فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
دوره ششم، شماره ۲۲، زمستان ۹۴، ص ۴۱ تا ۶۸

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 6, No. 22, Winter 2016

تحلیل تشخیصی سؤال‌های بخش درک مطلب زبان انگلیسی عمومی آزمون ورودی دوره‌های دکتری با استفاده از مدل غیر جبرانی فیوژن

اعظم مقدم[۱]، محمدرضا فلسفی نژاد[۲]، نورعلی فرخی[۳]، معصومه استاجی[۴]

تاریخ دریافت: ۹۴/۰۱/۲۹
تاریخ پذیرش: ۹۴/۱۰/۱۵

چکیده

رویکردهای سنتی و جاری سنجش در بازنمایی قابلیت‌های زبانی افراد با چالش‌های عملی و نظری همراه بوده و توانایی آن‌ها در سنجش و گزینش داوطلبان موردتردید قرار گرفته است. به‌گونه‌ای که متخصصان حوزه اندازه‌گیری آموزشی را به استفاده از روش‌های جدیدتر و کارآمدتر، سنجش تشخیصی شناختی، سوق داده است. هدف مقاله حاضر تحلیل تشخیصی سؤالات بخش درک مطلب آزمون ورودی دوره‌های دکتری زبان انگلیسی عمومی با استفاده از مدل غیر جبرانی فیوژن به‌منظور شناسایی مهارت‌های آزمون موردمطالعه، کفایت مدل ارائه‌شده (همگرایی و برازش مدل)، توان تشخیصی آزمون و میزان تسلط داوطلبان در هر یک از مهارت‌ها بود. جامعه موردبررسی عبارت از کلیه ۳۹۴۲ داوطلب کنکور در رشته‌های آموزش زبان انگلیسی، زبان‌شناسی، مترجمی و ادبیات انگلیسی در سال ۱۳۹۱ بود و اطلاعات مربوط به ۲۷۵۴ آزمودنی به‌عنوان گروه نمونه تحلیل شد. از تحلیل محتوای آزمون، کدگذاری آن و بررسی گزارش‌های کلامی برای تعیین مهارت‌های زیربنایی احتمالی هر یک از سؤالات استفاده شد. نتیجه

۱. دکترا سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران
۲. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران (نویسنده مسئول)
falsafinejad@yahoo.com
۳. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران
٤. استادیار گروه زبان و ادبیات انگلیسی، دانشگاه علامه طباطبائی، تهران

بررسی‌های بخش کیفی شش مهارت تدوین‌شده شامل استفاده از دانش واژگان، استفاده از دانش نحوی، استخراج اطلاعات صریح، استنتاج، اتصال و ادغام و استفاده از دانش عملی بود. تحلیل داده‌ها با استفاده از مدل غیر جبرانی فیوژن کاهش‌یافته مبتنی بر الگوریتم زنجیره مارکف مونته کارلو، نشان‌دهنده کفایت مدل تدوین‌شده و امکان استفاده از آن در داده‌های زبان انگلیسی بود. از بین مهارت‌های درک مطلب اتصال و ادغام، استفاده از دانش عملی، استخراج اطلاعات صریح، استنتاج، استفاده از دانش نحوی و استفاده از دانش واژگان به ترتیب آسان‌ترین تا دشوارترین مهارت‌ها بود. دقت و اعتبار طبقه‌بندی آزمون‌شوندگان در هر یک از مهارت‌های درک مطلب نیز مناسب ارزیابی شد. نتایج این مطالعه نشان داد که استفاده از مدل‌های تشخیصی شناختی، اطلاعات بیش‌تری را ارائه می‌دهد.

واژگان کلیدی: مدل‌های تشخیصی شناختی، مدل غیر جبرانی فیوژن، مهارت‌های سازه درک مطلب، زبان انگلیسی

مقدمه

آزمون‌های سرنوشت‌ساز ابزارهای مهمی برای ارزیابی پیشرفت فراگیران و نظارت بر کیفیت نظام‌های آموزشی هستند. بااین‌وجود، تقریباً در همه آن‌ها، با استفاده از رویکردهای مرسوم اندازه‌گیری مثل نظریه کلاسیک و مدل‌های نظریه سؤال پاسخ فقط اطلاعاتی کلی در قالب نمره و یا رتبه درصدی تهیه می‌شود و اطلاعات خاصی در مورد نقاط قوت و ضعف آزمودنی‌ها ارائه نمی‌شود تا به‌وسیله آن بتوان به یادگیرندگان در بررسی دلایل عدم موفقیت خود در آزمون و بهبود مهارت‌های شناختی خود و به معلمان در طراحی مداخلات آموزشی کمک کرد (لیتن و گیرل[1]، ۲۰۰۷ و لی[2]، ۲۰۱۱).

اگرچه رویکردهای سنتی اندازه‌گیری خدمت شایان توجهی در اندازه‌گیری سازه‌های علوم رفتاری داشته‌اند ولی این رویکردها در آزمون‌های چندبعدی و معمولاً دارای ساختار پیچیده که آزمودنی برای پاسخ‌گویی به سؤالات آن به مهارت‌های[3] مختلف نیاز دارد دچار

1. Leighton & Gierl
2. Li
۳. اسامی مختلفی مثل ویژگی، خصیصه، عناصر فرایند، صفت، شایستگی، مهارت و خرده مهارت برای متغیرهای پنهان پیشنهاد شده است.

دچار چالش‌اند، زیرا ارائه نمره کل حاصل و رتبه درصدی آزمودنی‌ها در دروس مختلف قابلیت بازنمایی واقعی توانایی افراد را با توجه به مهارت‌های مختلف موردنیاز ندارند. حصول این مهم مستلزم استفاده از رویکردهای جدیدتری است که مهارت‌های لازم برای ارائه پاسخ صحیح به سؤالات را بررسی کند و اطلاعات تشخیصی بیش‌تری به دست دهد (جانگ[۱]، ۲۰۰۵)

در این راستا استفاده از مدل‌های تشخیصی شناختی[۲] یا مدل‌های روان‌سنجی شناختی[۳]، مدل‌های تشخیص شناختی[۴]، مدل‌های پاسخ پنهان[۵]، مدل‌های طبقه پنهان مقید[۶]، مدل‌های طبقه پنهان طبقه‌بندی چندگانه[۷]، مدل‌های طبقه پنهان معین ساختارمند[۸] و مدل‌های نظریه سوال-پاسخ ساختاری[۹] پیشنهاد شده است (راپ، تمپلین و هنسن[۱۰]، ۲۰۱۰). هدف اصلی این مدل‌ها پیوند روانشناسی شناختی و اندازه‌گیری آموزشی برای شناسایی فرایندها، راهبردها و ساختارهای دانشی است که زیربنای ارائه پاسخ به سؤال‌های یک آزمون بوده و درنهایت منجر به شناسایی نقاط قوت و ضعف آزمودنی‌ها و طبقه‌بندی آن‌ها به مسلط و غیر مسلط بر اساس متغیرهای پنهان طبقه‌ای چندگانه (مهارت) می‌شود (هنسن، روسس، داگلاس و هی[۱۱]، ۲۰۰۸).

ضرورت سنجش تشخیصی شناختی در سال ۱۹۸۹ در دو فصل از کتاب اندازه‌گیری آموزشی روبرت لین[۱۲]، فصل روایی ساموئل مسیک[۱۳] و فصل دلالت‌های روانشناسی

1. Jang
2. cognitive diagnostic models
3. cognitive psychometric models
4. cognitive diagnosis models
5. latent response models
6. restricted latent class models
7. multiple classification latent class models
8. structured located latent class models
9. structured item response models
10. Rupp, Templin & Henson
11. Henson, Roussos, Douglas, & He
12. Linn
13. Messick

شناختی در اندازه‌گیری آموزشی اسنو و لهمن[1]، مطرح و بعدها توسط محققین دیگر بسط داده شد (لیتن و گیرل، ۲۰۰۷). تدوین این مدل‌ها ابتدا توسط فیشر[2] (۱۹۷۳) با استفاده از مدل آزمون لجستیک خطی[3] برای ارائه مدلی آماری به‌منظور سنجش تشخیصی شناختی آغاز شد. ادبیات مربوط به این مدل‌ها در دهه ۱۹۸۰ توسعه یافت تا این‌که در سال ۱۹۸۳، مدل فضای قاعده[4] توسط تاتسوکا[5] (۱۹۸۳) تدوین شد و از زمانی که اسنو و لهمن (۱۹۸۹) فصلی در زمینه رابطه متقابل روانشناسی شناختی با روان‌سنجی نوشتند، این حیطه به‌سرعت گسترش یافت.

