

مقایسه سؤال‌های چندگزینه‌ای و ترجمه‌ای از لحاظ پایایی و دشواری در آزمون دانش دستوری

حامد زندی*

دکترای آموزش زبان انگلیسی از دانشگاه تهران، مدرس دانشگاه تحصیلات تکمیلی

در علوم پایه زنجان، ایران

شیوا کیوان‌پناه**

دانشیار، گروه زبان و ادبیات انگلیسی، دانشکده زبان‌ها و ادبیات خارجی، دانشگاه تهران،

تهران، ایران

سید محمد علوی***

دانشیار، گروه زبان و ادبیات انگلیسی، دانشکده زبان‌ها و ادبیات خارجی، دانشگاه تهران،

تهران، ایران

(تاریخ دریافت: ۹۲/۴/۲۶، تاریخ تصویب: ۹۳/۲/۱۳)

چکیده

به طور معمول از سؤال‌های چندگزینه‌ای برای سنجش دانش زبانی استفاده می‌شود. اما از آنجا که در رویکرد نظام‌مند طراحی سؤال، که جدیداً مطرح شده است (شین، ۲۰۱۲)، استفاده از صورت‌های متنوع برای طراحی سؤال مطلوب است، همچنین به دلیل اینکه تحقیقات بسیار اندکی در مورد امکان استفاده از سؤال‌های ترجمه‌ای برای تنوع بخشیدن به سؤال‌های آزمون وجود دارد، پژوهش حاضر با هدف مقایسه دو کمیت پایایی و دشواری در صورت سؤال ترجمه‌ای و چندگزینه‌ای برای سنجش دانش دستوری طراحی شده است. ابتدا، سند مشخصات آزمون دارای دو بخش ترجمه‌ای و چندگزینه‌ای را هشت مدرس زبان با تحصیلات کارشناسی ارشد و دکتری بررسی کردند. بر اساس بازخورد آن‌ها، آزمون مورد استفاده در این پژوهش تهیه و با ۱۵۸ فراگیر زبان انگلیسی با توانش‌های مختلف برگزار شد. نتایج به دست آمده حاکی از آن است که هر دو صورت سؤال آزمون و نیز مدل راش تحلیل شد. نتایج به دست آمده حاکی از آن است که هر دو صورت سؤال دارای پایایی خوب بود ($0/884 = \alpha$ کرونباخ بخش ترجمه‌ای و $0/845 = \alpha$ کرونباخ بخش چندگزینه‌ای). براساس مدل راش، تمامی سؤال‌ها از نظر شاخصه‌های برازش، مطلوب ارزیابی شد. به علاوه، آزمون تی مستقل تفاوتی میان دشواری این دو صورت سؤال نشان نداد ($t=1/696, df=58 = P=0/095$ آزمون دو دامنه). بنابراین، به نظر می‌رسد بتوان از صورت ترجمه‌ای در کنار صورت متداول چندگزینه‌ای در رویکردی نظام‌مند برای طراحی سؤال‌های دانش دستوری استفاده کرد.

واژه‌های کلیدی: پایایی، دشواری، رویکرد طراحی سؤال، سؤال‌های ترجمه‌ای، سؤال‌های چندگزینه‌ای.

*. (نویسنده مسئول) تلفن: ۰۲۱-۶۱۱۱۹۰۱۶، دورنگار: ۰۲۱-۸۸۶۳۴۵۰۰، E-mail: zandi@iasbs.ac.ir

