؛ مرندی و دیگران[۱۷]، ۲۰۲۰؛ نوروزی و کرمی[۱۸]، ۲۰۲۴).

علی‌رغم تحقیقات فراوان بر روی جنبه‌های مختلف آزمون‌های ورودی دانشگاه، بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی، که مورد انتقاد بسیاری از دانشجویان و اساتید می‌باشد، مورد غفلت قرار گرفته و تحقیقات کافی در این زمینه وجود نداشته است. تحقیقات محدودی که در این زمینه وجود دارد مربوط به یک یا دو جنبه خاص مانند روایی

---

۱ Hughes & Hughes
۲ Sydorenko
۳ Schmitt et al.
۴ Choi & Moon
۵ Milton
۶ Bachman & Palmer
۷ Fulcher & Davidson
۸ high-stakes tests
۹ Elder & O'Loughlin
۱۰ Amirian et al.
۱۱ Pishghadam et al.
۱۲ Khodi et al.
۱۳ Khoii, 1998
۱۴ Razavipur, 2014
۱۵ Rafatbakhsh & Ahmadi
۱۶ Karami
۱۷ Marandi et al.
۱۸ Noroozi & Karami

پیشگو (قاسمی ورزنه[1]، ۲۰۰۵)، روایی ملاکی (شیخ الاسلامی[2]، ۱۹۹۹)، پایایی دو نیمه و تحلیل سؤالات (جمالی فر و دیگران[3]، ۲۰۱۴)، روایی سازه و پایایی (راوند و فیروزی[4]، ۲۰۱۶) و بسامد واژگان (رفعت بخش و احمدی، ۲۰۲۴) بوده و بخش واژگان آزمون را به‌طور مجزا و جامع مورد ارزیابی قرار نداده‌اند. علاوه‌بر این، تناقضهایی در یافته‌های تحقیقات مرتبط پیشین وجود داشته است. جمالی فر و دیگران و قاسمی ورزنه پایایی و روایی مناسبی را برای بخش واژگان آزمون کارشناسی ارشد رشته‌های زبان انگلیسی گزارش کردند، درحالی‌که راوند و فیروزی و شیخ الاسلامی روایی و پایایی بخش واژگان آزمون را مناسب نیافتند. بسامد واژگان آزمون مربوطه نیز به‌درستی مورد ارزیابی قرار نگرفته است. در تحقیق رفعت بخش و احمدی، بسامد واژگان آزمون گروه زبان انگلیسی در کنار واژگان آزمون کارشناسی ارشد رشته‌های دیگر در پیکره COCA مورد ارزیابی قرار گرفت؛ لذا اطلاعات کافی در مورد بسامد تمام واژگان آزمون رشته‌های زبان انگلیسی به دست نیامد و تنها تعداد واژگانی که در پیکره نبودند و یا بسامدی کمتر از ده داشتند گزارش شد. ازسوی‌دیگر، تحقیقات قبلی عمدتاً از شیوه‌های پیشین اعتبارسنجی مانند روایی سنجی محتوا–معیار[5] و روایی سازه[6] استفاده کرده‌اند و شیوه اعتبارسنجی مبتنی بر استدلال[7] برای ارزیابی بخش زبان عمومی کارشناسی ارشد زبان انگلیسی و بخش واژگان آن به کار نرفته است.

۲. پیشینه تحقیق

تحقیق حاضر سعی دارد تا بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی را به شیوه اعتبارسنجی مبتنی بر استدلال مورد ارزیابی قرار دهد. در این بخش موضوعات اعتبارسنجی، آزمون واژگان و تحقیقات عملی و مرتبط پیشین بررسی و ارائه می‌شوند.

۱،۲. اعتبارسنجی

روایی یا اعتبار[8] در دوره‌های مختلف معانی و تعاریف متفاوتی داشته است. در ابتدای قرن بیستم، روایی چنین تعریف می‌شد، اندازه‌ای که آزمون آنچه را که باید بسنجد می‌سنجد (کلی، ۱۹۲۷، ذکر شده در بقایی و امراهی[9]، ۲۰۱۱: ۱۰۵۲). روایی آزمون معمولاً با اندازه‌گیری رابطه بین نمرات آزمون و نمرات یک آزمون معیار به دست می‌آمد. بعدها در سال ۱۹۵۴، انجمن روانشناسی آمریکا چهار نوع روایی را مشخص کرد که شامل روایی محتوا، روایی پیشگو، روایی هم‌زمان، و روایی سازه بود. در حوزه آزمون‌سازی زبان مطابقت محتوای آزمون با مطالب و سرفصلهای آموزشی (روایی محتوا) و تطابق بین عملکرد زبان‌آموزان در آزمون طراحی شده و یک آزمون معیار (روایی ملاکی) کاربرد فراوانی داشتند. با وجود این، روایی، فهرستی از مقوله‌های مجزا بود تا یک مفهوم منسجم و یکپارچه. در اواخر قرن بیستم، مسیک[10] (۱۹۸۹: ۷۲) تِئوری یکپارچه روایی سازه را معرفی کرد. برای مسیک روایی یک مفهوم یکپارچه بود که در روایی سازه تبلور می‌یافت و شش بُعد داشت: بعد محتوا، بعد فرآیندی، بعد ساختاری، بعد قابلیت تعمیم، بعد بیرونی، و بعد پیامدها. بکمن[11] (۱۹۹۰، ۲۰۰۵) نیز با اتکا بر دیدگاه یکپارچه مسیک، چهارچوبی را پیشنهاد داد که در آن کارآمدی کلی آزمون زبان به‌عنوان تابعی از چند ویژگی به هم‌پیوسته (روایی، پایایی، اصالت، تعاملی بودن، تأثیر و کاربردپذیری) مفهوم‌سازی می‌شد.

آخرین مدل اعتبارسنجی[12]، اعتبارسنجی مبتنی بر استدلال می‌باشد که در دو دهه اخیر توسط کین[13] (۲۰۰۶، ۲۰۱۳)، چپل[14] (۲۰۲۰) و چپل و دیگران[15] (۲۰۰۸) ارائه و بسط داده شده است. این مدل رویکرد نظام‌مندی است که اعتبار آزمون را به‌عنوان زنجیره‌ای از استنتاج‌های موجه[16] در نظر می‌گیرد که نمرات آزمون را به تفسیرها و تصمیمات مبتنی بر نتایج آزمون مرتبط می‌سازد. برای انجام این مهم،

[1] Ghasemivarzaneh
[2] Sheikholeslam
[3] Jamalifar et al.
[4] Ravand & Firoozi
[5] content-criterion validity
[6] construct validity
[7] argument-based validation
[8] validity
[9] Baghaei & Amrahi
[10] Messick
[11] Bachman
[12] Validation
[13] Kane
[14] Chapelle
[15] Chapelle et al.
[16] justified inferences

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۵، شماره ۴، زمستان ۱۴۰۴، از صفحه ۳۴۹ تا ۳۷۵

محقق از انواع مختلف داده‌های کمی، کیفی، توصیفی، نظرسنجی و انواع مختلف ابزارهای جمع‌آوری و تحلیل داده شامل پرسش‌نامه، آزمون‌های معیار، تحلیل سؤالات و ابزارهای بررسی پیکره استفاده می‌کند و ارزیابی جامع و کاملی ازآزمون و استفاده از نتایج آن ارائه می‌کند. محقق در ابتدا چهارچوبی برای اعتبارسنجی ارائه می‌کند که در آن مراحل و ابعاد اعتبارسنجی و ابزارهای مربوطه را مشخص می‌نماید. سپس برای هر بعد شواهد و داده‌های لازم را جمع‌آوری و تجزیه‌وتحلیل می‌کند و در نهایت اعتبار استفاده از نمرات آزمون برای اهداف مد نظر مانند استخدام و پذیرش دانشگاه را تعیین و تبیین می‌کند.

طراحان این مدل (کین و چپل) تلاش کرده‌اند با حفظ کلیت و نقاط قوت مدل یکپارچه مسیک، چهارچوب ساده‌تر و عملی‌تری را برای اعتبارسنجی فراهم سازند (کین، ۲۰۱۳: ۴۵۰-۴۵۱). اعتبارسنجی مبتنی بر استدلال، اعتبار را به‌صورت زنجیره‌ای از دلایل و شواهد مفهوم‌سازی می‌کند، شواهدی که استفاده از نمرات آزمون را برای قضاوت و تصمیم‌گیری در مورد شرکت‌کنندگان مورد تأیید قرار می‌دهد یا به چالش می‌کشد (چپل و لی، ۲۰۲۱). در ارزشیابی زبان، اعتبارسنجی بر اساس زنجیره‌ای از هفت نوع استنتاج درهم تنیده می‌باشد (چپل، ۲۰۲۰؛ چپل و دیگران، ۲۰۰۸):

۱- تعریف دامنه[۱]: دامنه استفاده از زبان هدف را به محتوای آزمون یا مشاهده عملکرد فرد در آزمون پیوند می‌زند.

۲- ارزیابی[۲]: عملکرد شرکت‌کنندگان در آزمون را به‌دقت می‌سنجد و به شکل اطلاعات عددی ارائه می‌کند.

۳- تعمیم[۳]: نمرات مشاهده‌شده را به نمرات معیارهای دیگر پیوند می‌دهد.

۴- تبیین[۴]: نمرات مورد انتظار را به سازه زیربنایی مورد سنجش متصل می‌کند.

۵- برون افکنی[۵]: سازه عملیاتی‌شده را به وظایف واقعی و معیارهای دیگر در دامنه هدف پیوند می‌دهد.

۶- کاربست/تصمیم‌گیری[۶]: نمرات آزمون را به تصمیم‌ها یا کاربردهای معنادار مرتبط می‌سازد.

۷- پیامد[۷]: تصمیمات مبتنی بر نمره را به پیامدهای ناشی از استفاده از نمره پیوند می‌دهد.

۲,۲. آزمون واژگان

واژگان یکی از بنیادی‌ترین عناصر در یادگیری زبان دوم به شمار می‌آید و نقش آن در تمامی مهارت‌های زبانی، به‌ویژه خواندن درک مطلب و شنیدار، بارها مورد تأکید پژوهشگران قرار گرفته است (بگلار، ۲۰۱۰). در نتیجه، سنجش دانش واژگان بخش مهمی از ارزیابی توانایی زبانی محسوب می‌شود. طبق نظر بسیاری از پژوهشگران و طراحان آزمون، انتخاب واژگان به هدف آزمون بستگی دارد (یاناگیساوا و وب[۸]، ۲۰۲۰) و هدف بخش زبان عمومی آزمون‌های ورودی دانشگاه، عمدتاً ارزیابی توانایی افراد در خواندن و نگارش متون علمی در رشته تخصصی خود و ادامه تحصیل در دوره مربوطه می‌باشد (بکمن و پالمر، ۱۹۹۶؛ فولچر و دیویدسون، ۲۰۰۷). اشمیت و دیگران (۲۰۱۹) معتقدند ابتدا باید دامنه‌ای که سؤالات آزمون واژگان از آن انتخاب می‌شوند مشخص شود و بررسی شود که کلمات مورد سنجش با دامنهٔ مد نظر طراحان آزمون (مثلاً متون دانشگاهی) مرتبط و متناسب‌اند.

۳,۲. تحقیقات مرتبط پیشین

یکی از آزمون‌های سرنوشت‌ساز که توجه بسیاری از پژوهشگران ایرانی را به خود جلب کرده است بخش زبان عمومی آزمون‌های ورودی دانشگاه در ایران می‌باشد. محققین این آزمون را از جنبه‌های مختلف مانند محتوا (جعفرپور، ۱۳۷۶؛ خودی و دیگران، ۲۰۲۱)، روایی (خویی، ۱۹۹۸؛ شیخ الاسلامی، ۱۹۹۹؛ قاسمی ورزنه، ۲۰۰۵)، روایی سازه (راوند و فیروزی، ۲۰۱۶؛ مرندی و دیگران، ۲۰۲۰؛ نوروزی و کرمی، ۲۰۲۴)، روایی پیامدی (خسروانی و دیگران، ۱۴۰۱)، روایی فرآیندی و روایی پیشگو (رضوی پور، ۲۰۱۴)، پایایی (جمالی فر و دیگران، ۲۰۱۴؛ خویی، ۱۹۹۸؛ راوند و فیروزی، ۲۰۱۶)، تبعیض (کرمی، ۲۰۱۳)، عدالت (امیریان و دیگران، ۲۰۲۰)، بسامد واژگان (رفعت بخش و

[۱] Domain Definition
[۲] Evaluation
[۳] Generalization
[۴] Explanation
[۵] Extrapolation
[۶] Utilization
[۷] Consequence Implication
[۸] Yanagisawa & Web

احمدی، ۲۰۲۲، ۲۰۲۴)، نیازسنجی زبان‌آموزان (پیش قدم و دیگران، ۱۳۹۹) و موارد دیگر مورد ارزیابی قرار داده‌اند. آزمون‌های ورودی دانشگاه در کشورهای دیگر نیز مورد سنجش قرار گرفته است (اُبیوما و سالاو[۱]، ۲۰۰۷؛ پاردو-بالستر[۲]، ۲۰۱۰؛ جنینگز و دیگران[۳]، ۲۰۲۴؛ خامبونروانگ[۴]، ۲۰۲۵؛ کومازاوا و دیگران، ۲۰۱۶؛ یوجی و ونشیا[۵]، ۲۰۰۷).