این مدل‌های روان‌سنجی بر اساس سه ویژگی مقیاس اندازه‌گیری متغیرهای پاسخ مشاهده‌شده آزمونی‌ها (دو ارزشی در برابر چند ارزشی)، مقیاس اندازه‌گیری متغیرهای پنهان (دو ارزشی در برابر چند ارزشی) و شیوه ترکیب متغیرهای پنهان (روش جبرانی یا غیر عطفی[6] و غیر جبرانی یا عطفی[7]) به چند دسته تقسیم می‌شوند. در مدل‌های جبرانی، تسلط آزمودنی در یکی از مهارت‌های موردنیاز برای ارائه پاسخ درست به یک سؤال، عدم تسلط او را در مهارت‌های دیگر جبران می‌کند ولی در مدل‌های غیر جبرانی، عدم تسلط در یک مهارت، تسلط در مهارت دیگر را جبران نمی‌کند. ازجمله مدل‌های معروف غیر جبرانی می‌توان به مدل‌های فضای قاعده، روش سلسله مراتبی مهارت[8]، دینا[9]، نیدا[10] و مدل یکپارچه پارامتر بندی شده مجدد یا فیوژن[11] اشاره کرد و مدل‌های دینو[12]، نیدو[13] و پارامتر

1. Snow & Lohman
2. Fischer
3. Linear logistic test model (LLTM)
4. rule space model
5. Tatsuoka
6. compensatory or disjunctive models
7. non-compensatory or conjunctive models
8. Attribute hierarchy method (AHM)
9. DINA
10. NIDA
11. non-compensatory reparameterized unified or fusion model (NC-RUM)
12. DINO
13. NIDO

بندی مجدد[1] نیز ازجمله مدل‌های مرسوم جبرانی محسوب می‌شوند. علاوه بر این بعضی از مدل‌های تشخیصی شناختی مانند مدل تشخیصی شناختی لگ خطی[2] به‌صورت کلی هستند و می‌توان آن‌ها را هم به‌صورت جبرانی و هم غیر جبرانی در نظر گرفت (راپ و همکاران، ۲۰۱۰).

علی‌رغم چالش‌های عمده‌ای که در این مدل‌ها وجود دارد (مثل حجم نمونه، همبستگی بین مهارت‌ها و میزان دشواری مهارت‌ها (کیم، ۲۰۱۱) و مشکلات مربوط به تدوین ماتریس کیو روا (جانگ، ۲۰۰۵)) ولی ازآنجایی‌که امید است این مدل‌ها نتایج مفیدی‌تری نسبت به دیگر مدل‌های مرسوم اندازه‌گیری ارائه دهند، موردعلاقه بسیاری از متخصصان حوزه اندازه‌گیری به‌صورت ریتروفیت[3] آزمون‌های موجود (مثل لی، ۲۰۱۲ و تمپلین و هافمن، ۲۰۱۳)، تحلیل پرسشنامه‌های روانشناسی (مثل تمپلین و هنسن، ۲۰۰۶)، ساخت آزمون (مثل برادشاو، ایزساک، تمپلین و جاکبسن[4]، ۲۰۱۳) و ساخت پرسشنامه (کیم و کیم[5]، ۲۰۱۳) قرار گرفته است.

این مطالعه در صدد معرفی مدل‌های تشخیصی شناختی، امکان‌سنجی به‌کارگیری این مدل‌ها به‌منظور جمع‌آوری شواهد تجربی مبتنی بر کفایت و برتری آن‌ها نسبت به سایر مدل‌های سنتی موجود در سنجش توانایی و پیشرفت تحصیلی فراگیران در نظام آموزش عالی است. یکی از بهترین زمینه‌های استفاده از این مدل‌ها، داده‌های آزمون‌های سرنوشت‌ساز است که در این راستا داده‌های ورود به دوره‌های دکترا و حیطه درک مطلب در زبان انگلیسی عمومی بستر مناسبی برای اهداف این مطالعه محسوب می‌شود؛ چراکه درک مطلب ازنظر بسیاری از متخصصان زبان انگلیسی (جنگ، ۲۰۰۵؛ لی، ۲۰۱۲؛ گااو و روگر، ۲۰۱۰ و سوتینا، گرین و تاتسوکا، ۲۰۱۱) سازه‌ای چندبعدی با ساختار پیچیده است. بدین منظور ضمن شناسایی مهارت‌های احتمالی آزمون، نقاط قوت و ضعف آزمودنی‌ها

1. Compensatory reparameterized unified model (C-RUM)
2. log-linear cognitive diagnosis model (LCDM)
3. retrofitting
4. Bradshaw, Izsak, Templin, & Jacobson
5. Kim & Kim

مشخص و توان تشخیصی آزمون بررسی می‌شود. این تحقیق می‌تواند گامی مقدماتی در جهت ساخت آزمون‌های تکوینی مبتنی بر مهارت در درک مطلب زبان انگلیسی محسوب شود.

روش‌شناسی

این مطالعه ازنظر هدف جزو تحقیقات کاربردی و ازنظر روش جز تحلیل‌های ثانویه در چارچوب مدل‌سازی تشخیصی شناختی (رویکرد ریتروفیت) محسوب می‌شود. در این چارچوب مبتنی بر نظر باک و همکاران (۱۹۹۸) ابتدا لیست اولیه‌ای از مهارت‌های زیربنایی آزمون تهیه شد، سپس هر سؤال بر اساس مهارت (های) موردنیاز برای ارائه پاسخ درست کدگذاری شد و بعدازآن داده‌ها با استفاده از یک مدل آماری مناسب با ماتریس کیو اولیه تدوین‌شده از گام قبل تحلیل شد و درنهایت ماتریس کیو اولیه بر اساس نتایج آماری به‌دست‌آمده در چارچوب اهمیت مهارت بر اساس نظر متخصص تعدیل شد. گام سوم و چهارم تا زمان رسیدن به ماتریس کیو مناسب تکرار شد. درنهایت پارامترهای توانایی و سؤال به‌منظور مشخص کردن نقاط قوت و ضعف آزمودنی‌ها و بررسی توان تشخیصی آزمون نیز بررسی می‌شود.

در بخش تدوین ماتریس کیو این مطالعه ابتدا به‌منظور شناسایی مهارت‌های مربوط به درک مطلب پیشینه پژوهش در زمینه این سازه (مثل گِربه[1]، ۱۹۹۱؛ الدرسن[2]، ۲۰۰۰؛ جنگ، ۲۰۰۵؛ لی، ۲۰۱۱؛ گااو و روگر[3]، ۲۰۱۰؛ سوتینا، گرین[4] و تاتسوکا، ۲۰۱۱) مطالعه شد. سپس تمامی سؤالات توسط دو متخصص زبان انگلیسی تحلیل محتوا و طرح کدگذاری تعریف شد. پس‌ازآن از سه متخصص درخواست شد تا در چارچوب مهارت‌های تعیین‌شده، ماتریس کیو مربوط به آزمون را ارائه دهند. برای بررسی پایایی

1. Grabe
2. Alderson
3. Gao & Roger
4. Svetina, Gorin

توافق بین سه متخصص در مورد مهارت‌های زیربنایی هر یک از سؤالات تعیین با استفاده از ضریب کاپای محاسبه شد.

جدول ۱. طرح کدگذاری سؤال‌های بخش درک مطلب (مهارت‌ها و تعاریف آن‌ها)

مهارت‌ها	معادل فارسی	تعریف
Using Vocabulary Knowledge	استفاده از دانش واژگان	Understand academic texts with infrequently used vocabulary and specialized vocabulary
Using Syntactic Knowledge	استفاده از دانش نحوی	Understand the relationship of ideas within the sentence using knowledge of syntax, grammar, punctuation, or parts of speech
Extracting Explicit Information or Scan	استخراج اطلاعات صریح یا بررسی اجمالی	Locate the specific information requested in the question; scan the text for specific details
Drawing Inference	استنتاج	Ask about information not directly stated in the passage, but are understood by drawing conclusions from the information given in the passage
Connecting and Synthesizing	اتصال و ادغام	Integrate, relate, or summarize the information presented in different sentences or parts of the text to generate meaning
Using Pragmatic Knowledge	استفاده از دانش عملی	Understand pragmatic and rhetorical purposes of the text creator

(*** تعریف مهارت‌ها از گااو و روگر (۲۰۱۰) و لی (۲۰۱۱) گرفته شده است.)

(*** تعریف مهارت‌ها از گااو و روگر (۲۰۱۰) و لی (۲۰۱۱) گرفته شده است.)