ولی علی‌رغم تحقیقات فراوان در این زمینه تعداد محدودی از تحقیقات، مربوط به آزمون زبان عمومی کارشناسی ارشد رشته‌های زبان انگلیسی بوده و تعداد محدودتری به‌طور مجزا به بخش واژگان این آزمون پرداخته‌اند (جعفرپور، ۱۳۷۶؛ جمالی فر و دیگران، ۲۰۱۴؛ راوند و فیروزی، ۲۰۱۶؛ رفعت بخش و احمدی، ۲۰۲۴؛ شیخ الاسلامی، ۱۹۹۹؛ قاسمی ورزنه، ۲۰۰۵). هیچ‌کدام از تحقیقات مذکور، بخش واژگان آزمون کارشناسی ارشد رشته‌های زبان انگلیسی را به شکل ویژه و جامع مورد ارزیابی قرار نداده‌اند. علاوه‌بر این، تناقض‌هایی هم در یافته‌های تحقیقات مرتبط پیشین وجود داشته است. جعفرپور پس از بررسی نقادانه آزمون کارشناسی ارشد زبان انگلیسی اظهار داشت که سؤالات آزمون از نظر اصول آزمون‌سازی مشکلات بسیاری دارد و روایی آزمون زیر سؤال است. شیخ الاسلامی بخش زبان عمومی آزمون سالهای ۱۳۷۶ و ۱۳۷۷ را با آزمون تافل[۶] مقایسه کرد و نتایج نشان داد که بخش‌های خواندن و واژگان در دو آزمون ضریب همبستگی پایینی دارند؛ این نشان می‌داد بخش واژگان و خواندن روایی مناسبی ندارند. راوند و فیروزی بخش زبان عمومی آزمون سال ۱۳۸۸ را بررسی کردند و دریافتند که آزمون از پایایی پایین (۰,۵۳، ۰,۵۴ و ۰,۴۵ به ترتیب برای بخش‌های خواندن، گرامر و واژگان) برخوردار بوده و مشکل چندبعدی بودن[۷] دارد. در مقابل، قاسمی ورزنه با مقایسه عملکرد دانشجویان در آزمون کارشناسی ارشد سالهای ۱۳۸۱ و ۱۳۸۲ با نمرات آنها در آزمون آیلتس[۸] دریافت که آزمون زبان عمومی و زیرمجموعه‌های آن روایی مناسبی دارند. جمالی فر و دیگران نیز دریافتند که آزمون کارشناسی ارشد زبان انگلیسی دانشگاه آزاد ضریب پایایی (دو نیمه) قابل قبولی دارد. رفعت بخش و احمدی بسامد گزینه‌های بخش واژگان آزمون‌های کارشناسی ارشد شامل آزمون گروه زبان انگلیسی و دیگر رشته‌های کارشناسی ارشد را در پیکره عظیم COCA ارزیابی کردند و نشان دادند که کلمات آزمون‌شده بسامد مورد انتظار را در زیرپیکره دانشگاهی[۹] ندارند، اگرچه کلمات بیشترین بسامد را عموماً در زیرپیکره مذکور داشتند. البته بررسی کلمات آزمون گروه زبان انگلیسی درکنار آزمون رشته‌های دیگر باعث شد اطلاعات بیشتری در مورد بسامد کلمات آزمون گروه زبان انگلیسی به دست نیاید و فقط تعداد کلماتی که در پیکره موجود نبودند (۳ کلمه) و یا بسامدی کمتر از ده داشتند (۱۳ کلمه) مشخص شود. بررسی کلمات آزمون گروه زبان انگلیسی به‌طور مجزا و جدا از رشته‌های دیگر می‌توانست اطلاعات بیشتری درباره بسامد تمام کلمات آزمون و سطح سختی آنها ارائه کند.

۳. تحقیق حاضر

آزمون کارشناسی ارشد رشته‌های زبان انگلیسی یکی از آزمون‌های سرنوشت‌ساز در ایران می‌باشد که برای پذیرش دانشجویان کارشناسی ارشد رشته‌های ادبیات انگلیسی، مطالعات ترجمه و آموزش زبان انگلیسی به کار می‌رود. این آزمون دارای دو بخش آزمون زبان عمومی (شامل بخش‌های ساختار، واژگان و خواندن درک مطلب) و آزمون تخصصی (شامل آزمون درس‌هایی مانند روش تدریس، زبان‌شناسی، تئوری ترجمه، نقد ادبی و غیره) می‌باشد. بخش واژگان آزمون زبان عمومی همواره مورد انتقاد بسیاری از دانشجویان و اساتید زبان انگلیسی بوده و شرکت‌کنندگان را به مخاطره انداخته است، چنانکه برخی از شرکت‌کنندگان تصمیم می‌گیرند برای این بخش آزمون مطالعه نکنند و حتی در زمان آزمون سؤالات واژگان را پاسخ ندهند. مؤلف این مقاله به‌عنوان مدرس دوره‌های آمادگی آزمون کارشناسی ارشد رشته‌های زبان انگلیسی با مشکلات و چالش‌های بخش واژگان آزمون آشنا بوده و همواره به فکر کمک به ارتقای کیفیت آن بوده است. با توجه به کمبود تحقیقات بر

[۱] Obioma & Salau
[۲] Pardo-Ballester
[۳] Jennings et al.
[۴] Khamboonruang
[۵] Kumazawa et al.
[۶] TOEFL
[۷] Multidimentionality
[۸] IELTS
[۹] Academic Subcorpus

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۵، شماره ۴، زمستان ۱۴۰۴، از صفحه ۳۴۹ تا ۳۷۵

روی بخش واژگان این آزمون و تناقض در یافته‌های تحقیقات مرتبط پیشین و به منظور بررسی جامع و گسترده آزمون مذکور، محقق برآن شد تا با استفاده از شیوه اعتبارسنجی مبتنی بر استدلال بخش واژگان آزمون سالهای ۱۴۰۰ تا ۱۴۰۴ را مورد بررسی مجزا و جامع قرار داده و کیفیت و کارآیی آزمون را بررسی کند.

برای اعتبارسنجی بخش واژگان، محقق چهارچوب زیر (جدول ۱) را از مدلهای چپل و دیگران (۲۰۰۸) و چپل (۲۰۲۰) استخراج و استفاده کرد. این چهارچوب دارای پنج بُعد تعریف دامنه، ارزیابی ، تعمیم، برون افکنی (تعمیم به دنیای واقعی) و پیامد می‌باشد و اهداف و ابزارهای جمع‌آوری و تحلیل داده‌ها را مشخص می‌سازد. سؤالات تحقیق بر اساس این چهارچوب مطرح و فرضیات مربوطه منظور گردیده‌اند.

جدول ۱: چهارچوب اعتبارسنجی مبتنی بر استدلال برای ارزیابی بخش واژگان

بُعد ارزیابی	هدف	ابزارهای جمع آوری و تحلیل
تعریف دامنه	تطابق محتوای آزمون با دامنه هدف	بررسی واژگان درپیکره ها و فهرستهای واژگان بسامد–محور
ارزیابی	بررسی دقت و کیفیت سوالات آزمون	تحلیل سوالات آزمون[۱]
تعمیم	بررسی پایایی و قابلیت تعمیم نتایج آزمون	شیوه پایایی کرانباخ آلفا[۲]
برون فکنی	تطابق عملکرد در آزمون و معیارهای دیگر	روایی سنجی ملاکی[۳]
پیامد	بررسی استفاده از نتایج آزمون برای تصمیم گیری	پرسشنامه ارزیابی کیفی آزمون

3.1. سؤالات تحقیق

سؤالات تحقیق هرکدام به ترتیب به بعدهای اول تا پنجم چهارچوب اعتبارسنجی مرتبط هستند.

۱- چه درصدی از کلمات بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی بسامد مناسبی در پیکره‌های عمومی و پیکره‌های تخصصی رشته‌های زبان انگلیسی ندارند؟ چه درصدی ازکلمات در لیست‌های بسامد–محور (فهرست واژگان CEFR و Academic Word List) حضور و جایگاه مناسبی ندارند؟

۲- چه درصدی از سؤالات بخش واژگان ویژگی‌های یک سؤال خوب (شاخص دشواری، شاخص تمایز و توزیع گزینه) را ندارند؟

۳- آیا بخش واژگان آزمون‌های کارشناسی ارشد انگلیسی پایایی قابل قبولی دارد؟

۴- آیا بخش واژگان آزمون‌های کارشناسی ارشد انگلیسی روایی قابل قبولی دارد؟

۵- در نظر اساتید و دانشجویان زبان انگلیسی بخش واژگان آزمون برای سنجش دانش واژگانی داوطلبین و پذیرش دانشجویان کارشناسی ارشد مناسب است؟

۳,۲. فرضیات تحقیق

برای سؤالات ۳ و ۴ که سؤالات همبستگی وتجربی هستند فرضیات پوچ زیر مطرح شده است. سؤالات دیگر کیفی و بسامدی هستند و برای آنها معمولاً فرضیه مطرح نمی‌شود؛ برای این سؤالات انتظارات محقق ذکر شده است.

[۱] Item analysis

[۲] Cronbach's alpha

[۳] Criterion-related validity

۱- محقق با توجه به تجربه شخصی خود انتظار دارد کلمات بخش واژگان، بسامد و جایگاه مناسبی در پیکره‌ها و لیست‌های مذکور نداشته باشند.

۲- محقق انتظار دارد تعداد زیادی از سؤالات بخش واژگان شاخص دشواری، شاخص تمایز و توزیع گزینه مناسبی نداشته باشند.

۳- بخش واژگان آزمون‌های کارشناسی ارشد زبان انگلیسی پایایی قابل قبولی ندارد.

۴- بخش واژگان آزمون‌های کارشناسی ارشد زبان انگلیسی روایی قابل قبولی ندارد.

۵- محقق انتظار دارد شرکت‌کنندگان این آزمون را برای پذیرش دانشجویان کارشناسی ارشد مناسب ندانند.

۴. روش تحقیق

۱،۴. طرح تحقیق

تحقیق حاضر یک تحقیق ترکیبی کمی-کیفی می‌باشد که تلاش کرده است بخش واژگان زبان عمومی کارشناسی ارشد زبان انگلیسی را با استفاده از مدل اعتبارسنجی مبتنی بر استدلال مورد ارزیابی قرار دهد. چهارچوب تحقیق شامل پنج بُعد تعریف دامنه، ارزیابی ، تعمیم، برون افکنی (تعمیم به دنیای واقعی)، و پیامد می‌باشد. در بعد تعریف دامنه تلاش بر این بود تا محتوای آزمون (کلمات بخش واژگان) را با دامنه هدف که در این آزمون واژگان لازم برای خواندن و نگارش متون دانشگاهی رشته‌های زبان انگلیسی شامل مطالعات ترجمه، ادبیات انگلیسی و آموزش زبان انگلیسی می‌باشد مقایسه کند و مناسب بودن محتوا و کلمات آزموده‌شده را بررسی کند. برای انجام این کار، تمام کلمات (یعنی گزینه‌های سؤالات) بخش واژگان آزمون سالهای ۱۴۰۰ تا ۱۴۰۴ در فهرست واژگان BNC/COCA، فرهنگ لغتهای لانگمن و کمبریج، فهرست واژگان CEFR، فهرست واژگان دانشگاهی[۱] کاکس هد[۲] (۲۰۰۰) و همچنین در پیکره‌های تخصصی مترجمی، ادبیات انگلیسی و آموزش زبان انگلیسی، که برای انجام این تحقیق آماده شده بودند، مورد بررسی قرار گرفتند. در بعد ارزیابی، سؤالات واژگان آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ با استفاده از شیوه تحلیل سؤالات بررسی شد و شاخص دشواری، شاخص تمایز و توزیع گزینه‌های سؤالات مشخص گردید. برای بعد تعمیم، پایایی بخش واژگان آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ به شیوه کرانباخ آلفا محاسبه شد و برای بعد برون افکنی با استفاده از شیوه روایی ملاکی نمرات شرکت‌کنندگان در بخش واژگان آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ با نمرات آنها در آزمون اندازه واژگان[۳] نیشن و بگلار[۴] (۲۰۰۷) مقایسه شد. تعداد ۸۷ دانشجو در آزمون ۱۴۰۰ شرکت کردند و در آزمون‌های ۱۴۰۲ و ۱۴۰۳ به ترتیب ۷۴ و ۱۰۷ دانشجو حضور داشتند؛ تعدادی از دانشجویان در دو آزمون شرکت کردند. همه دانشجویان کارشناسی در آزمون اندازه واژگان شرکت کردند. و نهایتاً برای بعد پیامد، پاسخهای شرکت‌کنندگان در آزمون‌های ۱۴۰۰ و ۱۴۰۳ به سؤالات پرسش‌نامه محقق‌ساخته که در مورد مناسب بودن آزمون برای گزینش دانشجویان کارشناسی ارشد رشته‌های زبان انگلیسی بود به‌طور کیفی ارزیابی شدند.

۲،۴. شرکت‌کنندگان تحقیق

شرکت‌کنندگان در این پژوهش شامل ۱۹۴ دانشجوی کارشناسی زبان انگلیسی، ۲۴ دانشجوی کارشناسی ارشد، ۱۶ استاد زبان انگلیسی و شش بومی زبان انگلیسی بودند که به شیوه نمونه‌گیری در دسترس[۵] انتخاب شدند. دانشجویان کارشناسی شامل ۲۴ دانشجوی رشته زبان و ادبیات انگلیسی در دانشگاه قم، ۲۶ دانشجوی رشته آموزش زبان انگلیسی در دانشگاه مفید و ۳۷ دانشجوی رشته مترجمی زبان انگلیسی در دانشگاه حضرت معصومه (س) و دانشگاه مفید بودند. دامنه سنی دانشجویان کارشناسی ۲۱ الی ۲۵ سال بود و از هر دو جنسیت آقا و خانم حضور داشتند. این دانشجویان در سال چهارم تحصیل خود بودند و به‌زودی قرار بود در آزمون کارشناسی ارشد شرکت کنند. دانشجویان کارشناسی ارشد، دانشجویان رشته آموزش زبان انگلیسی در دانشگاه قم بودند؛ دامنه سنی آنها ۲۲ تا ۲۸ سال و از هر دو جنسیت بودند. این دانشجویان دانشجوی سال اولی بودند و به‌تازگی در آزمون کارشناسی ارشد شرکت کرده بودند. اساتید دانشگاه نیز بیش از ده سال سابقه تدریس در مقاطع

[۱] Academic Word List

[۲] Coxhead

[۳] Vocabulary Size Test

[۴] Nation & Beglar

[۵] convenience sampling

کارشناسی و کارشناسی ارشد در دانشگاه‌های قم را داشتند و با نیازهای دانشجویان کارشناسی ارشد آشنا بودند. بومی‌زبانان انگلیسی همه مدرس زبان انگلیسی در مدارس و آموزشگاه های کشور تایلند بودند؛ یکی از آنها استاد تمام دانشگاه بود و دیگران نیز تحصیلات دانشگاهی داشتند.

۳,۴. ابزار تحقیق

ابزار پژوهش در تحقیق حاضر شامل موارد زیر بود:

۱- بخش واژگان آزمون کارشناسی ارشد گرایش‌های زبان انگلیسی در سالهای ۱۴۰۰ تا ۱۴۰۴ (آزمون‌های ۱۴۰۰ و ۱۴۰۱ شامل بیست سؤال واژگان و آزمون‌های بعدی شامل ۱۵ سؤال)

۲- آزمون اندازه واژگان نیشن و بگلار (۲۰۰۷)، (بخش‌های شش و هفت)

۳- فهرست واژگان BNC/COCA، قابل دسترس در نرم‌افزار Range

۴- فهرست واژگان CEFR، قابل دسترس در سایت https://englishprofile.org

۵- پیکره عظیم زبان انگلیسی COCA، مشتمل بر بیش از یک میلیارد کلمه و متشکل از متون دانشگاهی، محاوره، روزمره، روزنامه و غیره

۶- پیکره‌های تخصصی زبان انگلیسی برای رشته‌های ادبیات، مترجمی و آموزش زبان انگلیسی، هرکدام با بزرگی ۳,۵ میلیون لغت، تهیه شده توسط محقق

۷- فرهنگ لغتهای معتبر زبان انگلیسی شامل:

الف- لانگمن (Longman Dictionary of Contemporary English, 5th edition)

ب- کمبریج (Cambridge Advanced Learners' Dictionary, 4th edition)

۸- نرم‌افزار تکست استت[۱] ۱,۵

۹- نرم‌افزار SPSS (ورژن ۲۷)

۱۰- پرسش‌نامه محقق‌ساخته در مورد کیفیت و کارآیی بخش واژگان آزمون کارشناسی ارشد رشته‌های زبان انگلیسی

آزمون اندازه واژگان برای سنجش دانش واژگانی توسط نیشن و بگلار در سال ۲۰۰۷ تهیه و تولید شد. این آزمون از ۱۴۰ سؤال چهارگزینه‌ای در قالب ۱۴ بخش ده سؤالی تشکیل شده است. هر بخش نماینده هزار خانواده لغت است و به ترتیب از اولین هزار خانواده لغت پرکاربرد تا چهاردهمین هزار خانواده لغت می‌باشند. طبق نظر نیشن و بگلار (۲۰۰۷) دو هزار خانواده اول شامل کلمات پرکاربرد است، سه تا نه هزار خانواده بعدی دربرگیرنده واژگانی با بسامد متوسط می‌باشند و خانواده‌های ده هزار به بالا کلمات کم کاربرد را در بر می‌گیرند. همچنین طبق نظر مؤلفین، یک دانشجوی کارشناسی غیرانگلیسی‌زبان برای تحصیل موفقیت‌آمیز در یک دانشگاه انگلیسی‌زبان به پنج الی شش هزار خانواده لغت نیازمند است و دانش واژگانی یک دانشجوی دکتری غیرانگلیسی‌زبان حدوداً نه هزار خانواده لغت است. لذا می‌توان نتیجه گرفت برای شروع موفقیت‌آمیز تحصیل در مقطع کارشناسی ارشد دامنه واژگان شش الی هفت هزار خانواده لغت مناسب می‌باشد. در تحقیق حاضر، بخشش ششم و هفتم آزمون اندازه واژگان، که باهم شامل ۲۰ سؤال چهارگزینه‌ای هستند، مورد استفاده قرار گرفت. به کار گرفتن کل آزمون که شامل ۱۴۰ سؤال می‌باشد به‌عنوان آزمون معیار نامناسب و دشوار به نظر می‌رسد و با توجه به نظرات نیشن و بگلار در مورد دانش واژگانی دانشجویان کارشناسی و دکترا می‌توان استدلال کرد که بخش‌های شش و هفت برای سنجش و ارزیابی دانش واژگانی داوطلبان آزمون کارشناسی ارشد مناسب می‌باشد. آزمون اندازه واژگان مورد ارزیابی‌های گسترده قرار گرفته و روایی و پایایی بالایی (ضریب روایی بالای ۰,۸۴ و ضریب پایایی بالای ۰,۹) برای آن گزارش شده است (بگلار[۲]،۲۰۱۰؛ کرمی[۳]، ۲۰۱۲؛ گیلستاد[۴]، ۲۰۱۲). بگلار نشان داد که بخش‌های مختلف آزمون اندازه واژگان به‌صورت مجزا دارای پایایی بالایی هستند؛ بخش‌های یک تا هشت ضریب پایایی ۰,۹۶ داشتند.

فهرست واژگان BNC/COCA، شامل لغات پرکاربرد زبان انگلیسی است که نیشن (۲۰۱۷) با تجزیه‌وتحلیل پیکره‌های BNC(British National Corpus) و COCA(Corpus of Contemporary American English) به دست آورد. این فهرست شامل ۳۴ هزار

[۱] TextSTAT 1.5

[۲] Beglar

[۳] Karami

[۴] Gyllstad

خانواده لغت است که در قالب ۳۴ لیست لغات پایه[۱]، هرکدام حاوی هزار خانواده لغت، طراحی شده است.

برای بررسی واژگان آزمون در متون تخصصی گرایش‌های زبان انگلیسی سه پیکره زبانی هرکدام با بزرگی بیش از ۳,۵ ملیون لغت برای رشته‌های مترجمی، ادبیات و آموزش زبان انگلیسی جمع‌آوری و تهیه شد. برای تولید پیکره‌ها ابتدا با متخصصین هر رشته مشورت و حیطه‌های هر رشته مشخص شد. سپس برای هر حیطه به‌اندازه کافی و مساوی متون گردآوری شد و بخش‌های اضافی متون از قبیل منابع، فهرست مطالب و ضمیمه‌ها از متون پاک شدند. در نهایت متون هر رشته در فایل‌های تکست[۲] در نرم‌افزار تکست استت بارگزاری شده و پیکره هر رشته توسط نرم‌افزار ساخته شد. اطلاعات مربوط به پیکره‌ها در جدول ۲ آمده است.

جدول ۲: اطلاعات پیکره‌های تخصصی گرایش‌های زبان انگلیسی

رشته	تعداد کلمات پیکره	تعداد نوع کلمه	تعداد کتابها
آموزش زبان انگلیسی	۳۵۳۶۰۳۰	۸۷۰۲۵	۱۸
ادبیات انگلیسی	۳۵۲۲۰۳۱	۱۱۷۲۴۹	۱۹
مترجمی انگلیسی	۳۵۲۵۶۰۷	۶۹۴۳۹	۲۳
پیکره کل رشته ها	۱۰۵۸۳۶۶۸	۲۷۳۷۱۳	۶۰

برای بررسی بسامد واژگان آزمون کارشناسی ارشد در پیکره‌های تخصصی و همچنین مشخص کردن اینکه هر کلمه در کدام‌یک از ۳۴ لیست لغات پایه BNC/COCA قرار دارد از نرم‌افزار تکست استت استفاده شد. متون با قالب تکست در حافظه نرم‌افزار بارگزاری شدند، سپس با استفاده از گزینه جستجو، واژگان در پیکره‌ها و لیست‌های لغات پایه جستجو شدند و از وجود و بسامد آنها اطلاعات مربوطه به دست آمد.

برای ارزیابی کیفی بخش واژگان آزمون محقق یک پرسش‌نامه شامل هشت سؤال تشریحی طراحی و تولید کرد. این پرسش‌نامه با کمک یکی از اساتید تهیه شد و سپس توسط سه تن از اساتید زبان انگلیسی با بیش از ده سال سابقه تدریس در مقاطع کارشناسی و کارشناسی ارشد مورد ارزیابی روایی محتوایی قرار گرفت و تأیید شد. در واقع پنج استاد دانشگاه با مدرک دکترای آموزش زبان انگلیسی و سابقه آموزشی کافی، در تولید و ارزیابی این پرسش‌نامه دخیل بوده‌اند. به علت اینکه سؤالات پرسش‌نامه سؤالات بازپاسخ[۳] بودند و قابلیت کدگذاری نداشتند، امکان بررسی پایایی پرسش‌نامه وجود نداشت. برخی از سؤالات نیز بسیار شخصی بودند (مثلاً، چه تعداد از گزینه‌ها را بلد نبودید؟ آیا برخی از سؤالات را تصادفی و شانسی زده بودید؟). سؤالات پرسش‌نامه در مورد مناسب بودن آزمون برای انتخاب دانشجویان کارشناسی ارشد می‌باشد (پیوست مقاله).

۴,۴. فرآیند گردآوری و تجزیه‌وتحلیل داده‌ها

داده‌های هدف در تحقیق حاضر شامل داده‌های کمی، کیفی و بسامد بودند، لذا از روشهای مختلفی برای گردآوری و تجزیه‌وتحلیل داده‌ها استفاده شد. برای سؤال اول تحقیق که مربوط به بُعد تعریف دامنه و تحلیل داده‌های بسامدی بود، تمام گزینه‌های سؤالات بخش واژگان آزمون‌های ۱۴۰۰ تا ۱۴۰۴ (آزمون‌های پنج سال) از نظر بسامد مورد بررسی قرار گرفتند. این کلمات در فهرست واژگان BNC/COCA، پیکره COCA (در صورت نبودن در فهرست واژگان BNC/COCA)، فرهنگ لغتهای لانگمن و کمبریج، فهرست واژگان CEFR و AWL وهمچنین پیکره‌های تخصصی مترجمی، آموزش زبان و ادبیات انگلیسی جستجو شدند و وجود و بسامد این کلمات در منابع فوق‌الذکر مورد کنکاش قرار گرفت. طبق نظر بسیاری از محققین برای خواندن متون دانشگاهی به کمتر از ده هزار خانواده لغت نیاز هست (اشمیت و اشمیت[۴]، ۲۰۱۴؛ نیشن و بگلار، ۲۰۰۷)، بنابراین کلماتی که در فهرست واژگان BNC/COCA

[۱] base word list

[۲] TEXT

[۳] Open-ended questions

[۴] Schmitt & Schmitt

وجود نداشتند یا در لیست لغتهای پایه (base word list) دهم تا سی چهارم قرار داشتند به‌عنوان لغات نامناسب در نظر گرفته شدند. در فهرست واژگان CEFR، طبق دستورالعمل مرکز ارزیابی زبان انگلیسی دانشگاه کمبریج (https://englishprofile.org/?menu=cefr-for-teachers-and-learners)، کلماتی که در سطوح C1 و C2 قرار دارند جزوه کلمات سطح پیشرفته در نظر گرفته می‌شوند، کلماتی که در سطوح A و B قرار دارند جزو کلمات مبتدی و متوسط هستند و کلماتی که در فهرست واژگان CEFR وجود ندارند در سطح بسیار دشوار قرار می‌گیرند. برای آزمون کارشناسی ارشد کلمات سطح C مناسب در نظر گرفته شد. تعداد کلمات در فهرست واژگان AWL نشان می‌دهد چه درصدی از کلمات آزموده‌شده جزء کلمات دانشگاهی می‌باشند.

برای بعد ارزیابی (سؤال دوم تحقیق)، ویژگی‌های تک تک سؤالات آزمون‌های ۱۴۰۰، ۱۴۰۲، و ۱۴۰۳، شامل شاخص دشواری (item facility)، شاخص تمایز (item discrimination) و توزیع گزینه‌ها (choice distribution) مورد سنجش قرار گرفتند. تحلیل سؤالات بر روی پاسخهای دانشجویان کارشناسی به سه آزمون انجام گرفت: آزمون ۱۴۰۰ (با ۸۷ دانشجوی شرکت‌کننده)، آزمون ۱۴۰۲ (با ۷۴ دانشجو) و آزمون ۱۴۰۳ (با ۱۰۷ دانشجو). با اتکا بر نظر متخصصین آزمون‌سازی و سنجش برای شاخص دشواری دامنه ضریب ۰,۳ تا ۰,۷(میلر و دیگران[۱]، ۲۰۱۳؛ هالادینا[۲]، ۲۰۰۴)، برای شاخص تمایز ضریب ۰,۳ (دولیس[۳]، ۲۰۱۶؛ لین و گرونلاند[۴]، ۲۰۰۰) و برای توزیع گزینه‌ها، انتخاب گزینه‌های نادرست توسط ۱۰ درصد الی ۳۰ درصد شرکت‌کنندگان (تارانت و ویر[۵]، ۲۰۱۲، رایت و استون[۶]، ۱۹۷۹) به‌عنوان معیار انتخاب سؤالات مناسب استفاده شدند. یعنی سؤالاتی که بیش از ۷۰ درصد از شرکت‌کنندگان جواب صحیح داده بودند به‌عنوان سؤال بسیار آسان و سؤالاتی که کمتر از ۳۰ درصد از شرکت‌کنندگان جواب صحیح داده بودند به‌عنوان سؤال بسیار سخت در نظر گرفته شدند. شاخص تمایز مناسب برای سؤالات نیز ضریب ۰,۳ در نظر گرفته شد و پاسخهای غلطی که توسط ۱۰ الی ۳۰ درصد از شرکت‌کنندگان انتخاب شده بودند به‌عنوان گزینه کارآمد و مابقی به‌عنوان گزینه‌های ناکارآمد ارزیابی شدند. برای شاخص تمایز معادله زیر به کار رفت:

$$ID = \frac{HC - LC}{HC + LC}$$

ID = شاخص تمایز سؤال

HC = تعداد دانشجویان قوی که سؤال را درست جواب داده‌اند

LC = تعداد دانشجویان ضعیفی که سؤال را درست جواب داده‌اند

برای ارزیابی قابلیت تعمیم آزمون، پایایی بخش واژگان آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ با شیوه کرانباخ آلفا در بسته نرم‌افزاری اس پی اس اس (ورژن ۲۷) محاسبه شد. همان‌طور که پیشتر گفته شد، در این آزمونها به ترتیب ۸۷، ۷۴ و ۱۰۷ دانشجوی کارشناسی شرکت کرده بودند. طبق نظر متخصصین آزمون‌سازی، کرانباخ آلفا یکی از بهترین و رایج‌ترین شیوه‌های سنجش پایایی آزمون می‌باشد (کوهن ودیگران[۷]، ۲۰۱۷؛ ویر[۸]، ۲۰۰۵). برای ارزیابی پایایی بخش‌های مختلف آزمون‌های زبان انگلیسی کمبریج عموماً از شیوه کرانباخ آلفا استفاده می‌شود (ویر، ۲۰۰۵: ۳۰). برای ارزیابی برون افکنی، از روایی ملاکی استفاده شد؛ نمرات دانشجویان در بخش واژگان آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ و نمرات آنها در آزمون اندازه واژگان به کمک نرم‌افزار اس پی اس اس، مقایسه شد و ضریب همبستگی دو آزمون استخراج شد تا روایی بخش واژگان محاسبه شود. متخصصین سنجش و آزمون‌سازی عموماً حداقل ضریب ۰,۷ را برای پایایی وروایی قابل قبول مطرح کرده‌اند (براون[۹]، ۲۰۰۵؛ بکمن و پالمر، ۲۰۱۰؛ کوهن و دیگران، ۲۰۱۷)، اگرچه برخی ضریب ۰,۸ را مطلوب‌تر دانسته‌اند (ویر، ۲۰۰۵). همچنین برای آزمون‌های سرنوشت‌ساز، انتظار می‌رود ضریب پایایی و روایی بالاتری لحاظ شود. در

[۱] Miller et al.
[۲] Haladyna
[۳] DeVellis
[۴] Linn & Gronlund
[۵] Tarrant & Ware
[۶] Wright & Stone
[۷] Cohen et al.
[۸] Weir
[۹] Brown

پژوهش حاضر، ضریب ۰,۷ به‌عنوان معیار سنجش برای هر دو شاخصِ پایایی و روایی آزمون در نظر گرفته شد و در نهایت، برای بُعد پیامد، پاسخهای شرکت‌کنندگان به سؤالات پرسش‌نامه مورد ارزیابی کیفی قرار گرفتند.