علاوه بر این با استفاده از روش تفکر با صدای بلند از پنج دانشجوی مقطع فوق‌لیسانس و دکترا رشته زبان انگلیسی درخواست شد تا به سؤالات پاسخ داده و توضیح دهند که در ارائه پاسخ به سؤالات از چه مهارت‌هایی استفاده کردند. سپس بر اساس تحلیل محتوای آزمون، نظر متخصصین، روش تفکر با صدای بلند، مهارت‌های احتمالی زیربنایی برای پاسخ از هر یک از سؤالات مشخص و ماتریس کیو اولیه تدوین شد. تعداد مهارت‌های اولیه تدوین‌شده توسط متخصصین، بیش‌تر بود ولی به دلیل این‌که در بعضی از مهارت‌ها کم‌تر از ۳ سؤال وجود داشت و به این دلیل که بسیاری از مهارت‌ها دارای هم‌پوشانی

زیادی بودند و به‌طور مشخصی از هم قابل‌تفکیک نبودند (مثل مهارت شناسایی ایده اصلی نویسنده و استنتاج) لذا مانند دیگر مطالعات انجام‌شده در زمینه کاربرد مدل‌های تشخیصی شناختی در درک مطلب (مثل باک و تاتسوکا، ۱۹۹۸ و لی، ۲۰۱۱) بعضی از آن‌ها در هم ادغام شد تا توافق نهایی متخصصین حاصل شود. علاوه بر این در نظر گرفتن مهارت‌های بیش‌تر در رویکرد ریتروفیت آزمون‌های اجراشده ممکن است منجر به توزیع نامتعادل سؤال‌ها برای مهارت‌ها شود و این امر مانعی در تدوین ماتریس کیو معقول می‌شود. به صورتی که در بعضی از مهارت‌ها تعداد زیادی سؤال و در برخی دیگر سؤالات خیلی کمی (کم‌تر از ۳ سؤال) وجود خواهد داشت (جنگ، ۲۰۰۹).

نمونه سؤال آزمون درک مطلب:

56- Based on the information in the passage, Wilson's letters can best described as----------.

1) witty 2) cynical 3) preachy 4) spontaneous

مهارت‌های مربوط به این سؤال استفاده از دانش واژگان، استفاده از دانش نحوی، و استخراج اطلاعات صریح یا بررسی اجمالی در نظر گرفته شد.

در بخش کمی این مطالعه نیز از بین مدل‌های غیر جبرانی که در حیطه درک مطلب در زبان انگلیسی مورداستفاده قرارگرفته (مثل جنگ، ۲۰۰۵، لی، ۲۰۱۱ و لی و ساواکی، ۲۰۰۹) از مدل فیوژن کمک گرفته شد. این مدل توسط هارتز (۲۰۰۲) که مدل یکپارچه دیبلو، استوت و روسس (۱۹۹۵) را دوباره پارامتر بندی کرد، پیشنهاد شده است. در مدل فیوژن احتمال ارائه پاسخ درست توسط آزمودنی n به سؤال i عبارت است از:

$$p(X_{ni} = 1 \mid \propto \alpha_n, \eta_n;\ \pi^*, r_{ik}^*, c_i) = \pi^* \prod_{k=1}^{k} r_{ik}^{*(1-\alpha_{nk})} P_{ci}(\eta_n).$$

در این مدل ویژگی‌های آزمودنی شامل تسلط یا عدم تسلط آزمودنی n در مهارت‌های ارائه‌شده توسط بردار نیمرخ مهارت ($\underline{\alpha} = (\alpha_{n1}, \dots, \alpha_{nk})$) و پارامتر آزمودنی ($\eta_n$)

می‌شود. علاوه بر این ویژگی‌های سؤال با استفاده از سه پارامتر سؤال که در زیر تعریف‌شده، ارائه می‌شود (دی‌بلو و استوت، ۲۰۰۸؛ جانگ، ۲۰۰۵ و ۲۰۰۹؛ لی، ۲۰۱۱، تمپلین و همکاران، ۲۰۱۰ و روسس و همکاران، ۲۰۰۷).

پارامتر سؤال π_i^* خط پایه[1] عبارت است از احتمال ارائه پاسخ درست به یک سؤال با این فرض که آزمودنی در همه مهارت‌های موردنیاز به تسلط رسیده‌اند. پارامتر سؤال r_{ik}^* یا جریمه[2] به رابطه سوال-مهارت مربوط می‌شود. پارامتر جریمه، یک عامل ضربی[3] است که در اثر آن اگر فرد در مهارت k در سؤال i به تسلط نرسیده باشد احتمال پاسخ درست کاهش می‌یابد. پارامتر سؤال c_i یا تعامل پنهان یا عبارت کامل[4] به عبارت η_n (پارامتر آزمودنی) مربوط می‌شود. η_n بیانگر همه مهارت‌ها، دانش‌ها و توانایی‌هایی است که در مجموعه مهارت‌های تدوین‌شده در ماتریس کیو ارائه نشده است. $P_{ci}(\eta_j)$ مربوط به مدل راش با پارامتر دشواری منفی c_i ($-c_i$) است. اگر c_i زیاد باشد، $P_{ci}(\eta_j)$ نیز بزرگ خواهد بود.

جامعه این پژوهش را کلیه داوطلبان شرکت‌کننده در آزمون ورودی دوره دکترا داخل در رشته زبان‌های خارجی در سال ۱۳۹۱ تشکیل می‌دهد. تعداد کل این شرکت‌کننده‌ها، ۱۵۰۳۳۳ داوطلب است که ۳۹۴۲ داوطلب مربوط به گروه زبان‌های خارجی هستند. طبق آمار منتشرشده توسط سازمان سنجش از این تعداد ۲۰۱۳ داوطلب مربوط به رشته آموزش زبان انگلیسی (۱۰۹۷ زن و ۹۱۶ مرد)، ۵۴۱ داوطلب مربوط به رشته زبان و ادبیات انگلیسی (۲۹۹ زن و ۲۴۲ مرد)، ۴۶۷ داوطلب مربوط به رشته ترجمه (۲۴۷ زن و ۱۹۳ مرد)، و ۹۲۱ داوطلب مربوط به رشته زبان‌شناسی (۵۴۹ زن و ۳۷۲ مرد) می‌باشند (دفتر طرح و آمار سازمان سنجش، ۱۳۹۲). به دلیل وجود تعداد زیاد آزمودنی‌ها با داده گمشده و با این استدلال که ممکن است در آزمودنی‌های سرنوشت‌ساز به دلیل زمان کم آزمون یا حدس زدن، فرایند موردنیاز برای حل مسئله را طی نکرده‌اند (لی، ۲۰۱۱) آزمودنی‌هایی که به

1. baseline
2. penalty
3. multiplicative factor
4. latent interaction or completeness term

بیش از ۳۰ درصد سؤالات (حداقل ۱۳ سؤال از ۴۵ سؤال) پاسخ نداده بودند، حذف شدند. ضرورت حذف این آزمودنی‌ها به دلیل قرار گرفتن سؤال‌های بخش درک مطلب در پایان آزمون، بیش‌تر حس شد.

برای بخش تدوین ماتریس کیو روش نمونه‌گیری قضاوتی به کار گرفته شد. علاوه بر استفاده از دو متخصص به‌منظور تحلیل محتوای آزمون و سه کدگذار برای اختصاص هر یک از سؤالات به مهارت‌های تدوین‌شده که هر پنج نفر در زمینه آموزش زبان انگلیسی متخصص بوده و دارای سابقه تدریس حداقل ده سال بودند. به‌منظور تدوین ماتریس کیو روا و تأکید بر استفاده از منابع چندگانه در تدوین آن گزارش کلامی‌های پنج آزمودنی استفاده شد.

ابزار مورداستفاده در این مطالعه، سؤالات آزمون ورودی دوره دکترا داخل در سال ۱۳۹۱ است. این آزمون دارای ۱۰۰ سؤال است که ۲۰ سؤال آن مربوط به دستور زبان، ۳۵ سؤال مربوط به واژگان و ۴۵ سؤال آن مربوط به بخش درک مطلب است. در این مطالعه فقط ۴۵ سؤال مربوط به بخش درک مطلب بررسی شد. در سنجش تشخیصی شناختی، برای بررسی روایی و اعتبار، روش‌های مرسوم مورداستفاده مستقیماً قابل کاربرد نیست زیرا در آن‌ها فرض می‌شود که یک صفت پنهان تک‌بعدی بررسی می‌شود. در سنجش تشخیصی که از مدل‌های طبقه پنهان استفاده می‌شود (دی‌بلو و همکاران، ۲۰۰۷، تمپلین و برادشاو[1]، ۲۰۱۳) فرمول‌های جدید برای اندازه‌گیری اعتبار تدوین‌شده است. در این مدل‌ها روایی ماتریس کیو و پایایی طبقه‌بندی موردتوجه قرار می‌گیرد. ازآنجایی‌که روش‌های کمی برای بررسی روایی ماتریس کیو فقط روش دلتا برای مدل غیر جبرانی دینا (دی‌لتره، ۲۰۰۸) تدوین‌شده است لذا برای بررسی روایی ماتریس کیو در این مطالعه با استفاده منابع چندگانه تلاش شد ماتریس کیو روایی تدوین شود. به‌منظور بررسی اعتبار درون‌درجه‌بندی کننده[2] نیز از آماره کاپای فلیس[3] (فلیس، ۱۹۷۱) در بسته irr نرم‌افزار R استفاده شد. دلیل