۵. نتایج تحقیق

داده‌های جمع‌آوری شده برای هرکدام از سؤالات تحقیق به شیوه‌ای که در بخش پیشین توضیح داده شد مورد تجزیه‌وتحلیل قرار گرفتند و پاسخ سؤالات تحقیق به دست آمد. در این بخش، نتایج تحقیق برای هر سؤال به تفکیک ارائه می‌شوند.

۱,۵. بُعد تعریف دامنه (بسامد کلمات در پیکره)

برای بررسی بعد تعریف دامنه و پاسخ به سؤال اول تحقیق، تمام کلمات در گزینه‌های بخش واژگان آزمون‌های پنج سال استخراج شد و حضور و بسامد آنها در پیکره‌ها و لیست‌های واژگان مذکور بررسی شد. از مجموع ۳۴۰ کلمه در گزینه‌های این پنج آزمون، ۹,۴ درصد کلمات (یعنی ۳۲ کلمه) در لیست BNC/COCA وجود نداشتند و ۳۹,۷ درصد کلمات (یعنی ۱۳۵ کلمه) در لیست‌های لغات پایه دهم تا سی و چهارم قرار داشتند. درواقع قریب به نیمی (۴۹,۱ درصد) از کلمات، طبق معیار ارائه‌شده در بخش ۴,۴ (نبودن در لیست BNC/COCA و یا حضور در لیست‌های لغات پایه دهم تا سی چهارم) برای ارزیابی دانش واژگانی داوطلبان آزمون کارشناسی ارشد نامناسب شناخته شدند. چهار مورد از کلمات مذکور در پیکره عظیم یک میلیارد لغتی COCA یافت نشدند (*delitiologist, in actus me invito factus, morior invictus, omnifarious*). تعداد ۳۸ کلمه (یعنی ۱۱,۱ درصد کلمات) نیز در فرهنگ لغتهای لانگمن و کمبریج موجود نبودند. کلماتی که در فرهنگ لغتهای لانگمن و کمبریج و فهرست واژگان BNC/COCA وجود نداشتند در جدول ۳ ارائه شده‌اند.

در خصوص حضور و بسامد کلمات در پیکره‌های تخصصی، ۴۰,۲۹ درصد کلمات در پیکره آموزش زبان انگلیسی، ۳۸,۲۳ درصد در پیکره مترجمی و ۱۸,۸۲ درصد در پیکره ادبیات انگلیسی وجود نداشتند. بیش از ۲۸,۸ درصد کلمات (یعنی ۹۸ کلمه) نیز در پیکره تخصصی کل (شامل هر سه زیرپیکره تخصصی و حاوی بیش از ۱۰,۵ میلیون لغت) بسامدی کمتر از پنج داشتند، یعنی یک تا چهار بار به کار رفته بودند. به‌طور متوسط بیش از ۶۰ درصد کلمات یا در پیکره‌های تخصصی وجود نداشتند و یا بسامدی کمتر از پنج داشتند. همچنین، تنها ۳۱ کلمه (یعنی ۹,۱ درصد کلمات) در سطح C CEFR قرار داشتند و تعداد کلماتی که در فهرست واژگان AWL حضور داشتند فقط ۹ کلمه بود (یعنی تنها ۲,۶ درصد کلمات). نتایج تحقیق برای آزمون هر سال به‌طور مجزا و مفصل در ادامه ارائه شده است. اطلاعات این بخش در جدول ۴ خلاصه شده است.

۱,۱,۵. *آزمون کارشناسی ارشد ۱۴۰۰*

در آزمون سال ۱۴۰۰، که شامل بیست سؤال واژگان و در نتیجه ۸۰ گزینه بود، یازده کلمه در فهرست واژگان BNC/COCA وجود نداشت و چهل واژه هم در لیست‌های دهم به بالا قرار داشتند. درنتیجه ۵۱ کلمه، یعنی بیش از ۶۰ درصد واژگانی که مورد سنجش قرار گرفتند، طبق معیار ارائه‌شده در بخش ۴,۴، برای این آزمون نامناسب ارزیابی شدند. از یازده کلمه‌ای که در لیست BNC/COCA وجود نداشتند بسامد پنج کلمه در پیکره COCA کمتر از پنج بود (*bemire, caliginous, hypermodernist, oppugnant, silviculturist*) و یک کلمه نیز در پیکره وجود نداشت (*omnifarious*). به‌علاوه، هجده مورد از کلمات نیز در فرهنگ لغتهای لانگمن وکمبریج یافت نشدند. در خصوص پیکره‌های تخصصی، حدود نیمی از واژگان در پیکره‌های آموزش زبان (۳۹ کلمه) و مترجمی (۳۷ کلمه) به کار نرفته بودند و ۱۹ کلمه نیز در پیکره ادبیات انگلیسی وجود نداشتند. شانزده کلمه هم در پیکره تخصصی کل بسامدی کمتر از پنج داشتند. فقط دو کلمه در سطح C لیست CEFR (*cramp, unquaslified*) و یک کلمه در لیست AWL (*prohibitive*) حضور داشتند.

۲,۱,۵. *آزمون کارشناسی ارشد ۱۴۰۱*

در آزمون سال ۱۴۰۱ با ۲۰ سؤال واژگان، تعداد نه کلمه در لیست BNC/COCA موجود نبود و ۲۸ کلمه در لیست‌های دهم تا سی چهارم قرار داشتند، یعنی قریب به نیمی از کلمات (تعداد ۳۷ کلمه) برای این آزمون نامناسب بودند. پنج کلمه (*atrabilious, brummagem, homologate, hortative, temerarious*) کمتر از ۵ بار در پیکره COCA به کار رفته بودند و دو کلمه هم بسامدی بین ۱۰ تا ۲۰ داشتند. یازده کلمه هم در فرهنگ لغتهای لانگمن وکمبریج موجود نبود. همچنین، ۳۵ کلمه در پیکره آموزش زبان انگلیسی، ۳۲ کلمه در پیکره مترجمی و ۱۵ کلمه در

پیکره ادبیات انگلیسی به کار نرفته بودند و ۲۶ کلمه نیز در پیکره تخصصی کل بسامدی کمتر از ۵ داشتند. تنها پنج کلمه در سطح C لیست CEFR (*claim, irony, rank, refrain, tender*) قرار داشتند و هیچ‌کدام از کلمات در لیست AWL نبودند.

۳،۱،۵. *آزمون کارشناسی ارشد ۱۴۰۲*

در آزمون سال ۱۴۰۲، که بخش واژگان ۱۵ سؤال و در نتیجه ۶۰ گزینه داشت، ۲۲ کلمه در لیست‌های دهم به بالای BNC/COCA قرار داشتند و ۴ کلمه در لیست نبود. در واقع بیش از۴۰ درصد واژگان برای آزمون کارشناسی ارشد زبان انگلیسی نامناسب بودند. یکی از این کلمات در پیکره عظیم COCA به کار نرفته بود (*delitiologist*) و سه کلمه هم در فرهنگ لغتها پیدا نشدند. بیست و چهار کلمه در پیکره آموزش زبان، ۱۷ کلمه در پیکره مترجمی و ۱۲ کلمه در پیکره ادبیات انگلیسی یافت نشدند و هجده کلمه کمتر از پنج بار در پیکره تخصصی کل به کار رفته بودند. هفت کلمه در سطح C لیست CEFR بودند (*indispensable, immune, spark, rank, stake, snap, trace*) و سه کلمه در لیست AWL (*unpredictable, range, trace*).

۴،۱،۵. *آزمون کارشناسی ارشد ۱۴۰۳*

در آزمون سال ۱۴۰۳، ۲۳ کلمه در لیست‌های دهم به بالای BNC/COCA قرار داشتند و دو کلمه در کل لیستهای BNC/COCA موجود نبودند. در واقع ۴۱،۶۶ درصد کلمات برای این آزمون نامناسب تشخیص داده شدند. همچنین چهار کلمه در فرهنگ لغتها یافت نشدند و یک عبارت هم در پیکره COCA وجود نداشت (*in actus me invito factus*). علاوه‌بر این، ۲۲ کلمه در پیکره مترجمی، ۱۸ کلمه در پیکره آموزش زبان و ۱۰ کلمه در پیکره ادبیات انگلیسی یافت نشدند و ۱۸ کلمه در پیکره تخصصی کل کمتر از پنج بار به کار رفته بودند. پنج کلمه در سطح C لیست CEFR (*plight, declaration, inclusive, collaborative, serve up*) قرار داشتند و دو کلمه در لیست AWL (*stress, hypothetical*).

۵،۱،۵. *آزمون کارشناسی ارشد ۱۴۰۴*

و در نهایت، در آزمون سال ۱۴۰۴، شش کلمه در فهرست واژگان BNC/COCA وجود نداشتند و ۲۲ کلمه در لیست‌های دهم تا سی و چهارم قرار داشتند، یعنی ۴۶،۶۶ درصد کلمات برای این آزمون بسیار سخت و نامناسب شناخته شدند. یکی از این کلمات در پیکره COCA یافت نشد (*morior invictus*) و دو کلمه هم در فرهنگ لغتهای لانگمن و کمبریج موجود نبودند. همچنین، ۲۱ کلمه در پیکره آموزش زبان انگلیسی، ۲۲ کلمه در پیکره مترجمی و ۸ کلمه در پیکره ادبیات انگلیسی یافت نشدند و ۲۰ کلمه کمتر از پنج بار در پیکره تخصصی کل به کار رفته بودند. دوازده کلمه در سطح C لیست CEFR قرار داشتند (*crumble, trace, drain, tread, undermine, plead, vicious, dole out,* explicitly, endow, linger, sensational) و سه کلمه در لیست AWL (*trace, correspond, explicitly*).

جدول ۳.واژگانی که در فرهنگ لغتها و فهرست BNC/COCA وجود نداشتند

منبع / آزمون	فرهنگ لغت لانگمن و کمبریج	*فهرست واژگان* BNC/COCA
آزمون ۱۴۰۰	*arriviste, bemire, benison, bloviate, caliginous, demarche, deracinated, desuetude, discursion, dubiety, duende, flummeries, hypermodernist, inculpate, numinous, occluded, omnifarious, oppugnant, silviculturist*	*anathematize, bemire, bloviate, caliginous, discursion, hypermodernist, inculpate, omnifarious, oppugnant, self-effacing, silviculturist*
آزمون ۱۴۰۱	*aggrandize, aleatory, atrabilious, brummagem, homologate, hortative, palpate, plebian, pluvial, temerarious, tumid*	*atrabilious, brummagem, chimerical, homologate, hortative, plebian, prepossession, temerarious, tumid*
آزمون ۱۴۰۲	*delitiologist, lepidopterist, naysayer*	*delitiologist, expatriated, rehashed, valedictorian*

آزمون ۱۴۰۳	*procrustean, undergrid, in actus me invito factus, in flagrante delico*	*in actus me invito factus, undergrid,*
آزمون ۱۴۰۴	*fortuity, morior invictus*	*bona fide, fortuity, morior infictus, non sequitur, semper fidelis, willfully*

جدول ۴.اطلاعات واژگان آزمونها در پیکره‌ها، فهرست‌ها و فرهنگ لغتها

تعداد واژگان ...	آزمون ۱۴۰۰	آزمون ۱۴۰۱	آزمون ۱۴۰۲	آزمون ۱۴۰۳	آزمون ۱۴۰۴	کل آزمون‌ها
در لیست‌های ۱۰ تا ۳۴ BNC/COCA	۴۰	۲۸	۲۲	۲۳	۲۲	۱۳۵
غایب در فهرست واژگان BNC/COCA	۱۱	۹	۴	۲	۶	۳۲
غایب در فرهنگ لغتها	۱۸	۱۱	۳	۴	۲	۳۸
غایب در پیکره آموزش	۳۹	۳۵	۲۴	۱۸	۲۱	۱۳۷
غایب در پیکره مترجمی	۳۷	۳۲	۱۷	۲۲	۲۲	۱۳۰
غایب در پیکره ادبیات	۱۹	۱۵	۱۲	۱۰	۸	۶۴
با بسامد کمتر از پنج در پیکره تخصصی کل	۱۶	۲۶	۱۸	۱۸	۲۰	۹۸
موجود در سطح C لیست CEFR	۲	۵	۷	۵	۱۲	۳۱
موجود در لیست AWL	۱	۰	۳	۲	۳	۹

۵,۲. بُعد ارزیابی (نتایج تحلیل سؤالات)

برای بررسی بعد ارزیابی و پاسخ به سؤال دوم تحقیق، ویژگی تک تک سؤالات در آزمون‌های ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ (که به ترتیب ۸۷، ۷۴ و ۱۰۷ شرکت‌کننده داشتند) بررسی شدند. از مجموع پنجاه سؤال در این سه آزمون، ۳۵ سؤال شاخص دشواری نامناسب، ۱۹ سؤال شاخص تمایز نامناسب و ۲۹ سؤال توزیع نامناسب گزینه داشتند (جدول ۵). درواقع ۷۰ درصد سؤالات در شاخص دشواری، ۳۸ درصد در شاخص تمایز و ۵۸ درصد در توزیع گزینه‌ها دچار مشکل بودند.

سی و سه سؤال ضریب دشواری بالای ۰,۷ داشتند و دو سؤال ضریب دشواری کمتر از ۰,۳، یعنی دو سؤال بسیار آسان و ۳۳ سؤال بیش از حد دشوار بودند. در خصوص شاخص تمایز، ۱۹ سؤال ضریبی پایین‌تر از ۰,۳ داشتند و به‌خوبی بین دانشجویان قوی و ضعیف تمییز قایل نمی‌شدند. البته، استفاده از شیوه‌های دیگر (مثلاً، $ID = IF\ high - IF\ low$) احتمالاً برای تعداد کمتری از سؤالات مشکل شاخص تمایز را نشان می‌داد، ولی به علت اینکه برخی از سؤالات را تعداد بسیار محدودی از شرکت‌کنندگان جواب داده بودند این شیوه‌ها مناسب به نظر نمی‌رسید. مثلاً در سؤال ششم آزمون ۱۴۰۲ با ۷۴ نفر شرکت‌کننده، پنج نفر از گروه دانشجویان قوی و سه نفر از گروه دانشجویان ضعیف جواب صحیح داده بودند و در سؤال هشتم و دهم آزمون ۱۴۰۰ با ۸۷ شرکت‌کننده هفت نفر از دانشجویان قوی و چهار نفر از دانشجویان ضعیف پاسخ صحیح داده بودند. استفاده از شیوه‌های مذکور شاخص تمایز این سؤالات را مناسب نشان خواهد داد، درصورتی‌که قدرت تمایز مناسبی ندارند (تعداد دانشجویان قوی و ضعیفی که این سؤالات را جواب نداده بودند تقریباً مساوی بود). در خصوص توزیع گزینه‌ها، ۴۲ گزینه (در ۲۹ سؤال) عملکرد مناسبی نداشتند؛ ۳۳ گزینه را کمتر از ۱۰ درصد از شرکت‌کنندگان انتخاب کرده بودند و ۹ گزینه غلط انتخاب بیش از ۳۰ درصد از شرکت‌کنندگان بود. نتایج تحقیق برای آزمون هر سال به‌طور مجزا و مفصل در ادامه ارائه شده است و نتایج به‌طور خلاصه در جدول ۵ آمده است.

۱,۲,۵. *آزمون کارشناسی ارشد ۱۴۰۰*

از بین سؤالات بخش واژگان این آزمون که شامل بیست سؤال بود، ۱۳ سؤال شاخص دشواری نامناسبی داشتند، یعنی درجه سختی آنها بیش از ۰,۷ بود و بیش از هفتاد درصد از

شرکت‌کنندگان به این سؤالات پاسخ نادرست داده بودند (سؤالات ۳، ۴، ۵، ۶، ۷، ۸، ۹، ۱۰، ۱۳، ۱۵، ۱۷، ۱۸، ۱۹). سطح دشواری برخی از سؤالات آنچنان بالا بود که بومی‌زبانان و اساتید دانشگاه نیز جواب آنها را نمی‌دانستند. سؤالهای ۱۱ و ۱۷ را فقط یک نفر از شش بومی‌زبان جواب صحیح داده بودند و سؤالات ۴، ۸، ۹، ۱۵ را فقط دو بومی‌زبان. برای اساتید دانشگاه نیز ۴ سؤال ضریب دشواری بسیار بالایی داشت. برای دانشجویان کارشناسی ارشد نیز سیزده سؤال ضریب دشواری بسیار بالایی داشتند. در خصوص شاخص تمایز، هشت سؤال (شامل سؤالهای ۱، ۲، ۵، ۸، ۹، ۱۰، ۱۱ و ۱۲) ضریب تمایز کمتر از ۰,۳ داشتند و به‌طور مناسبی بین داوطلبان قوی و ضعیف تمییز ایجاد نمی‌کردند. توزیع گزینه‌ها نیز در نه سؤال (شامل سؤالهای ۱، ۳، ۴، ۷، ۸، ۱۰، ۱۴، ۱۸، ۲۰) مناسب نبود. شش مورد از گزینه‌های غلط را کمتر از ده درصد از شرکت‌کنندگان انتخاب کرده بودند و شش گزینه غلط انتخاب بیش از سی درصد از شرکت‌کنندگان بود. برای مثال، در سؤال سوم، گزینه غلط سوم را فقط پنج درصد از شرکت‌کنندگان انتخاب کرده بودند، درحالی‌که گزینه غلط دوم انتخاب ۴۹ درصد از دنشجویان بود. هیچ‌کدام از سؤالات آزمون هر سه ویژگی (شاخص دشواری، شاخص تمایز و توزیع گزینه‌ها) را باهم نداشت و در یک تا سه شاخص دچار مشکل بود.

۲,۲,۵. آزمون کارشناسی ارشد ۱۴۰۲

در آزمون سال ۱۴۰۲، که ۱۵ سؤال واژگان داشت، تعداد نه سؤال شاخص دشواری بالای ۰,۷ داشتند (سؤالهای ۱۳،۱۰،۹،۷،۶،۵،۴،۲،۱). پنج مورد از این سؤالها برای اساتید دانشگاه نیز بسیار سخت بود و دو سؤال را هیچ‌کدام از اساتید جواب صحیح نداده بودند (سؤالهای ۱۰ و۱۳). دانشجویان کارشناسی ارشد و بومی‌زبانان انگلیسی در این آزمون شرکت نکرده بودند. در خصوص شاخص تمایز، شش سؤال ضریبی کمتر از ۰,۳ داشتند و به‌خوبی بین دانشجویان قوی و ضعیف تمییز قایل نمی‌شدند (سؤالهای ۳، ۶، ۸، ۹، ۱۱، ۱۴). در نه سؤال هم توزیع گزینه مناسب نبود. دوازده گزینه را کمتر از ۱۰ درصد دانشجویان به‌عنوان جواب صحیح انتخاب کرده بودند و یک گزینه غلط انتخاب تعداد بسیار زیادی از دانشجویان بود. برای مثال، در سؤال دوم، گزینه غلط اول را فقط پنج درصد دانشجویان انتخاب کرده بودند، درحالی‌که گزینه غلط آخر انتخاب ۴۲ درصد از دانشجویان بود. در بین سؤالات، فقط سؤال ۱۵ هر سه ویژگی یک سؤال خوب را داشت و بقیه در یک تا سه ویژگی دچار مشکل بودند.

۳,۲,۵. آزمون کارشناسی ارشد ۱۴۰۳

در آزمون ۱۴۰۳، تعداد سیزده سؤال از ۱۵ سؤال شاخص دشواری نامناسبی داشتند؛ دو سؤال بسیار راحت و ۱۱ سؤال بیش از حد سخت بود. سؤالهای ۲ و۳ ضریب دشواری کمتر از ۰,۳ داشتند، درحالی‌که ضریب دشواری سؤالهای ۵، ۶، ۷، ۹، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴، ۱۵، رقمی بالای ۰,۷ بود. چهار سؤال برای اساتید دانشگاه و هشت سؤال برای دانشجویان کارشناسی ارشد ضریب دشواری بسیار بالایی داشتند. پنج سؤال نیز ضریب تمایزی زیر ۰,۳ داشتند (سؤالهای ۱۴،۴،۳،۲،۱). یازده سؤال هم مشکل توزیع گزینه داشت. پانزده گزینه غلط را تعداد بسیار کمی از دانشجویان انتخاب کرده بودند و دو گزینه نادرست بیش از حد انتخاب شده بودند. در سؤال دوم، گزینه غلط اول را فقط یک نفر از ۱۰۷ دانشجو انتخاب کرده بود، درحالی‌که گزینه غلط سوم در سؤال هفت، انتخاب ۵۵ درصد از شرکت‌کنندگان بود. هیچ‌کدام از سؤالات تمام ویژگی‌های یک سؤال خوب را هم‌زمان نداشتند.

جدول ۵. نتایج تحلیل سؤالات بخش واژگان آزمون‌های کارشناسی ارشد

تعداد سوالات	آزمون ۱۴۰۰	آزمون ۱۴۰۲	آزمون ۱۴۰۳	کل آزمونها
با شاخص دشواری نامناسب	۱۳	۹	۱۳	۳۵
با شاخص تمایز نامناسب	۸	۶	۵	۱۹
با توزیع گزینه نامناسب	۹	۹	۱۱	۲۹
کل آزمون	۲۰	۱۵	۱۵	۵۰

۳،۵. بعد تعمیم (پایایی آزمون)

برای بعد تعمیم (سؤال سوم تحقیق)، متوسط ضریب پایایی سه آزمون ۰،۴۸ به دست آمد. ضریب پایایی بخش واژگان آزمون ۱۴۰۰، برابر ۰،۵۸ بود و این شاخص برای آزمون‌های ۱۴۰۲ و۱۴۰۳ به ترتیب رقمهای ۰،۴۱ و۰،۴۷ محاسبه شدند (جدول ۶)؛ هر سه ضریب پایایی از شاخص معیار (ضریب ۰،۷) بسیار پایین‌تر بودند. البته به علت اینکه آزمون زبان عمومی کارشناسی ارشد کوتاه شده است و پایایی بخش واژگان آزمون که ۱۵ الی ۲۰ سؤال دارد به‌طور مجزا محاسبه شده است، ضریب پایایی پایین‌تری را باید انتظار داشت. ولی ازسوی‌دیگر چون آزمون حاضر آزمون سرنوشت‌ساز و پرخطری محسوب می‌شود، پایایی قابل قبول نباید بسیار پایین در نظر گرفته شود. تصور می‌شود حداقل ضریب پایایی قابل قبول برای بخش واژگان این آزمون ۰،۶۰ باشد و ضریبهایی مانند ۰،۴۱ و ۰،۴۷ نامناسب می‌باشند. راوند و فیروزی (۲۰۱۶) نیز ضریب پایایی ۰،۵۳، ۰،۵۴ و ۰،۴۵ که برای بخش‌های خواندن، دستور و واژگان آزمون کارشناسی ارشد ۱۳۸۸ به دست آورده بودند را پایین و نامناسب در نظر گرفتند. در خصوص دانشجویان کارشناسی ارشد و اساتید در تحقیق حاضر، ضریب پایایی به‌دست‌آمده مشابه ضریب پایایی به‌دست‌آمده برای دانشجویان کارشناسی و یا حتی پایین‌تر از آن بود (ضریبهای ۰،۴۴ و ۰،۳۱)، اگرچه در برخی موارد ارقام بالاتری را نشان می‌داد (ضریب ۰،۶۲).

۴،۵. بعد برون افکنی (روایی ملاکی آزمون)

برای بعد برون افکنی (سؤال چهارم تحقیق)، میانگین ضریب روایی برای سه آزمون ۰،۳۲ به دست آمد، که از ضریب معیار (۰،۷) فاصله زیادی داشت و قابل قبول نبود. روایی آزمون ۱۴۰۰ برابر با رقم ۰،۴۱ و برای آزمون‌های ۱۴۰۲ و۱۴۰۳ به ترتیب رقمهای ۰،۳۴ و ۰،۲۲ محاسبه شدند. روایی به‌دست‌آمده برای دانشجویان کارشناسی ارشد و اساتید دانشگاه نسبتاً بهتر بودند، اگرچه با شاخص معیار فاصله زیادی داشتند (به ترتیب ۰،۵۶، ۰،۴۸ و ۰،۳۶). جدول ۶ ضریب روایی و پایایی بخش واژگان آزمون‌هایی که توسط دانشجویان کارشناسی پاسخ داده شده بود را نمایش می‌دهد.

جدول ۶. **پایایی و روایی بخش واژگان آزمون‌های کارشناسی ارشد**

شاخص	آزمون ۱۴۰۰	آزمون ۱۴۰۲	آزمون ۱۴۰۳	کل آزمونها
پایایی	۰،۵۸	۰،۴۱	۰،۴۷	۰،۴۸
روایی	۰،۴۱	۰،۳۴	۰،۲۲	۰،۳۲

۵،۵. بعد پیامد (تحلیل پاسخها به پرسش‌نامه)

برای بعد پیامد که مربوط به مناسب بودن استفاده از نتایج آزمون در تصمیم‌گیری‌ها و قضاوتها است یکی از ابزارهای ممکن نظرسنجی و پرسش‌نامه است. در این تحقیق، گروهی از دانشجویان شرکت‌کننده در آزمون ۱۴۰۰ (۲۹ نفر) و آزمون ۱۴۰۳ (۲۲ نفر) و شش نفر از اساتید به سؤالات پرسش‌نامه (پیوست مقاله) پاسخ دادند و بررسی نظرات آنان پاسخ سؤال آخر تحقیق را فراهم ساخت.

در خصوص سؤال اول، ۲۵ نفر در آزمون ۱۴۰۰ و ۲۰ نفر در آزمون ۱۴۰۳ معتقد بودند که این آزمون برای هدف مورد نظر مناسب نمی‌باشد. دلایل مطرح‌شده عمدتاً سختی بیش از حد واژگان و غیرکاربردی بودن آنها بود. در مورد سؤال دوم، اکثر دانشجویان اذعان داشتند که برخی از سؤالات را تصادفی و بدون دانش مورد نیاز پاسخ داده اند. در پاسخ به سؤال بعدی (چه تعداد از گزینه‌ها را بلد نبودید) شرکت‌کنندگان در آزمون ۱۴۰۰ معنی حدوداً ۶۵ کلمه (از ۸۰ گزینه) را نمی‌دانستند و شرکت‌کنندگان آزمون ۱۴۰۳ به‌طور متوسط ۳۵ کلمه از ۶۰ گزینه را بلد نبودند. اکثر دانشجویان آزمون اندازه واژگان را آزمون بهتر و قوی‌تری تشخیص دادند (۲۲ و ۱۸ نفر به ترتیب). در خصوص روایی و پایایی (سؤال پنجم)، ۲۲ نفر در آزمون ۱۴۰۰ و ۱۹ نفر در آزمون ۱۴۰۳ بخش واژگان آزمون را فاقد روایی و پایایی مناسب دانستند و محتوای آزمون را برای هدف مورد نظر شایسته ارزیابی نکردند. ۲۱ نفر در آزمون ۱۴۰۰ و ۱۶ نفر در آزمون دیگر معتقد بودند که این کلمات را در متون تخصصی کارشناسی ارشد نخواهند دید یا بسیار به‌ندرت می‌بینند؛ البته برخی وجود این کلمات را در متون تخصصی محتمل دانستند. در خصوص دو سؤال آخر، ۸۵ درصد دانشجویان این درجه از دشواری را نامناسب و این آزمون را برای تمییز

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۵، شماره ۴، زمستان ۱۴۰۴، از صفحه ۳۴۹ تا ۳۷۵

بین دانشجویان قوی و ضعیف و پذیرش دانشجویان کارشناسی ارشد ناکارآمد تلقی کردند.

اساتید نیز نظرات مشابهی داشتند و دشواری، تمایز، کارآمدی گزینه‌ها، روایی و پایایی آزمون را نامناسب ارزیابی کردند و در توجیه پاسخ خود به دلایل زیر اشاره نمودند: ضعف روایی محتوایی به علت عدم تطابق بین محتوای آزمون و محتوای دروس کارشناسی و کارشناسی ارشد، عدم تطابق بین هدف (انتخاب دانشجوی کارشناسی ارشد) و محتوای آزمون، عدم تطابق بین سطح دشواری واژگان و سطح دانش زبانی دانشجویان سال آخر کارشناسی، ضعف پایایی درونی آزمون به‌دلیل تفاوت در سطح دشواری سؤالات و ضعف پایایی زمانی به علت تصادفی بودن برخی از پاسخهای دانشجویان و در نتیجه ناپایداری نمرات.