1. Bradshaw
2. inter-rater reliability
3. Fleiss' Kappa statistic

استفاده از این آماره وجود بیش از دو کدگذار بود. مقادیر کم‌تر از ۰/۲، توافق کم، بین ۰/۲۱ تا ۰/۴ دارای توافق مناسب، بین ۰/۴۱ تا ۰/۶۰ توافق متوسط، بین ۰/۶۱ تا ۰/۸، توافق زیاد و بین ۰/۸۱ تا ۱ توافق تقریباً کامل تفسیر می‌شود (لندیس و کچ[1]، ۱۹۷۷). در این مطالعه، میزان توافق بین کدگذاران مناسب ارزیابی شد. در آزمون درک مطلب مهارت استفاده از دانش واژگان (۰/۸۹) دارای بیش‌ترین توافق بود و کم‌ترین توافق در مهارت استفاده از دانش عملی (۰/۵۹) به دست آمد. اگرچه در سنجش تشخیصی شناختی روش‌های مرسوم بررسی روایی و اعتبار مستقیماً استفاده نمی‌شود و در عوض روایی ماتریس کیو و دقت و اعتبار طبقه‌بندی آزمون‌شوندگان در هر یک از مهارت‌ها حائز اهمیت است ولی بررسی روایی و اعتبار آزمون به شیوه‌های مرسوم نیز می‌تواند اطلاعاتی در زمینه کیفیت سؤالات تدارک ببیند. نتایج مربوط به بررسی روایی و اعتبار آزمون به شکل مرسوم بررسی شد. ازآنجایی‌که این ابزار توسط متخصصین دانشگاهی و آزمون سازان باتجربه سازمان سنجش کل کشور ساخته شده است دارای روایی محتوایی مناسبی است. اعتبار آزمون نیز به‌وسیله آلفای کرونباخ ۰/۷۷ با خطای استاندارد اندازه‌گیری ۳/۱۰ به دست آمد. نتایج مربوط به بررسی روایی ماتریس کیو با استفاده از روش‌های کیفی و اعتبار طبقه‌بندی آزمون‌شوندگان با استفاده از شبیه‌سازی داده‌ها نیز در بخش یافته‌ها خواهد آمد.

در این تحقیق متغیرهای پنهان (مهارت‌ها) و متغیرهای مشاهده‌شده (سؤال‌ها)، دو ارزشی در نظر گرفته شد. هم‌چنین، با فرض رابطه غیر جبرانی بین مهارت‌ها در سازه درک مطلب از مدل غیر جبرانی یکپارچه پارامتر بندی مجدد کاهش‌یافته یا مدل فیوژن با کمک سیستم آرپجیو[2] که دارای رویکردی بیزی و مبتنی بر زنجیره‌های مارکف مونته کارلو است برای تحلیل داده‌ها استفاده شد. در این مطالعه دو زنجیره با ۳۰۰۰۰ تکرار که ۱۳۰۰۰ تکرار آن در مرحله داغیدن[3] کنار گذاشته شد به کار گرفته شد.

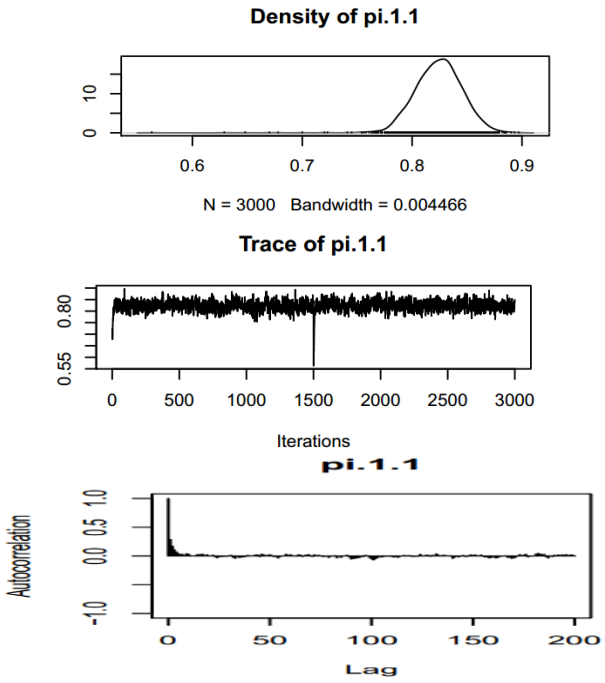
1. Landis & Koch
2. Arpeggio
3. Burn-in phase

یافته‌های تحقیق

بررسی همگرایی زنجیره‌های مارکف مونته کارلو. مانند همه مدل‌های روان‌سنجی که از رویکرد بیز استفاده می‌کند اولین نکته‌ای که پس از اجرای مدل باید مدنظر قرار گیرد، همگرایی زنجیره‌ها به یک راه‌حل مانا[1] (زنجیره مارکف مونته کارلو به سطح دشواری واحد در هر مهارت برسد) است. در مدل فیوژن به‌منظور بررسی همگرایی مدل از چهار روش استفاده می‌شود: نمودارهای زنجیره‌ای (نمودار سری زمانی[2])، توزیع پسین برآورد شده[3] (نمودار چگالی[4])، نمودار خودهمبستگی[5] برآوردهای زنجیره و $\hat{R}$ گلمن و روبین[6] (روسس و همکاران، ۲۰۰۷).

بررسی چشمی نمودارها نشان داد که همه پارامترها دارای همگرایی عالی بودند. بررسی نمودار زنجیره برای هر پارامتر در هر گام زنجیره نشان داد که زنجیره‌ها به توزیع ثابتی رسیده است. نمودارهای زنجیره سری‌های زمانی نوساناتی را نشان نمی‌دهد که نشان‌دهنده همگرایی است و نمودارهای چگالی نیز تک نمایی[7] هستند و خودهمبستگی کمی بین برآوردهای زنجیره وجود دارد. به‌عنوان نمونه یکی از نمودارهای ایجاد شد توسط بسته Coda در نرم‌افزار R در زیر ارائه می‌شود. مقادیر $\hat{R}$ گلمن و روبین در همه پارامترها کم‌تر از ۱/۲ بود که به معنای همگرایی است.

1. Stationary solution
2. Time series
3. Estimated posterior distribution
4. density
5. Auto-correlation
6. Gelman & Robin
7. unimodal



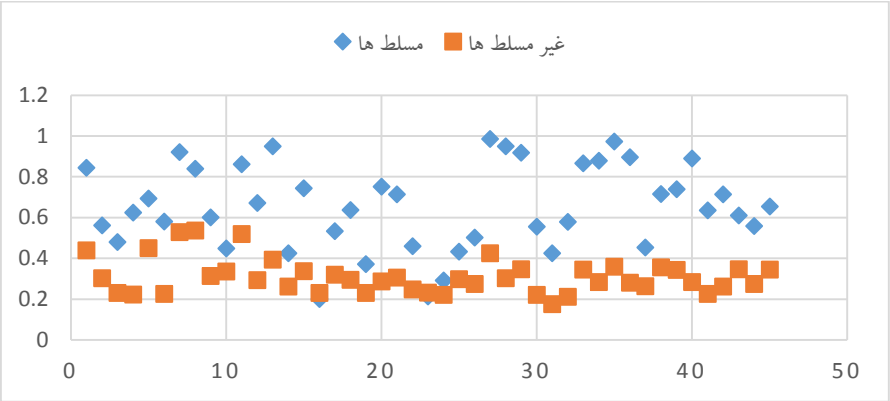
شکل ۱. نمودارهای چگالی، سری‌های زمانی و خودهمبستگی یکی از پارامترهای مدل

بررسی برازش مدل. برای بررسی برازش مدل با داده‌ها از روش‌های مختلفی آماره تسلط در سؤال[1] (روایی درونی[2])، مقایسه مقادیر مشاهده‌شده و پیش‌بینی‌شده و مقایسه احتمال‌ها تراکمی مشاهده‌شده و برآورد شده به شرح ذیل استفاده شد.