۶. بحث و استدلال

پس از تجزیه‌وتحلیل داده‌ها نوبت به بحث در مورد نتایج به‌دست‌آمده، پاسخ به سؤالات تحقیق و توضیح و توجیه نتایج و ارائه راهکارها می‌رسد. درخصوص سؤال اول تحقیق، بررسی واژگان در پیکره‌ها، فرهنگ لغتها و لیست‌های مهم واژگان نشان داد که واژگان آزمونها بسامد مناسبی در پیکره‌های عمومی و تخصصی ندارند. حدود نیمی از این واژگان در پیکره‌های BNC و COCA بسامد مناسبی نداشتند و یا حتی موجود نبودند و بیش از ۶۱ درصد در پیکره‌های تخصصی وجود نداشتند و یا بسامد بسیار کمی داشتند. بنابراین، در پاسخ به سؤال اول تحقیق باید گفت درصد بسیار بالایی از کلمات بخش واژگان در منابع مذکور وجود نداشتند یا بسامد ناچیزی داشتند. نبود یا بسامد بسیار پایین این کلمات در متون دانشگاهی رشته‌های زبان انگلیسی بدین معناست که دانشجویان در متون تخصصی خود با این کلمات روبرو نخواهند شد و محتوای آزمون با دامنه هدف که واژگان متون دانشگاهی مربوطه می‌باشد تناسبی ندارد. در نتیجه بخش واژگان آزمون در بُعد تعریف دامنه قابل دفاع نمی‌باشد و این آزمون برای سنجش دانش واژگانی مورد نیاز برای مطالعه و نگارش متون علمی این رشته‌ها مناسب نمی‌باشد. پاسخهای دانشجویان و اساتید زبان به سؤالات ۶ و ۷ پرسش‌نامه نیز مؤید دشواری بیش از حد کلمات بخش واژگان می‌باشد و محتوای آزمون را با دامنه هدف متناسب نمی‌داند. دلیل این مسئله این است که ظاهراً تمرکز طراحان بر گنجاندن کلمات بسیار دشوار در گزینه‌ها می‌باشد و سعی می‌شود کلماتی که داوطلبان کمتر با آنها برخورد داشتند مورد سنجش قرار گیرد. همچنین بسامد واژگان در پیکره‌ها و لیست‌های واژگان پرکاربرد بررسی نمی‌شود و کلمات صرفاً بر اساس نظر طراحان آزمون انتخاب می‌شوند، درحالی‌که بسامد واژگان یکی از مهم‌ترین شاخصه‌ها و نشانگرهای سطح دشواری کلمات می‌باشد (اشمیت و دیگران، ۲۰۱۹؛ چویی و مون، ۲۰۱۹). علت استفاده نکردن طراحان از پیکره‌ها و لیست‌های واژگان می‌تواند عدم دسترسی آنان به این منابع و یا عدم توانایی آنان در استفاده از ابزارهای بسامدسنجی باشد. دلیل دیگر ممکن است پیروی طراحان از شیوه‌های سنتی طرح سؤال و آزمون‌های پیشین که صرفاً کلمات بسیار سخت و کم‌کاربرد را می‌سنجیدند باشد. همان‌طور که مدسن[۱] (۱۲:۱۹۸۳) بیان می‌دارد درطراحی آزمون واژگان «فقط کلمات سخت را انتخاب کردن یا انتخاب از فهرست تصادفی کلمات خیلی منطقی نیست؛ ما باید دریابیم زبان‌آموزان ما چه کلماتی را نیاز هست یاد بگیرند». کرمل[۲] (۲۰۱۶) نیز معتقد است که نباید از شیوه‌های پیشین طراحی سؤال بدون بررسی اصول زیربنایی آنها پیروی کرد. لذا پیشنهاد می‌شود به‌جای پیروی از سبک آزمون‌های سابق و انتخاب کلمات کم‌بسامد، طراحان آزمون با بررسی لیست‌های واژگان پرکاربرد و پیکره‌های عمومی و تخصصی، کلمات مناسب‌تری را برای بخش واژگان برگزینند. همچنین طراحان می‌توانند واژگان پرکاربرد در متون دانشگاهی مربوطه را با کمک پیکره‌ها و ابزارهای بسامدسنجی استخراج کرده و کلمات آزمون‌های واژگان را از آن برگزینند.

پژوهش‌های مرتبط معدودی در این حیطه وجود دارد که اکثراً مؤید یافته این بخش از تحقیق می‌باشند (جنینگز و دیگران، ۲۰۲۴؛ رفعت بخش و احمدی، ۲۰۲۲، ۲۰۲۴). رفعت بخش و احمدی (۲۰۲۲) دریافتند که فقط ۲٫۹ درصد کلمات بخش واژگان آزمون دکترای گروه انگلیسی از فهرست واژگان AWL می‌باشد؛ در تحقیق حاضر نیز تنها ۲٫۶ درصد کلمات آزمون در لیست AWL وجود داشتند. رفعت بخش و احمدی (۲۰۲۴) نیز متوجه شدند کلمات

[۱] Madsen

[۲] Kremmel

بخش واژگان آزمون‌های کارشناسی ارشد عمدتاً بسامد مورد انتظار را در زیرپیکره دانشگاهی COCA ندارند و با متون دانشگاهی چندان مرتبط نیستند. آنها، همانند محقق حاضر، دریافتند که برخی از کلمات آزمون کارشناسی ارشد بسامد مناسبی در پیکره عظیم COCA ندارند: سه کلمه در پیکره COCA وجود نداشت (*nefandous, saporous, containerport*) و ۱۳ کلمه نیز بسامدی کمتر از ده داشتند (کلماتی مانند *defalcate, clamant, banausic, animadversion, torpidity, perorate, depredate, paralogism, acidulous*). محقق حاضر با بررسی این کلمات در فرهنگ لغت‌های معتبر دریافت که هیچ‌کدام از این کلمات در فرهنگ لغت لانگمن وجود ندارند و فقط یک کلمه در فرهنگ لغت کمبریج موجود می‌باشد (*torpidity*). وجود چنین کلماتی در آزمون کارشناسی ارشد اعتبار و کارآیی آزمون را خدشه‌دار خواهد کرد. در تحقیق دیگری، جنینگز و دیگران دریافتند که واژگان ارزیابی‌شده در آزمون سرنوشت‌ساز GCSE در انگلستان عمدتاً کم‌بسامد هستند و بیشتر در متون ادبی قدیمی‌یافت می‌شوند. بیش از ۷۰ درصد از واژگان این آزمون در پیکره BNC بسامدی کمتر از پنج داشتند، درحالی‌که بسامد کمتر از ۵ در هر یک ملیون لغت در زبان عمومی بسامد پایینی محسوب می‌شود (برایسبرت و دیگران، ۲۰۱۸، ذکر شده در جنینگز و دیگران).

در پاسخ به سؤال دوم تحقیق، بررسی ویژگی‌های سؤالات آزمون نشان داد که ۷۰ درصد از سؤالات شاخص دشواری مناسبی نداشتند، ۳۸ درصد از شاخص تمایز مناسبی برخوردار نبودند و توزیع گزینه‌ها هم در ۵۸ درصد از سؤالات مناسب نبود. فقط یک سؤال (سؤال ۱۵ در آزمون ۱۴۰۲) هر سه شاخص را باهم داشت و ۴۹ سؤال دیگر در یک تا سه شاخص دچار مشکل بودند. تعداد قابل توجهی از سؤالات بخش واژگان تا حد زیادی ویژگی‌های یک سؤال خوب را نداشتند و در یک تا سه ویژگی دچار مشکل بودند، در نتیجه بُعد ارزیابی بخش واژگان دچار مشکل بوده و قابلیت دفاع ندارد. این یافته نشان می‌دهد که سؤالات این آزمون برای سنجش دانش واژگانی و گزینش داوطلبان کارشناسی ارشد زبان انگلیسی مناسب نمی‌باشند و باید اصلاح شوند. علت این مسئله گنجاندن واژگان بیش از حد سخت، کم‌کاربرد و نادر در گزینه‌های سؤالات می‌باشد که سطح دشواری سؤال را بسیار بالا می‌برد و ازآنجایی‌که نه دانشجویان ضعیف با این کلمات آشنا هستند و نه دانشجویان قوی، شاخص تمایز نیز بسیار پایین خواهد بود. از طرف دیگر چون سطح دشواری گزینه‌های هر سؤال متفاوت بود توزیع گزینه‌ها در سؤالات مشکل پیدا می‌کردند. در برخی از سؤالات، یک گزینه بسیار ناآشنا وجود داشت که افراد بسیار کمی آنرا انتخاب کرده بودند و یا یک گزینه بسیار آسان که انتخاب افراد زیادی بود. مثلاً گزینه‌های *askance* و *undergrid* در سؤالهای هفتم و نهم آزمون ۱۴۰۳ نسبت به گزینه‌های *suspicious* و *propose* بسیار سخت‌تر بوده و بسیار کمتر انتخاب شده بودند. در آزمون ۱۴۰۲ نیز کلمات *rue* و *rotund* در سؤالهای اول و دوم از کلمات *emulate* و *condescending* خیلی کمتر انتخاب شده بودند. بررسی این کلمات در فهرست واژگان BNC/COCA نشان داد که کلماتی که کمتر انتخاب شدند بسامد بسیار پایین‌تری نسبت به کلمات دیگر دارند.

ظاهراً فرض طرحان این آزمون بر این است که گنجاندن کلمات بسیار سخت در آزمون واژگان دانشجویانی که دامنه واژگان قوی‌تری دارند را بهتر شناسایی خواهند کرد، غافل از اینکه بسیاری از این کلمات برای دانشجویان قوی نیز بیش از حد سخت بوده و این سؤالات نمی‌توانند دانشجویان قوی‌تر را به‌خوبی شناسایی کنند. براون و آبی ویکراما[۱] (۲۰۱۹: ۷۵) معتقدند که «سؤالات بسیار دشوار حتی برای آزمون‌شوندگان با بالاترین سطح توانایی نیز می‌توانند چالش برانگیز باشند» و فرهادی و دیگران[۲] (۱۹۹۴) سؤالات بسیار سخت و بسیار آسان را توصیه نمی‌کنند، چراکه این سؤالات اطلاعات مفیدی را در مورد آزمون دهندگان فراهم نمی‌کنند. لذا توصیه می‌شود طراحان آزمون درجه سختی کلمات را در طراحی سؤالات بخش واژگان در نظر بگیرند و کلمات بسیار سخت را در گزینه‌ها نگنجانند. چون اجرای آزمایشی آزمون‌های سرنوشت‌ساز به دلایل امنیتی ممکن است میسر نباشد، طراحان آزمون می‌توانند بسامد کلمات را به‌عنوان یکی از نشانگرهای درجه سختی کلمات مورد استفاده قرار دهند. اجرای مجدد آزمون‌های پیشین با دانشجویان سال آخر کارشناسی و انجام تحلیل سؤالات نیز

[۱] Brown & Abeywickrama

[۲] Farhady et al.

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۵، شماره ۴، زمستان ۱۴۰۴، از صفحه ۳۴۹ تا ۳۷۵

می‌توانند اطلاعات و آگاهی‌هایی را در خصوص سؤالهای خوب و واژگان مناسب به طراحان آزمون ارائه کنند.

یافته‌های این بخش با نتایج بسیاری از پژوهشهای پیشین که بخش واژگان آزمون‌های سرنوشت‌ساز دیگر را بررسی کرده‌اند همسویی و تطابق دارد (جمالی فر و دیگران، ۲۰۱۴؛ رفعت بخش و احمدی؛ ۲۰۲۲؛ قهرکی و دیگران[1]، ۲۰۲۲؛ مرندی و دیگران، ۲۰۲۰؛ یوجی و ونشیا، ۲۰۰۷). این تحقیقات نیز مانند تحقیق حاضر نشان دادند که سؤالات واژگان برخی از آزمون‌های سرنوشت‌ساز ویژگی‌های یک سؤال خوب را دارا نمی‌باشند. در تحقیق جمالی فر و دیگران، ۱۲ سؤال از بیست سؤال بخش واژگان آزمون کارشناسی ارشد دانشگاه آزاد ضریب دشواری بسیار بالایی داشتند و دو سؤال نیز بیش از حد آسان بودند؛ دوازده سؤال نیز از توزیع گزینه‌های مناسبی برخوردار نبودند. در بخش واژگان آزمون دکترای رشته‌های زبان انگلیسی که توسط رفعت بخش و احمدی بررسی شد، میانگین نمرات دانشجویان در پنج آزمون (آزمون‌های ۱۳۹۴ تا ۱۳۹۸) نمره دو از ۱۲ بود، که نشان‌دهنده دشواری بیش از حد سؤالات این بخش می‌باشد. در پژوهش مرندی و دیگران بر روی آزمون‌های زبان انگلیسی وزارت بهداشت یازده سؤال از ۲۴ سؤال بخش واژگان شاخص تمایز پایین‌تر از ضریب معیار ۰,۳ داشتند و ۴ سؤال هم شاخص دشواری مناسبی نداشتند. قهرکی و دیگران نیز نشان دادند که در آزمون زبان عمومی وزارت علوم برای دانشجویان دکترا (MSRT)، ۴۸ درصد از سؤالات شاخص دشواری نامناسب و ۲۴ درصد شاخص تمایز نامناسب داشتند. در آزمون زبان انگلیسی دانشجویان دکترا در چین (۲۰۰۶) که توسط یوجی و ونیشا ارزیابی شد، فقط ده درصد از سؤالات دو شاخص دشواری و تمایز مناسب را توامان داشتند.

در پاسخ به سؤال سوم تحقیق، بررسی‌ها نشان داد که ضریب پایایی بخش واژگان آزمون بسیار پایین و غیرقابل قبول می‌باشد و در نتیجه بعد تعمیم برای بخش واژگان آزمون قابل دفاع نیست. بنابراین، پاسخ سؤال سوم تحقیق خیر می‌باشد و فرضیه پوچ مربوطه (بخش واژگان آزمون‌های کارشناسی ارشد زبان انگلیسی پایایی قابل قبولی ندارد) پذیرفته می‌شود. این یافته نشان می‌دهد که نمرات داوطلبان در این آزمون پایداری مناسبی ندارند و برای ارزیابی دانش واژگانی و پذیرش داوطلبان مناسب نیستند. علت این مسئله دشواری (و در برخی موارد محدود سهولت) بیش از حد واژگان آزمونها بوده است. سطح دشواری سؤالات و گزینه‌ها هم متفاوت بود که منجر به کاهش پایایی درونی آزمون می‌شد.