آماره تسلط در سؤال (آماره برازش سؤال). تفاوت عملکرد آزمودنی‌هایی مسلط و غیر مسلط (آماره تسلط در سؤال) به‌منظور بررسی ظرفیت تشخیصی سؤالات آزمون و برازش مدل در نمودار ۲ ارائه شده است. هر چه فاصله بین نقاط خطوط برای یک سؤال بیش‌تر باشد نشان‌دهنده روایی درونی سؤال برای تمیز بین مسلط‌ها از غیر مسلط‌ها است. در برخی از سؤالات، آزمودنی‌های مسلط در مهارت‌های موردنیاز برای ارائه پاسخ به سؤالات دارای عملکرد بهتری نسبت به افراد غیر مسلط بوده‌اند و در برخی سؤال‌ها تفاوتی بین این دو

1. ImStats
2. Internal validity

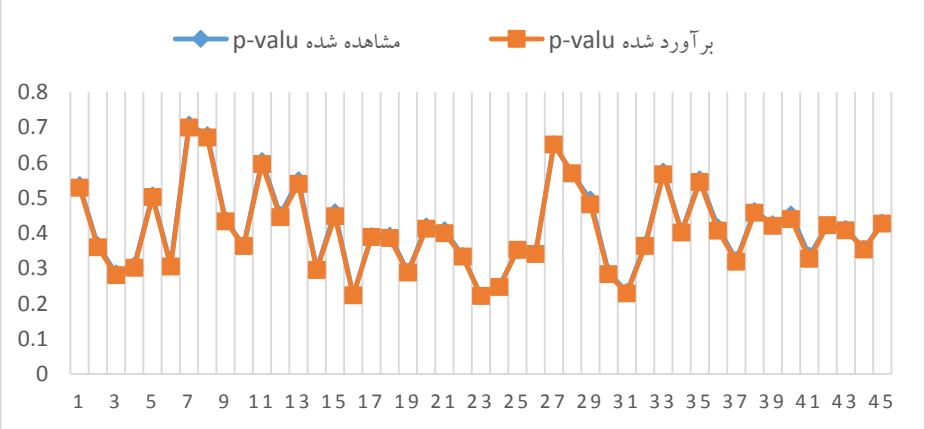
گروه تفاوتی مشاهده نمی‌شود. مثلاً سؤال ۱۶ و ۲۳ قادر به تمایز افراد مسلط و غیر مسلط نیست و لذا دارای ظرفیت تشخیصی مناسبی نیست. تفاوت بین میانگین نمره‌های افراد مسلط (۶۵/.) و غیر مسلط (۳۱/.) در کل سؤالات ۳۴/. به دست آمد. آماره تسلط در سؤال به‌عنوان شاهد روایی درونی مورداستفاده قرار می‌گیرد، زیرا از خود داده‌های آزمون برای کمک به بررسی اصالت مدل استفاده می‌کند.



شکل ۲. مقایسه عملکرد افراد مسلط و غیر مسلط در سؤالات آزمون درک مطلب

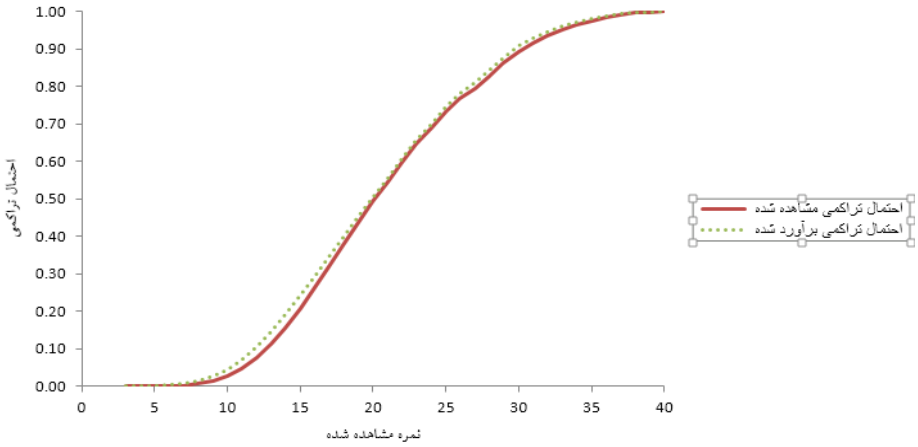
مقایسه مقادیر دشواری مشاهده‌شده و پیش‌بینی‌شده در سؤالات. نمودار ۳ مقادیر p (نسبت آزمودنی‌هایی که به سؤال پاسخ درست داده‌اند) مشاهده‌شده و پیش‌بینی‌شده را مقایسه می‌کند. تقریباً در تمامی سؤال‌ها دو خط به هم نزدیک هستند. میانگین قدر مطلق تفاوت[1] بین مقادیر مشاهده‌شده و پیش‌بینی‌شده ۰۰۴۸/. به است آمد که این اختلاف کم نشان‌دهنده برازش خوب مدل است.

1. Mean absolute difference (MAD)



شکل ۳. مقادیر دشواری مشاهده‌شده و پیش‌بینی‌شده سؤالات

مقایسه توزیع احتمال تراکمی نمرات کل مشاهده‌شده و برآورد شده. همان‌طور که در نمودار ۴ قابل‌مشاهده است بین توزیع احتمال تراکمی نمرات کل مشاهده‌شده و برآورد شده تفاوت چندانی وجود ندارد که این حاکی از برازش مدل با داده‌هاست. همان‌طور که قابل‌ملاحظه است فقط بین نمرات در کران پایین تا حدودی بیش برآورد شده است ولی این تفاوت بسیار ناچیز است.



شکل ٤. مقایسه توزیع احتمال تراکمی مشاهده‌شده و برآورد شده

پارامترهای سؤال. در مدل فیوژن، پارامترها (π_i^*، r_{ik}^* و c_i) هر سؤال ارائه می‌شود. در اجرای اول پارامترهای c_i تعداد زیادی از سؤالات بیش‌تر از ۲ بود لذا در اجرای دوم، مقادیر آن‌ها ثابت فرض شد. در اجرای دوم نیز مقادیر r_{ik}^* ۱۶ پارامتر نسبتاً بزرگ بود (بیش از ۹/.) که در اجرای سوم درایه‌های ماتریس کیو تمامی این موارد، از یک به صفر تبدیل شد. درواقع در اجرای سوم، تحلیل دوباره با ماتریس کیو اصلاح‌شده مبتنی بر اجرای اول و دوم انجام شد. مقادیر هر یک از پارامترها به تفکیک مهارت‌ها و نمره نسبت درست هر یک از سؤالات در جدول ۳ ارائه شده است.

جدول ۲. مقادیر پارامترهای سؤال به تفکیک مهارت‌ها در هر یک از سؤالات

سؤال	p_i^*	r_1^*	r_2^*	r_3^*	r_4^*	r_5^*	r_6^*	نمره نسبت درست
۱	/۸۲	/۷۱	/۸۷	/۸۱				/۵۹
۲	/۵۳		/۶۳		/۸۹			/۳۸
۳	/۴۶	/۵۴	/۹۰					/۳۳
۴	/۶۲	/۵۴	/۶۹		/۸۷			/۳۲
۵	/۶۸	/۷۴			/۸۸			/۵۳
۶	/۵۷	/۵۹	/۷۶	/۸۲				/۳۳
۷	/۸۹			/۶۳				/۷۶
۸	/۸۲			/۶۹				/۷۲
۹	/۵۷					/۵۸		/۴۶
۱۰	/۴۴		/۸۶		/۸۷			/۴۰
۱۱	/۸۳		/۷۱		/۸۷			/۶۵
۱۲	/۶۱			/۵۳				/۵۰
۱۳	/۹۳	/۶۶		/۵۸				/۶۰
۱۴	/۴۱		/۶۹		/۸۹			/۳۱
۱۵	/۶۸		/۵۲					/۵۴
۱۶	/۲۳				/۹۶			/۲۳
۱۷	/۵۱			/۶۷		/۸۸		/۴۴
۱۸	/۵۸				/۵۲			/۴۵
۱۹	/۳۵			/۷۰				/۳۰

۲۰	./۶۹			./۴۸		./۸۴		./۴۸
۲۱	./۶۶	./۴۸						./۴۴
۲۲	./۴۴				./۵۸			./۳۸
۲۳	./۲۳			./۹۶				./۲۴
۲۴	./۲۸			./۷۸				./۲۶
۲۵	./۴۳		./۸۲			./۸۳		./۴۰
۲۶	./۴۷			./۶۰				./۳۹
۲۷	./۹۷		./۷۰			./۵۹		./۷۱
۲۸	./۹۲			./۹۰		./۳۲		./۶۳
۲۹	./۸۵	./۴۲						./۵۱
۳۰	./۵۲	./۵۴	./۸۶	./۸۲				./۲۹
۳۱	./۳۹	./۴۹		./۸۸				./۲۴
۳۲	./۵۴				./۴۴			./۴۱
۳۳	./۸۳				./۴۵			./۶۲
۳۴	./۸۵	./۷۶	./۸۳	./۶۷		./۵۸		./۴۵
۳۵	./۹۴			./۷۶	./۴۲			./۳۲
۳۶	./۸۷	./۶۲		./۴۹				./۴۸
۳۷	./۴۴			./۶۱				./۴۶
۳۸	./۶۹			./۶۴	./۷۹			./۴۹
۳۹	./۶۹	./۵۷		./۸۷		./۸۶		./۳۷
۴۰	./۸۳		./۴۴		./۷۶			./۴۸
۴۱	./۵۷		./۴۵	./۸۸				./۳۷
۴۲	./۶۸			./۶۱		./۵۵		./۴۸
۴۳	./۵۹	./۷۷			./۶۸			./۳۲
۴۴	./۵۴			./۶۷	./۷۱			./۳۹
۴۵	./۶۴			./۷۹	./۵۸			./۴۶
میانگین	./۶۳	./۶۰	./۷۲	./۷۵	./۶۵	./۶۵		./۴۲

(***خانه‌های سیاه شده، خانه‌های سیاه شده، خانه‌هایی هستند که در اجرای دوم پارامتر جریمه آن‌ها بیش از ۹/. بوده و در اجرای سوم درایه‌های ماتریس کیو از یک به صفر تغییر داده شد.)