برای سؤال چهارم تحقیق مقایسه نمرات داوطلبان در این آزمون و آزمون اندازه واژگان نشان داد که آزمون واژگان کارشناسی ارشد روایی مناسبی ندارد و آنچه را که باید بسنجد نمی‌سنجد. لذا پاسخ سؤال چهارم، همانند سؤال قبلی، خیر بوده و فرضیه پوچ مربوطه (بخش واژگان آزمون‌های کارشناسی ارشد زبان انگلیسی روایی قابل قبولی ندارد) پذیرفته می‌شود. این یافته نشانگر این است که استفاده از این آزمون برای پذیرش دانشجویان کارشناسی ارشد رشته‌های زبان انگلیسی روا نمی‌باشد و باید از آزمون دیگری استفاده شود. علت این مسئله نیز دشواری بیش از حد کلمات آزمون، بسامد بسیار پایین آنها در متون عمومی و تخصصی و نامرتبط بودن آنها به متون تخصصی رشته‌های مربوطه می‌باشد. واژگان این آزمون نسبت به کلمات بخش‌های ششم و هفتم آزمون اندازه واژگان، که مناسب دانشجویان کارشناسی ارشد هستند، بسامد بسیار پایین‌تری دارند. علاوه‌بر این، مقایسه نمرات دانشجویان در دو آزمون نشان داد که آزمون واژگان کارشناسی ارشد به‌طور معناداری از آزمون اندازه واژگان سخت‌تر می‌باشد (Sig = 0.00). میانگین نمرات دانشجویان در بخش واژگان آزمون‌های سالهای ۱۴۰۰، ۱۴۰۲ و ۱۴۰۳ به ترتیب ۴,۸۰، ۳,۷۷ و ۴,۳۳ بود، درحالی‌که میانگین نمرات همان دانشجویان در آزمون اندازه واژگان به ترتیب ۱۰,۱۷، ۱۱/۵۰ و ۱۱/۶۱ بود. این نشان می‌دهد که سطح آزمون واژگان کارشناسی ارشد بسیار فراتر از دانش واژگانی دانشجویان (و حتی بومی‌زبانان و اساتید) است و آنچه را که باید بسنجد نمی‌سنجد و درنتیجه روایی لازم را ندارد. ارزیابی آزمون توسط برخی از اساتید و متخصصین و پاسخهای ایشان به سؤالات ۴ و ۵ پرسش‌نامه نیز مؤید ضعف روایی و پایایی این آزمون بوده است. ظاهراً طراحان آزمون تلاش دارند درجه سختی این آزمون را در حد آزمون‌هایی که برای

[1] Ghahraki et al.

پذیرش دانشجویان تحصیلات تکمیلی در کشورهای انگلیسی زبان مثل آمریکا و کانادا استفاده می‌شود، مثلاً آزمون جی آری ای[1]، و یا حتی بالاتر قرار دهند تا دانشجویان قوی را بهتر شناسایی کنند، درحالی‌که شرایط متفاوت هست و نیاز دانشجویان تحصیلات تکمیلی ایرانی متفاوت از نیازهای دانشجویان انگلیسی‌زبان می‌باشد. برخلاف دانشجویان انگلیسی‌زبان، دانشجویان ایرانی با بسیاری از کلمات به‌کاررفته در متون تخصصی رشته‌های زبان انگلیسی آشنا نیستند. تصور می‌شود آزمونی که واژگان به‌کاررفته در متون تخصصی مرتبط را بسنجد روایی و پایایی بهتری خواهد داشت. توصیه می‌شود طراحان بخش واژگان آزمون با بررسی ویژگی‌های سؤالات واژگان در آزمون‌های معیار مانند آزمون اندازه واژگان و آزمون‌های استادارد بین المللی مانند تافل و آیلتس، سؤالات و آزمون‌های بهتر و موثرتری را تهیه و تولید نمایند.

یافته‌های این دو بخش با نتایج برخی از تحقیقات پیشین که پایایی و روایی آزمون کارشناسی ارشد زبان انگلیسی یا آزمون‌های سرنوشت‌ساز دیگر در ایران یا کشورهای دیگر را بررسی کرده‌اند همسو و مطابق است و با برخی دیگر متناقض. برخی از این تحقیقات، همانند تحقیق حاضر، نشان دادند که آزمون‌های سرنوشت‌ساز روایی و پایی لازم را ندارند، ولی تحقیقات دیگر روایی و پایایی مناسبی گزارش کردند (اُبیوما و سالاو، ۲۰۰۷؛ پاردو–بالستر، ۲۰۱۰؛ جمالی فر و دیگران، ۲۰۱۴؛ خامبونروانگ، ۲۰۲۵؛ خودی و دیگران، ۲۰۲۱؛ راوند و فیروزی، ۲۰۱۶؛ شیخ الاسلامی، ۱۹۹۹؛ علوی و دیگران، ۲۰۲۱؛ قاسمی ورزنه، ۲۰۰۵؛ کومازاوا و دیگران، ۲۰۱۶؛ مرندی و دیگران، ۲۰۲۰). راوند و فیروزی نشان دادند که بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی پایایی ضعیفی دارد (۰,۴۵) و مناسب سطح آزمون دهندگان نمی‌باشد. شیخ الاسلامی رابطه همبستگی ضعیفی بین عملکرد دانشجویان در بخش واژگان آزمون کارشناسی ارشد زبان انگیسی و بخش واژگان آزمون تافل به دست آورد و در نتیجه روایی ضعیفی برای بخش واژگان آزمون کارشناسی ارشد گزارش کرد. مرندی و دیگران نیز دریافتند که آزمون زبان انگلیسی وزارت بهداشت از روایی سازه مناسبی برخوردار نیست. در مقابل، در تحقیق جمالی فر و دیگران پایایی آزمون کارشناسی ارشد قابل قبول و مناسب ارزیابی شد و قاسمی ورزنه رابطه همبستگی بالایی بین نمرات آزمون کارشناسی ارشد و آزمون آیلتس به دست آورد. علوی و دیگران نیز پایایی و روایی مناسبی را برای زبان عمومی کنکور کارشناسی به دست آوردند. در خصوص آزمون‌های سرنوشت‌ساز کشورهای دیگر، کومازاوا و دیگران روایی و پایایی مناسبی برای آزمون‌های ورودی دانشگاه در ژاپن پیدا کردند ولی اُبیوما و سالاو نشان دادند که آزمون‌های ورودی نیجریه روایی و پایایی لازم را برای انتخاب دانشجویان ندارند.

مقایسه عملکرد گروه‌های مختلف شرکت‌کنندگان به علت اینکه برخی گروه‌ها فقط در یک آزمون شرکت کرده بودند، به‌طور کامل قابل انجام نبود. تنها در آزمون سال ۱۴۰۰ همه شرکت‌کنندگان حضور داشتند و فقط پنج دانشجوی کارشناسی ارشد در آزمون ۱۴۰۳ شرکت کرده بودند. مقایسه نمرات شرکت‌کنندگان در آزمون ۱۴۰۰ نشان داد که تفاوت معناداری بین عملکرد گروه‌های شرکت‌کننده وجود دارد ($F(3,129) = 17.26, p < 0.001$). میانگین نمرات بومی‌زبانان، اساتید، دانشجویان کارشناسی ارشد و دانشجویان کارشناسی به ترتیب ۱۱,۳۳، ۸,۸۷، ۵,۴۱ و ۴,۸۰ بود. تفاوت معنادارای بین عملکرد دانشجویان کارشناسی و کارشناسی ارشد وجود نداشت ولی عملکرد اساتید و بومی‌زبانان به‌طور معناداری بهتر از دانشجویان بود. تفاوت بین میانگین نمرات اساتید و بومی‌زبانان از نظر آماری معنادار نبود، اگرچه نمرات بومی‌زبانان به‌طور قابل مشاهده‌ای بهتر بود. در آزمون‌های ۱۴۰۲ و ۱۴۰۳ نیز میانگین نمرات اساتید به‌طور معناداری بهتر از میانگین دانشجویان کارشناسی بود، اگرچه فاصله نمرات نسبت به آزمون ۱۴۰۰ کمتر بودند. میانگین نمرات اساتید و دانشجویان کارشناسی در آزمون ۱۴۰۲ به ترتیب ۶,۳۳ و ۳,۷۷ و در آزمون ۱۴۰۳ به ترتیب ۷,۵۸ و ۴,۳۳ بود. البته نمرات تعداد قابل توجهی از دانشجویان کارشناسی از نمرات برخی اساتید بسیار بالاتر بود و تفاوت میانگینها بیشتر به خاطر دانشجویان بسیار ضعیف بود که هیچ‌کدام از سؤالات را جواب صحیح نداده بودند و یا یک یا دو سؤال را پاسخ صحیح داده بودند. از طرف دیگر، اگرچه این آزمون اساتید و دانشجویان را متمایز

[1] GRE

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۵، شماره ۴، زمستان ۱۴۰۴، از صفحه ۳۴۹ تا ۳۷۵

می‌سازد ولی به‌خوبی بین دانشجویان کارشناسی و کارشناسی ارشد تمایز ایجاد نمی‌کند و درنتیجه برای شناسایی دانشجویان قوی‌تر و پذیرش دانشجویان کارشناسی ارشد مناسب به نظر نمی‌رسد. انتظار می‌رود دانشجویان کارشناسی ارشد که با یکی از همین آزمونها پذیرفته شده‌اند عملکرد بسیار بهتری از دانشجویان کارشناسی داشته باشند.

در خصوص سؤال آخر تحقیق، بررسی پاسخهای شرکت‌کنندگان به سؤالات پرسش‌نامه نشان داد که از نظر آنها این آزمون برای پذیرش دانشجویان کارشناسی ارشد مناسب نمی‌باشد و پاسخ سؤال پنجم تحقیق (در نظر اساتید و دانشجویان زبان انگلیسی بخش واژگان آزمون برای سنجش دانش واژگانی داوطلبین و پذیرش دانشجویان کارشناسی ارشد مناسب است؟) خیر می‌باشد. این یافته نشان می‌دهد که بخش واژگان آزمون باید اصلاح شود و آزمون‌های مناسب‌تری برای سنجش دانش واژگانی و پذیرش داوطلبان کارشناسی ارشد زبان انگلیسی طرح شوند. دلیل این مسئله نیز دشواری بیش از حد کلمات می‌باشد؛ شرکت‌کنندگان معتقد بودند برای شناسایی دواطلبین قوی باید درجه دشواری کلمات در حد مناسب‌تری باشد. محتوای آزمون نیز از نظر اکثر شرکت‌کنندگان با دامنه هدف تناسب خوبی ندارد و واژگان ارزیابی‌شده در متون تخصصی کارشناسی ارشد دیده نخواهد شد. بررسی واژگان در پیکره‌های تخصصی مؤید نظرات شرکت‌کنندگان بوده است. بررسی نظرات دانشجویان و اساتید در خصوص آزمون‌های پیشین و استفاده از پیکره‌های تخصصی و ابزارهای سنجش بسامد می‌تواند کمک خوبی برای طراحان آزمون در انتخاب واژگان مناسب‌تر برای آزمون‌های ورودی رشته‌های زبان انگلیسی باشد.

تحقیقات مرتبطی که با استفاده از پرسش‌نامه و نظرسنجی آزمون‌های سرنوشت‌ساز دیگری را مورد ارزیابی و اعتبار سنجی قرار داده‌اند نتایج مشابهی ارائه کرده‌اند (اقلیدی و طباطبایی[1]، ۲۰۱۸؛ پیش قدم و دیگران، ۱۳۹۹؛ قربانی و دیگران[2]، ۲۰۲۲). پیش قدم و دیگران نشان دادند که دانشجویان دکترا و اساتید دانشگاه رضایت چندانی از آزمون بسندگی زبان انگلیسی وزارت علوم ندارند و روایی، پایایی، تأثیر و عدالت آزمون را نامناسب ارزیابی می‌کنند. در تحقیق اقلیدی و طباطبایی معلمین زبان انگلیسی روایی آزمون نهایی زبان انگلیسی مدارس را متوسط ارزیابی کردند. و شرکت‌کنندگان در تحقیق قربانی و دیگران معایبی را برای آزمون بسندگی وزارت علوم بر شمردند که شامل موارد زیر بود: عدم تطابق بین محتوای آزمون و نیازهای دانشجویان دکترا، اثر بازگشتی منفی[3]، محتوای غیر مبتنی بر نظریه و فقدان اصالت در سؤالات آزمون.

۷. نتیجه‌گیری ، محدودیتها و پیشنهادات

نتایج تحقیق حاضر نشان داد که بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی در بعدهای مختلف اعتبارسنجی مبتنی بر استدلال (تعریف دامنه، ارزیابی، تعمیم، بیرون افکنی و پیامد) دچار مشکل می‌باشد و استفاده از نمرات و نتایج این آزمون برای هدف مدنظر (پذیرش دانشجویان کارشناسی ارشد) موجه و قابل دفاع نمی‌باشد. بررسی‌ها نشان داد که کلمات آزمون در پیکره‌های عمومی و تخصصی بسامد مناسبی ندارند و بسیاری از کلمات در فهرست‌های مهم واژگان عمومی و دانشگاهی و فرهنگ لغتها وجود ندارند، سؤالات آزمون از ویژگی‌های یک سؤال مناسب (شاخص دشواری، شاخص تمایز و توزیع گزینه‌ها) برخوردار نیستند، روایی و پایایی آزمون مناسب نمی‌باشند و از نظر اساتید و دانشجویان این آزمون برای پذیرش دانشجویان کارشناسی ارشد مناسب نیست. لذا توصیه می‌شود طراحان آزمون با بررسی علل و شیوه‌های برطرف کردن این کاستی‌ها، که به برخی از آنها در بخش پیش اشاره شد، در ارتقاء کیفیت این آزمون که سرنوشت بسیاری از افراد به نتایج آن گره خورده است بکوشند. همچنین توصیه می‌شود طراحان آزمون از طرح سؤالات بر اساس شیوه آزمون‌های سابق و بدون ارزیابی عملی فرضهای زیربنایی اصول آزمون‌سازی بپرهیزند (کرمل، ۲۰۱۶). اشمیت و دیگران (۲۰۱۹: ۱) بر اعتبارسنجی نظام‌مند و دقیق‌تر آزمون‌های واژگان و ارتقای سواد ارزیابی و سنجش در میان طراحان این آزمونها تأکید دارند. ازآنجایی‌که برگزاری پیش‌آزمون و اجرای آزمایشی برای ارزیابی این آزمون به

[1] Eghlidi & Tabatabaei

[2] Ghorbani et al.

[3] negative washback

لحاظ امنیتی ممکن نیست، پیشنهاد می‌شود طراحان آزمون با بررسی نظری و محتوایی گزینه‌ها و استفاده از اطلاعات در مورد بسامد و کاربرد واژگان در پیکره‌های عمومی و تخصصی، گزینه‌های مناسب‌تری را برای سؤالات بیابند. همچنین، پیشنهاد می‌شود لغات و اصطلاحاتی که در نگارش متون علمی و پژوهشی رشته‌های دانشگاهی ضروری و در سطح دشواری مناسبی هستند و می‌توانند از طریق بررسی پیکره به دست آیند (مانند باقری نویسی و دیگران[۱]، ۲۰۲۳؛ صفری[۲]، ۲۰۱۸، ۲۰۱۹؛ فراهانی و دیگران[۳]، ۱۳۹۹) در آزمون‌های واژگان گنجانده شوند.

مانند هر پژوهش دیگری، تحقیق حاضر حدود و محدودیت‌هایی داشت که تعمیم نتایج تحقیق را محدود می کند و مستلزم پژوهشهای بیشتر خواهد بود. شرکت‌کنندگان در این تحقیق ۱۹۴ دانشجوی کارشناسی، ۲۴ دانشجوی کارشناسی ارشد و ۱۶ استاد زبان انگلیسی در دانشگاه‌های شهر قم و شش بومی زبان بودند. امید است تحقیقات بعدی با تعداد بیشتری از شرکت‌کنندگان از مکانهای جغرافیایی دیگر آزمون کارشناسی ارشد گروه انگلیسی و آزمون‌های سرنوشت‌ساز دیگر را مورد ارزیابی قرار دهند. به علت محدودیت، تعداد شرکت‌کنندگان بومی زبان محدود به شش نفر بودند که البته بیشتر برای نشان دادن سطح سختی واژگان آزمون به کار گرفته شدند و همین تعداد تا حد کافی بازتاب‌دهنده سختی بیش از حد این واژگان بوده است. از طرف دیگر، چون در این تحقیق آزمون‌های چندین سال بررسی شدند تعداد شرکت‌کنندگان در هر آزمون بسیار کمتر بود و به علت محدودیت در تعداد شرکت‌کنندگان از شیوه‌های پیچیده‌تر مانند مدل رش استفاده نشد. محققین بعدی می‌توانند این آزمون را با استفاده از شیوه‌ها و مدلهای نوین و پیچیده‌تر بررسی کنند. البته اعتبارسنجی مبتنی بر استدلال تأکیدی بر روشها و ابزار پیچیده ندارد و شیوه‌های ساده‌تر نیز مقبول می‌باشند (ایم و دیگران[۴]، ۲۰۱۹؛ کین، ۲۰۱۳). همچنین تحقیق حاضر محدود به بخش واژگان آزمون کارشناسی ارشد زبان انگلیسی بود؛ تحقیقات بعدی می‌توانند بخش واژگان آزمون دکترا و بخش‌های خواندن و درک مطلب و دستور زبان را در آزمون‌های کارشناسی ارشد و دکترا مورد ارزیابی قرار دهند. برخی از جنبه‌های آزمون‌های سرنوشت‌ساز مانند اثر بازگشتی آزمون[۵]، تبعیض[۶]، وعدالت[۷] در تحقیق حاضر بررسی نشدند؛ توصیه می‌شود محققین بعدی این جنبه‌ها را در آزمون ورودی کارشناسی ارشد و دکترای گروه زبان انگلیسی بررسی کنند. بسامد واژگان آزمون ورودی دکترا در پیکره‌های عمومی و تخصصی هم می‌تواند موضوع تحقیقات آینده باشد. همچنین مطالعات تطبیقی برای مقایسه بخش واژگان آزمون‌های ورودی دانشگاه در ایران و کشورهای دیگر می‌تواند موضوع پژوهش محققین دیگر باشد. در آخر، امید است نتایج تحقیق حاضر و تحقیقات دیگر در ارتقاء کیفیت آزمون‌های سرنوشت‌ساز مؤثر باشند.

با تشکر از تمام اساتید و دانشجویانی که در جمع آوری داده های این پژوهش همکاری داشتند.

منابع فارسی

پیش قدم، رضا، ابراهیمی، شیما، شایسته، شقایق، طباطبایی فارانی، سحر و جاجرمی، هانیه (۱۳۹۹). بررسی و آسیب‌شناسی آزمون‌های بسندگی زبان انگلیسی وابسته به وزارت عتف و دانشگاه‌های ایران و نیازسنجی زبانی ذی‌نفعان. *پژوهشهای زبان شناختی در زبانهای خارجی*، ۱۰ (۴)، ۷۰۵-۶۸۶.
https://doi.org/10.22059/jflr.2021.317539.801

جعفرپور، عبدالجواد (۱۳۷۶). نقدی بر آزمون انگلیسی ورودی تحصیلات تکمیلی (مجموعه زبان انگلیسی). *مجله پژوهشی دانشگاه اصفهان (علوم انسانی)*، ۱ (۲)، ۱۵-۲۲.

خسروانی، محمود، رستمیان، مرتضی، اشرف، حمید و خدابخش زاده، حسین (۱۴۰۱). پدیدارشناسی،

[۱] Bagheri Nevisi et al.

[۲] Safari

[۳] Farahani et al.

[۴] Im et al.

[۵] washback effect

[۶] bias

[۷] fairness

ساخت و اعتباربخشی انگارهٔ روایی پسینی سنجهٔ زبان در آزمون ورودی دانشگاه: کاربردهایی برای آزمون‌سازی در تحصیلات تکمیلی. *پژوهش و نگارش کتب دانشگاهی، ۲۶*(۵۱)، ۲۱۷-۱۸۹.
https://doi.org/10.30487/rwab.2023.1982793.1540

فراهانی، الهام، یزدانی، هوشنگ، احمدیان، موسی و عامریان، مجید (۱۳۹۹). بررسی کاربرد اصطلاحات انگلیسی در مقاله‌های پژوهشی زبان‌شناسی کاربردی: پژوهش پیکره محور. *پژوهشهای زبان شناختی در زبانهای خارجی، ۱۰*(۲)، ۴۰۵-۳۹۰.
https://doi.org/10.22059/jflr.2020.291411.694

منابع انگلیسی

Alavi, S. M., Karami, H., & Khodi, A. (2021). Examination of factorial structure of Iranian English language proficiency test: An IRT analysis of Konkur examination. *Current Psychology, 42*(10), 8097–8111. https://doi.org/10.1007/s12144-021-01922-1.

Amirian, S. M. R., Ghonsooly, B., & Amirian, S. K. (2020). Investigating fairness of reading comprehension section of INUEE: Learner's attitudes towards DIF sources. *International Journal of Language Testing, 10*(2), 88–100.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice. Oxford University Press.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5), 1052–1060. https://doi.org/10.4304/jltr.2.5.1052-106.

Bagheri Nevisi, R., Safari, M., Hosseinpur, R. M., & Mousakazemi, R.S. (2023). A high frequency word list for political sciences. *Journal of Modern Research in English Language Studies, 10*(4), 21-43.

Beglar, D. (2010) A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118. https://doi.org/10.1177/0265532209340194

Brown, D. H. & Abeywickrama, P. (2019). *Language assessment: Priciples and classroom practices (3rd ed.)*. Pearson Publications.

Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment (2nd ed.). McGraw-Hill College.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage Publications.

Farhady, H., Jafarpur, A. and Birjandi, P. (1994). *Testing language skills: From theory to practice*. SAMT Publications.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. Routledge.

Ghahraki, S., Tavakoli, M., Ketabi, S. (2022). Applying a two-parameter item response model to explore the psychometric properties: The case of the ministry of Science, Research and Technology (MSRT) high-stakes English Language Proficiency test. *Journal of English Language Teaching and Learning*, *14*(29), 1-26. https://doi.org/10.22034/ELT.2021.46325.2396

Ghasemivarzaneh, S. (2005). On the predictive validity of the proficiency section of the M.A. entrance examination in English language major [Unpublished master's thesis]. Allameh Tabatabaee University.

Ghorbani, M. R., Abbassi, H., & Razali, A. B. M. (2021). Exploring the Shortcomings of the Iranian MSRT English Proficiency Test. *Pertanika Journal of Social Sciences & Humanities*, *29*(S3), 115-132.

Gyllstad, H. (2012, June). *Validating the vocabulary size test: A classical test theory approach*. 9th EALTA Conference, Innsbruck, Austria.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items (3rd ed.). Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C.A. & Lee, H. (2021). Understanding argument-based validity in language testing. In C. A. Chapelle & E. Voss (Ed.), *Validity argument in language testing* (pp. 19-44). Cambridge University Press.

Choi, I. C., & Moon, Y. (2019). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, *17*(1), 18–42. https://doi.org/10.1080/15434303.2019.1674315

Cohen, L., Manion, L., & Morrison, K. (2017). Research methods in education *(8th ed.)*. Routledge.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

DeVellis, R. F. (2016). Scale development: Theory and applications (4th ed.). Sage Publications.

Eghlidi, M. & Tabatabaei, O. (2018). Evaluating the construct validity of dinal test of grade three of junior high school in Iran. *Journal of Applied Linguistics and Language Research*, *5*(2), 275-285.

Elder, C. & O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gains on IELTS. *IELTS Research Reports*, *4*, 207-254.

the vocabulary size test. *RELC Journal, 43*(1), 53-67. https://doi.org/10.1177/0033688212439359

Karami, H. (2013). An investigation of the gender differential performance on a high-stakes language proficiency test in Iran. *Asia Pacific Education Review, 14*(3), 435–444. https://doi.org/10.1007/s12564-013-9272-y

Khamboonruang, A. (2025). Argument-based validation of Chulalongkorn University Language Institute (CULI) test: A Rasch-based evidence investigation. *Language Testing in Asia, 15*(10), 1-24. https://doi.org/10.1186/s40468-025-00346-z

Khodi, A., Alavi, S. M., & Karami, H. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Language Testing in Asia, 11*(14), 1–10. https://doi.org/10.1186/s40468-021-00125-6.

Khoii, R. (1998). *A qualitative and quantitative evaluation of the English subtests of the entrance examinations of universities using Rasch model* [Doctoral dissertation]. Islamic Azad University, Science and Research Campus, Tehran.

Kumazawa, T., Shizuka, T., Mochizuki, M. & Mizumoto, A. (2016). Validity argument for the VELC Test® score interpretations and uses. *Language Testing in Asia, 6*(2), 1-18.

Hughes, A., & Hughes, J. (2020). Testing for language teachers (3rd ed.). Cambridge University Press.

Im, G., Shin, D. & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14). https://doi.org/10.1186/s40468-019-0089-4.

Jamalifar, G., Heidari Tabrizi, H. & Chalak, A. (2014). Islamic Azad University entrance examination of master program in TEFL: An analysis of its reliability of the general English section. *The Asian EFL Journal, 10*(5), 386-403.

Jennings, B., Powel, D., Jaworska, S. & Joseph, H. (2024). A corpus study of English language exam texts: Vocabulary difficulty and the impact on students' wider reading (or Should students be reading more texts by dead white men?). *Journal of Adolescent & Adult Literacy, 67*, 303-316. https://doi.org/10.1002/jaal.1331

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 17-64). American Council on Education/Praeger.

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 44-457.

Karami, H. (2012). The development and validation of a bilingual version of

cognitive load and difficulty estimates of language test items. *Language Testing in Asia, 12*(1), 1-19 https:// doi.org/10.1186/s40468-022-00163-8

Obioma, G. & Salau, M. (2007, September 16-21). *The predictive validity of public examinations: A case study of Nigeria*. The 33rd Annual Conference of International Association for Educational Assessment (IAEA) Baku, Azerbaijan.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*(2), 137-159, https://doi.org/10.1080/15434301003664188

Rafatbakhsh, E., & Ahmadi, A. (2022). The argument-based validation of a large-scale high-stakes vocabulary test. *Practical Assessment, Research, & Evaluation, 27*(28). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/28

Rafatbakhsh, E., & Ahmadi, A. (2024). A Corpus-based Evaluation of a High-stakes EFL Exam. *Journal of Studies in Language Learning and Teaching, 1*(2), 211-225.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing, 6*(1), 1–23.

https://doi.org/10.1186/s40468-015-0023-3.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly, 50*, 976-987. https://doi.org/10.1002/tesq.329.

Linn, R. L., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8th ed.). Prentice Hall.

Madsen, H. S. (1983). *Techniques in testing*. Oxford University Press.

Marandi, S., Tajik. L. & Zohali., L. (2020). On the construct validity of the Iranian Ministry of Health Language Exam (MHLE). *Journal of Language Horizons, 4*(2), 9-36.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (pp. 13–103). American Council on Education/Macmillan.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). Measurement and assessment in teaching (11th ed.). Pearson.

Nation, I.S.P. (2017). *The BNC/COCA Level 6 word family lists* (Version 1.0.0) [Data file]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between

Razavipur, K. (2014). On the substantive and predictive validity facets of the university entrance exam for English majors. *Research in Applied Linguistics, 5*(1), 77-90.

Safari, M. (2018). Do university students need to master the GSL and AWL words? A psychology word list. Journal of Modern Research in English Language Studies, 5(2), 101-122.

Safari, M (2019). English vocabulary for equine veterans: How different from GSL and AWL words. Iranian Journal of English for Academic Purposes, 8 (2), 51-65.

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*, 109 - 120. https://doi:10.1017/S0261444819000326

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

Sheikholeslami, A. (1999). *On the validity of the TEFL MA Entrance Examination proficiency tests in Iran* [MA thesis]. Allamehtabatabaee University.

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly, 8*(1), 34-52. https://doi.org/10.1080/15434303.2010.536924

Tarrant, M. & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today, 30*, 539-543. https://doi.org/10.1016/j.nedt.2009.11.002.

Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Palgrave Macmillan.

Wright, B. & Stone, M. (1979). *Best test design*. Mesa Press.

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371-386), Routledge.

Yujie, J. & Wenxia. Z. (2007). Evaluating the construct validity of an EFL test for PhD candidates: A quantitative analysis of two versions. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 2-16.

پیوست

ارزیابی بخش واژگان آزمون کارشناسی ارشد

۱- آیا به نظر شما این آزمون برای انتخاب دانشجویان کارشناسی ارشد رشته‌های زبان انگلیسی (مترجمی، ادبیات – انگلیسی، آموزش زبان انگلیسی) مناسب می‌باشد؟ توضیح مختصر با ذکر دلایل.

۲- برخی از لغات که بسیار سخت بودند را شما درست انتخاب کرده بودید؟ آیا شانسی و تصادفی زده بودید؟

۳- چه تعداد از گزینه‌ها (از هشتاد/شصت گزینه) را بلد نبودید؟
۴- در مقایسه با آزمون دیگر واژگان (Vocabulary Size Test) آزمون بهتر یا ضعیف‌تری بود؟ چرا؟
۵- به نظر شما این آزمون روایی (تا چه حد آنچه که باید ارزیابی کند را ارزیابی می‌کند) و پایایی (تا چه حد عملکرد آزمون دهندگان در تمام سؤالات یکسان یا مشابه می‌باشد) لازم را دارد؟
۶- آیا فکر می‌کنید دانشجویان کارشناسی ارشد زبان انگلیسی (مترجمی، ادبیات انگلیسی، آموزش زبان انگلیسی) این واژگان را در متون تخصصی خود خواهند دید؟
۷- به نظرتان این درجه از سختی برای سؤالات بخش واژگان آزمون کارشناسی ارشد مناسب است؟ چرا؟
۸- به نظرتان این آزمون می‌تواند از نظر زبان عمومی دانشجویان قوی، متوسط و ضعیف را مشخص کند؟ چرا؟