در چارچوب مدل فیوژن حالت ایده‌آل ارائه مدلی است که منتج به پارامترهایی با مقادیر π_i^* بالا (بیش از ۶/.)، r_{ik}^* پایین (بین ۵/.-۰)، که در این حالت آزمون دارای ساختار شناختی بالا[1] در نظر گرفته می‌شود، و c_i کم (۱.۵-۰) است (روسس و همکاران، ۲۰۰۷). نزدیک بودن مقادیر پارامتر π_i^* به ۱ نشان‌دهنده مقاوم بودن مدل‌سازی تشخیص مهارت‌ها[2] است و پایین بودن مقادیر پارامتر r_{ik}^* نشان‌دهنده قدرت تشخیصی بالا در تمایز افراد مسلط و غیر مسلط است (کیم، ۲۰۱۰). با توجه به جدول فوق مقادیر پارامتر π_i^*، در بعضی از سؤالات کم‌تر از ۶/. است (مانند سؤال ۲ و ۳). این امر نشان‌دهنده این است که آن سؤال برای مهارت (های) اختصاص داده‌شده به آن، سخت است یا باید مهارت‌های بیش‌تر یا متفاوت‌تری (احتمالاً سخت‌تری) به چنین سؤالاتی اضافه شود تا پاسخ درست بیانگر تبحر مهارت‌های اختصاص داده‌شده به آن سؤال باشد. میانگین این پارامتر در کل سؤالات برابر با ۶۳/. است. با توجه به جدول فوق، مقادیر پارامتر r_{ik}^* در مهارت‌های مختلف در اجرای سوم به‌جز در سؤالات ۲۸ و ۲۳ کم‌تر از ۹/. است. بنابراین می‌توان گفت که بعضی از سؤالات در مهارت‌های ارائه‌شده در ماتریس دارای قدرت تشخیص نسبتاً بالایی است ولی بعضی دیگر که ۹/. یا به ۹/. نزدیک است دارای قدرت تشخیص مناسبی نیست. میانگین نمرات از ۶۰/. برای مهارت واژگان تا ۷۵/. برای مهارت استخراج اطلاعات صریح متغیر است. هم‌چنین مقادیر c_i نیز ثابت در نظر گرفته شد زیرا در ۱۰ سؤال مقدار آن زیاد بود و این بدان معناست که مهارت‌های اختصاص داده‌شده به سؤالات، برای ارائه پاسخ صحیح تا حدود زیادی کافی هستند و لازم نیست مهارت‌های دیگری در مدل وارد شوند. دامنه نمرات نسبت درست یا ضریب دشواری (نسبت آزمودنی‌هایی که به یک سؤال پاسخ درست داده‌اند) نیز از ۲۳/. برای سؤال ۱۶ و ۷۶/. برای سؤال ۷ متغیر است و میانگین آن ۴۱/. به دست آمد که نشان‌دهنده دشواری آزمون است. نکته قابل‌تأمل در پارامترهای سؤال مربوط به آزمون درک مطلب این است که تمامی سؤالات مربوط به متن ۷ (سؤالات ۷۱ تا ۸۱) به‌جز دو مورد از مقدار پارامتر خط

1. high cognitive structure
2. Robustness of skills diagnosis modeling

پایه کمی (۲۳/. تا ۶۹/.) برخوردارند. هم‌چنین سؤالات ۷۱ و ۷۸ که دارای کم‌ترین مقدار پارامتر خط پایه و بیش‌ترین مقدار پارامتر جریمه هستند هر دو مربوط به یک متن و ازنظر نوع تکلیف شبیه هم می‌باشند. بدنه سؤال ۷۱ و سؤال ۷۸ که در زیر ارائه شده است نشان‌دهنده شباهت نوع این دو سؤال است.

71- It is NOT true that-------.
78- All of the following are claims made by author EXPECT that ----.

در بررسی کیفی نیز سؤالات این بخش ازنظر متخصصین و آزمون‌شوندگان بسیار سخت ارزیابی شد. هم‌چنین سؤالات مربوط به این متن دارای همبستگی دو رشته‌ای نقطه‌ای تعدیل‌یافته بسیار کم و حتی منفی هستند.

جدول ۳. شاخص‌های اعتبار طبقه‌بندی برحسب مهارت‌های آزمون درک مطلب و افراد مسلط یا غیر مسلط

مهارت	اعتبار طبقه‌بندی درست		ضریب	اعتبار طبقه‌بندی آزمون-باز آزمون	
	مسلط	غیر مسلط	کاپای کوهن	مسلط	غیر مسلط
استفاده از دانش واژگان	/۸۳	/۹۷	/۸۲	/۷۲	/۹۳
استفاده از دانش نحوی	/۷۶	/۹۴	/۷۲	/۶۳	/۸۹
استخراج اطلاعات صریح	/۸۲	/۹۰	/۷۳	/۷۸	/۸۳
استنتاج	/۷۸	/۹۵	/۷۵	/۶۶	/۹۱
اتصال و ادغام	/۸۸	/۹۳	/۸۲	/۷۹	/۸۷
استفاده از دانش عملی	/۸۸	/۹۰	/۷۸	/۷۹	/۸۳

در مدل فیوژن به‌منظور بررسی اعتبار طبقه‌بندی، بر اساس مجموعه داده‌های شبیه‌سازی‌شده، از اعتبار طبقه‌بندی درست[1] (نسبت دفعاتی که آزمودنی به‌درستی در یک مهارت طبقه‌بندی‌شده) که نشان می‌دهد تا چه حد یک مدل شناختی، آزمودنی‌ها را بر اساس تسلط در مهارت در طبقات درست طبقه‌بندی می‌کند و اعتبار طبقه‌بندی آزمون

1. correct classification reliability (CCR)

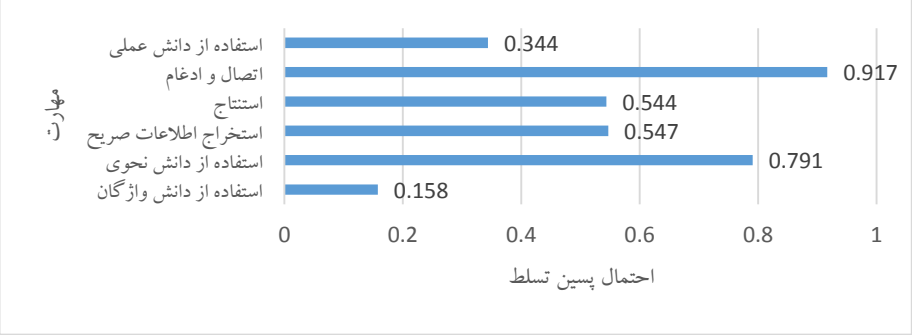
بازآزمون[1] (نسبت دفعاتی که یک آزمودنی در دو آزمون موازی در طبقه یکسانی قرار گرفته است) استفاده می‌شود (دی‌بلو و استوت، ۲۰۰۸). همان‌طور که در جدول ۴ مشاهده می‌شود، مقادیر اعتبار طبقه‌بندی درست و اعتبار طبقه‌بندی آزمون-باز آزمون در شش مهارت موردبررسی در این آزمون نسبتاً مطلوب است. این مقادیر در افراد غیر مسلط بیش‌تر از افراد مسلط است. ملاک طبقه‌بندی مقادیر احتمال پسین تسلط[2] کم‌تر از ۴/. به‌عنوان افراد غیر مسلط و مقادیر بیش‌تر از ۶/. به‌عنوان افراد مسلط در نظر گرفته شد. مقادیر بین ۴/. و ۶/. نیز مربوط به آزمودنی‌هایی بودند که در ناحیه خنثی[3] قرار داشتند یعنی نه جز افراد مسلط بودند و نه غیر مسلط.

جدول ٤. نسبت تسلط آزمودنی‌ها در هر یک از مهارت‌های درک مطلب

مهارت	مسلط		غیر مسلط		خنثی	
	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد
استفاده از دانش واژگان	۶۳۱	۲۲/۹	۲۰۰۳	۷۲/۷	۱۲۰	۴/۴
استفاده از دانش نحوی	۶۹۱	۲۵/۱	۱۸۳۱	۶۶/۵	۲۳۲	۸/۴
استخراج اطلاعات صریح	۱۰۲۲	۳۷/۱	۱۴۴۲	۵۲/۴	۲۹۰	۱۰/۵
استنتاج	۷۱۴	۲۵/۹	۱۸۵۰	۶۷/۲	۱۹۰	۶/۹
اتصال و ادغام	۱۰۷۷	۳۹/۱	۱۴۸۴	۵۳/۹	۱۹۳	۷
استفاده از دانش عملی	۱۰۱۳	۳۶/۸	۱۵۴۸	۵۶/۲	۱۹۳	۷

طبق جدول۴ اتصال و ادغام (۳۹/۱ درصد)، استفاده از دانش عملی (۳۶/۸)، استخراج اطلاعات صریح (۳۷/۱ درصد)، استنتاج (۲۵/۹ درصد)، استفاده از دانش نحوی (۲۵/۱) و استفاده از دانش واژگان (۲۲/۹ درصد) به ترتیب آسان‌ترین تا دشوارترین مهارت‌های آزمون درک مطلب است.

1. test-retest classification reliability (TCR)
2. posterior p probability of mastery (PPM)
3. indifference region



شکل ۵. نمودار نمونه نیمرخ مهارت آزمودنی (۰۱۰۰۱۰) در آزمون درک مطلب
(آزمودنی شماره ۳۲)

در شکل ۵ نیمرخ مهارت (۰۱۱۱۱۰) آزمودنی شماره ۳۲ ارائه شده است یعنی این آزمودنی در آزمون درک مطلب در مهارت‌های اول و ششم به تسلط نرسیده، در مهارت‌های سه و چهار در ناحیه خنثی قرار دارد و در مهارت‌های دوم و پنجم به تسلط رسیده است.

بحث و نتیجه‌گیری

به‌منظور مقابله با چالش‌های عملی و نظری رویکردهای سنتی و جاری سنجش قابلیت‌های زبانی و شناسایی نقاط ضعف و قوت آزمودنی‌ها به‌منظور ارائه نیمرخ تسلط در مهارت از مدل‌های تشخیصی شناختی با رویکرد ریتروفیت و مدل غیر جبرانی فیوژن استفاده شد. این مطالعه نشان داد که تدوین مهارت‌های زیربنایی آزمون و طراحی ماتریس کیو دارای مشکلاتی است. ازیک‌طرف به نظر متخصصان، آزمون مهارت‌های بیش‌تری را می‌سنجید و از طرف دیگر محدودیت‌های نظری در تدوین ماتریس کیو مثل ساده طراحی کرن آن، متوازن بودن آن و گنجاندن حداقل سه سؤال در هر مهارت (هارتز، ۲۰۰۲) و در نظر نگرفتن بیش از شش مهارت برای یک آزمون (تمپلین و هافمن، ۲۰۱۳) مانع از گنجاندن مهارت‌های بیش‌تر و ماتریس پیچیده‌تر می‌شد. این امر منجر به ادغام بعضی از مهارت‌ها پس از مشورت با متخصصین حیطه بود. هم‌راستا با دیگر مطالعات انجام‌شده با رویکرد

ریتروفیت، پیچیدگی شناسایی مهارت‌های یک آزمون در آزمون‌های درک مطالب بیش‌تر است. مثلاً جانگ (۲۰۰۸) نشان داد که مهارت‌های پردازش در آزمون‌های درک مطلب به‌صورت همزمان (استفاده همزمان از چندین راهبرد) و تعاملی (ارائه پاسخ با استفاده از منابع مختلف مثل متن، سؤال‌ها و دانش و تجربه قبلی) است که این ویژگی‌ها منجر به ایجاد چالش‌هایی در شناسایی مهارت‌های شناختی می‌شود. علاوه بر این مهارت‌های دیگری به‌غیراز مهارت‌های شناختی مثل مدیریت زمان و تیز آزمون بودن (شامل استفاده از سرنخ‌ها در سؤال‌های دیگر، حدس زدن، و استفاده از ویژگی‌های گزینه‌ها) نیز به‌عنوان مهارت‌های احتمالی اثرگذار بر ارائه پاسخ درست در آزمون‌های زبان شناسایی شده‌اند (گاآو و روگر،۲۰۱۰). سه عامل اصلی اثرگذار در درک مطلب شامل آزمودنی (عوامل فیزیکی مانند جنسیت، عوامل عاطفی، انتقال از زبان اول به زبان دوم، دانش زمینه‌ای، فرایندهای فراشناختی)، متن (سازمان معنایی، موضوع متن و متغیرهای زبانشناسی) و تکلیف (چندگزینه‌ای در مقایسه با تشریحی) است (جانگ، ۲۰۰۵). به‌طورکلی فرایندهای شناختی را می‌توان به دو دسته کلی فرایندهای مبتنی بر آزمون و فرایندهای مبتنی بر متن (جانگ، ۲۰۰۹). که در مطالعه حاضر صرفاً فرایندهای مبتنی بر متن مدنظر قرار گرفت. نکته دیگری که در تدوین ماتریس کیو مدنظر قرار گرفت ارائه خرد یا کلان مهارت‌هاست. این‌که مهارت‌ها را تا چه حد خرد یا کلان تعریف کنیم بحثی است که به میزان دقت نظریه‌های شناختی موجود در حیطه و نوع سؤال‌های آزمون برمی‌گردد. عوامل موردنیاز برای شناسایی مهارت‌های خرد مناسب به‌خوبی تعریف نشده‌اند. ارائه مهارت‌ها به‌صورت خردتر به‌منظور استنباط‌های شناختی بیش‌تر مستلزم سؤال‌های بیش‌تر است (گیرل، ونگ[1] و زو[2]، ۲۰۰۸). در این مطالعه نوع سؤال‌های طراحی‌شده به‌گونه‌ای بود که امکان خردتر طراحی کردن مهارت‌ها ممکن نشد. به نظر می‌رسد مدل‌های تشخیصی شناختی هنوز قادر نیستند به‌خوبی پیچیدگی سازه درک مطلب را ترسیم کنند. همچنین همگرایی زنجیره‌های مارکف مونته کارلو با استفاده از مدل فیوژن در داده‌های مطالعه

1. Wang
2. Zhou

حاضر در راستای دیگر مطالعات (مثل جانگ، ۲۰۰۵؛ چیو، ۲۰۰۸؛ لی و ساواکی، ۲۰۰۹؛ لی، ۲۰۱۱؛ ژانگ، ۲۰۱۳؛ فنگ، ۲۰۱۳؛ کیم، ۲۰۱۵) در زمینه کاربرد این مدل در داده‌های مربوط به آزمون‌های زبان انگلیسی بود. این مطلب برمی‌گردد به انعطاف‌پذیری الگوریتم مورداستفاده در این مدل که حتی در صورت عدم همگرایی نیز می‌توان با افزایش تعداد زنجیره‌ها و طول آن‌ها، به همگرایی رسید (روسس و همکاران، ۲۰۰۷). نتیجه این بخش از مطالعه بیانگر استفاده از دو زنجیره با طول زیاد در مدل‌های پیچیده‌ای مانند مدل‌های تشخیصی شناختی است. علاوه بر این برازش مناسب مدل با داده‌ها نیز تأیید شد ولی آزمون دارای روایی درونی بالایی نبود و تا حدی توانسته بود بین افراد مسلط و غیر مسلط تمایز قائل شود. این امر از جملات مشکلات رویکرد ریتروفیت آزمون‌های هنجار محور اجراشده موجود است که در آن آزمون با اهداف رتبه‌بندی طراحی‌شده و نه اهداف تشخیصی. نتایج حاصل از برازش مدل و روایی درونی نیز هم‌راستا با مطالعات فوق‌الذکر بود. پارامترهای سؤال به‌دست‌آمده در این مطالعه نشان‌دهنده آزمون با ساختار شناختی بالا نیست. میانگین مقادیر پارامتر خط پایه ۶۳/. بود و میانگین مقادیر پارامتر جریمه در هیچ‌یک از شش مهارت احتمالی آزمون کم‌تر از ۵/. به دست نیامد. دلیل اصلی این امر را می‌توان دشواری آزمون در نظر گرفت. وقتی آزمون دشوار باشد حتی افرادی که در مهارت‌ها به تسلط نرسیده‌اند نیز قادر به پاسخگویی به سؤالات مربوط به آن مهارت نیستند. شاید اضافه کردن مهارت‌های دشوارتر یا مهارت‌های مربوط به آزمون مثل تیز آزمون بودن و مدیریت زمان به برآورد بهتر پارامترها کمک کند. هم‌چنین اضافه کردن سؤالاتی با ضریب تشخیص بیش‌تر نیز مفید است. این امر نیاز به بررسی روایی بیش‌تر ماتریس کیو و شناسایی مهارت‌های متفاوت برای آزمون را نشان می‌دهد. بااین‌وجود باید در نظر داشت که آزمون موردبررسی در این مطالعه یک آزمون هنجاری، غیرتشخیصی و با اهداف رتبه‌بندی بوده است که در آن صرفاً مهارت‌های شناختی مربوط به متن نمی‌تواند منجر به پاسخ درست شود. علاوه بر این نتایج حاصل از این مطالعه نشان داد که مهارت کاربرد واژگان، سخت‌ترین مهارت محسوب می‌شود. این نتیجه در راستای نتیجه به‌دست‌آمده از مطالعه لی (۲۰۱۱) و جانگ (۲۰۰۵) است. دلیل این مطلب آن است که در بسیاری از

آزمون‌های درک مطلب، دانستن معنی واژه‌ها از اهمیت زیادی برخوردار است. به‌طوری‌که الدرسون (۲۰۰۰) بر این باور است که سازه خواندن مستلزم دو مهارت درک و دانش بالا در زمینه واژگان است. به‌زعم گارسیا (۱۹۹۱) دانش کم در زمینه واژگان مانعی جدی در درک مطلب است. هم‌چنین گربه (۲۰۰۹) در طی بررسی‌های خود نشان داده است که برای درک یک متن خواننده باید با ۹۵٪ واژگان آن متن آشنا باشد. هم‌راستا با نتیجه مطالعه حاضر، در پژوهش لی (۲۰۱۱) در زمینه درک مطلب نیز مهارت استفاده از دانش واژگان سخت‌ترین مهارت و درک اطلاعات مستتر ساده‌ترین مهارت شناسایی شد. نتیجه این بخش از مطالعه در راستای نتایج جانگ (۲۰۰۵) در زمینه شناسایی نقاط قوت و ضعف آزمون‌شوندگان در آزمون درک مطلب بود. در مطالعه او مهارت‌ها مانند مربوط به واژگان، گرامر و خلاصه کردن ایده‌های اصلی متن ساده‌ترین و ترسیم ایده‌های متضاد در یک چارچوب دشوارترین مهارت‌ها شناسایی شد. هم‌چنین سوتینا و همکاران (۲۰۱۱) نشان دادند که فرایندهای شناختی پیچیده مثل درک ایده‌های ضمنی سخت‌ترین مهارت و فرایندهای شناختی پایه مثل معنی واژگان ساده‌ترین مهارت شناسایی بود. نتیجه مطالعه رواند و روبیتز (۲۰۱۵) نیز نشان داد که صرف و نحو، ساده‌ترین و استنتاج دشوارترین مهارت بوده‌اند. بر اساس نتایج حاصل از این مطالعه پیشنهاد می‌شود که تحقیقات بیش‌تری برای تدوین دقیق روایی ماتریس کیو انجام شود. نکته دیگر این‌که این مطالعه به بررسی ویژگی‌های مدل غیر جبرانی فیوژن محدود بود و نتایج آن با دیگر مدل‌های تشخیصی شناختی و بخصوص مدل‌های جبرانی در آن مقایسه نشده است. پیشنهاد می‌شود نتایج حاصل از مدل جبرانی با نتایج حاصل از این مطالعه بررسی شود. علاوه بر این‌یکی مشکلات موجود در تحلیل و تفسیر داده‌ها ناشی از استفاده از این مدل‌ها برای ریتروفیت آزمون طراحی‌شده برای اهداف غیرتشخیصی بود. لذا پیشنهاد می‌شود که آزمونی با اهداف تشخیصی تدوین و با استفاده از این مدل‌ها موردبررسی قرار گیرد تا به‌طور دقیق‌تری بتوان میزان تسلط در مهارت‌های مختلف را نشان داد.

منابع

دفتر طرح و آمار سازمان سنجش (۱۳۹۲). کارنامه آزمون نیمه‌متمرکز دکتری، معاونت فنی و آماری، دفتر طرح و آمار.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Bradshaw, L, Izsák, A., Templin, J and Jacobson, E. (2014). Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test within the Diagnostic Classification Framework, Educational measurement: Issues and practice, 23(1), 2-14.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test, *Language Testing*, 15(2), 119–157.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section* (Research Rep. No. RR-98-19), Princeton, NJ: Educational Testing Service.

Chiu, C. (2008). *Cluster analysis for cognitive diagnosis: theory and application*, unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

de la Torre, J. (2008). An empirically based method of Q-Matrix validation for the DINA model: development and applications, *Journal of Educational Measurement*, 45(4), 343–362.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques, chapter cognitively diagnostic assessment, pages 361-390. Hillsdale, NJ: Erlbaum.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.

Garccia, G. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. *Reading Research Quarterly*, 26, 371–392.

Gao, L. & Roger, T. (2010). Use of tree-based regression in the analyses of L2 reading test items, *Language Testing, 28(1) 77–104*, DOI: 10.1177/0265532210364380

Gierl, M.J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive

skills in algebra on the SAT, Journal of technology, learning, and assessment, 6(6). Retrieved from http://www.jtla.org.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375–406.

Feng, Y. (2013). Estimation and Q-matrix validation for diagnostic classification models. (Master's thesis). University of South Carolina, Retrieved from http://scholarcommons.sc.edu/etd/2611.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica 37, 359-374.*

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality*, Doctoral dissertation. University of Illinois, Urbana-Champaign.

Henson, R; Roussos, L.; Douglas, J. & He, X. (2008). Cognitive diagnostic attribute-level discriminate indices. *Applied psychological measurement*, 32(4), 275-288, DOI: 10.1177/0146621607302478.

Jang, E.E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG-TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign, Urbana, IL.

Jang, E. E. (2008). *A framework for cognitive diagnostic assessment*. In C. A . Chapelle, Y. Chung, & J. Xu,
Towards adaptive CALL: Natural language processing for diagnostic l anguage assessment (pp. 117-131). Ames, IA: Iowa State University.

Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment, *Language Testing*, 26(1), 31-73.

Jang, E.E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills, *Language Assessment Quarterly*, 6(3), 210-238, DOI: 10.1080/15434300903071817.

Kim, A. Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability, *Language Testing*, doi:10.1177/0265532214558457

Kim, S. H. & Kim, S. (2013). Incorporating diagnostic aspects to mathematical affects inventory development, *International Journal of Evaluation and Research in Education (IJERE)*, 2(4), 163-174.

Kim, Y., H., (2010). *An argument-based validity inquiry into empirically-driven descriptor-based diagnostic (EDD) assessment in ESL academic writing, unpublished doctoral dissertation*, University of Toronto.

Kim, H. S. (2011). *Diagnosing examinees' attributes-mastery using the Bayesian inference for binomial proportion: a new method for*

cognitive diagnostic assessment, unpublished doctoral dissertation, Georgia Institute of Technology.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Lee, Y.-W., & Sawaki, Y. (2009b). *Cognitive diagnosis approaches to language assessment: An overview*, Language Assessment Quarterly, 6(3), 172-189. doi: 10.1080/15434300902985108

Leighton, J. & Gierl, M. (2007). Cognitive diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan fellow working papers in second or foreign language assessment*, 9(???), 17–46.

Ravand, H. & Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research & Evaluation*, 20 (11). Available online: http://pareonline.net/getvn.asp?v=20&n=11.

Roussos, L.A., DiBello, L.V., Stout, W.F., Hartz, S.M., Henson, R.A., & Templin, J.H. (2007). *The fusion model skills diagnostic system*. In J. Leighton & M. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

Rupp, A., Templin, J. & Henson, R. (2010). *Diagnostic measurement: theory, methods, and applications*. New York: The Guilford Press.

Snow, R. E., & Lohman, D. F. (1989). Implication of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education/Macmillan.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30,* 251-275.

Templin, J. & Henson, R. A. (2006). Measurement of Psychological disorders using cognitive diagnostic models. *Psychological Methods*, 11(3), 287–305, DOI: 10.1037/1082-989X.11.3.287.

Templin, J. L., & Hoffman, L. (2013). Obtaining Diagnostic Classification Model Estimates Using Mplus, *Educational Measurement: Issues and Practice*, 32 (2), 37–50.

Svetina, D., Gorin, J. S., Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric

modeling approach, *International Journal of Testing*, 11 (1), 1-23, DOI: 10.1080/15305058.2010.518261.

Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: A new networking model in language testing and experiment with a new psychometric model and task type* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Zhang, J. (2013). Relationships between missing responses and skill mastery profiles of cognitive diagnostic assessment, unpublished doctoral dissertation, University of Toronto.