** تلفن: ۰۲۱-۶۱۱۱۹۰۱۶، دورنگار: ۰۲۱-۸۸۶۳۴۵۰۰، E-mail: shkaivan@ut.ac.ir

*** تلفن: ۰۲۱-۶۱۱۱۹۰۱۶، دورنگار: ۰۲۱-۸۸۶۳۴۵۰۰، E-mail: smalavi@ut.ac.ir

مقدمه

در مراکز آموزشی زبان در سراسر دنیا از مدرسان زبان انگلیسی انتظار می‌رود برای فراگیران کلاس خود آزمون‌های میان‌ترم و پایان‌ترم طراحی کنند. نوشتن سؤال‌های آزمون در منظر نخست کاری پیچیده به نظر نمی‌آید، ولی با تحلیلی عمیق و بررسی تاریخی درمی‌یابیم که نحوه و نوع نگارش سؤال‌ها بیانگر گرایش طراح به رویکرد طراحی سؤال است، مانند روان‌سنجی^۱، اصل‌محوری^۲، نظام‌مندی^۳، و انتقادی^۴. این پژوهش در نظر دارد دو گونه طراحی صورت سؤال^۵ یعنی چندگزینه‌ای و ترجمه‌ای را از نظر مکتب نظام‌مند و مقوله پایایی^۶ و دشواری^۷ بررسی کند. این تتبع از آن جهت اهمیت دارد که امروزه سؤال‌های چندگزینه‌ای در آزمون‌های مراکز آموزش زبان بسیار متداول است، ولی ظرفیت و امتیازهای سؤال‌های ترجمه‌ای مورد غفلت واقع شده است. این درحالی است که روش‌های آموزش زبان انگلیسی که در آن استفاده از زبان مادری فراگیران در آموزش جایگاهی ندارد، در مجموع به ضرر یادگیری زبان دوم است (کوک، ۲۰۱۰). متأسفانه در دهه‌های اخیر ترویج این روش‌های آموزش تک‌زبانی اثر منفی خود را بر نحوه ارزیابی زبان انگلیسی نیز بر جای گذاشته است. اغلب قریب به اتفاق آزمون‌های زبان انگلیسی چه آن‌هایی که نقش تعیین‌کننده در سرنوشت افراد^۸ دارند مانند امتحانات ورودی دانشگاه‌ها و چه آزمون‌های غیرسرنوشت‌ساز^۹ تک‌زبانی‌اند و به امتیازات استفاده از ترجمه برای سنجش بهایی داده نشده است. این امر باعث شده است که تنها تحقیقات پراکنده و معدودی در زمینه کاربرد ترجمه به منزله وسیله سنجش دانش و توانش زبانی انجام بگیرد.

در این راستا، در این پژوهش سعی بر آن است امتیازات این دو صورت سؤال از جهت پایایی و دشواری با یکدیگر مقایسه شود و راهکارهایی برای طراحی و استفاده بهینه از سؤال‌های ترجمه‌ای برای مهندسی متوازن آزمون در رویکردهای نوین سنجش پیشنهاد شود. از

-
1. psychometric
 2. authentic
 3. systematic
 4. critical
 5. item format
 6. reliability
 7. item difficulty
 8. testshigh takess
 9. low stakes tests

آنجا که جو غالب در آموزش و سنجش عدم اقبال به ترجمه است، دلایل این موضوع بحث می‌شود تا علاقه‌مندان به استفاده از رویکردهای نوین سنجش بتوانند تصمیم آگاهانه‌تری در استفاده از نوع سؤال در آزمون اتخاذ کنند.

پیشینه تحقیق

نحوه و نوع طراحی سؤال متأثر از عوامل گوناگونی است، از جمله چارچوب فلسفی حاکم، فرهنگ ارزیابی، یا دیگر پیش‌فرض‌هایی که در عرف طراحی سؤال وجود دارد (شین، ۲۰۱۲). برای نمونه، در نیمه قرن بیستم با تأثیر از مکتب روان‌شناسی رفتارگرایی و اثبات‌گرایی (جانسون و جانسون، ۱۹۹۹) آزمون‌سازی زبان رویکردی کمی و عینی به خود گرفته بود (فرهادی، ۲۰۰۷). بنابراین، در دهه شصت میلادی مبحث پایایی آزمون یکی از موضوعات عمده در مباحث آزمون‌سازی زبان محسوب می‌شد که از قابلیت اندازه‌گیری به صورت کمی و عینی برخوردار است (اسپالسکی، ۲۰۱۰). در همین دوره بود که روان‌سنجی الهام‌بخش آزمون‌سازی زبان شد (مک‌نامارا و روور، ۲۰۰۶) و با کمک زبان‌شناسی ساختارگرا نکات مختلف دانش زبان استخراج و برای سنجش در قالب سؤال‌های چندگزینه‌ای قرار گرفت. شین (۲۰۱۲: ۲۳۸) اصل حاکم بر طرح سؤال در رویکرد روان‌سنجی را این‌گونه خلاصه می‌کند: «سؤال‌ها را همانند هم و به‌صورت چندگزینه‌ای تهیه کنید. این کار را به صورت عینی انجام دهید.»

در اواخر قرن بیستم، رویکرد اصل‌محوری در طراحی سؤال مطرح شد که در آن سنجش مهارت‌های ارتباطی به گونه‌ای که زبان به صورت روزمره در جامعه استفاده می‌شود اهمیت خاصی دارد. صورت‌های مختلف آزمون از قبیل بررسی پوشه عملکرد، دفترچه خاطرات، ارزیابی از خود مطرح شد که مطابق با دیدگاه‌های کیفی، ذهنی، فرایندگرا و در ارتباط با جهان واقعی بود (براون و هادسن، ۱۹۹۸). اصل حاکم بر طراحی سؤال در این رویکرد جست‌وجوی سؤال‌های اصل‌محور و پرهیز از سؤال‌های چندگزینه‌ای بود (شین، ۲۰۱۲). دلیل عدم استفاده از صورت چندگزینه‌ای انتقادهای متعددی بود که بر آن وارد شده بود، از جمله اینکه امکان سنجش مهارت‌های تعاملی به گونه‌ای وجود ندارد که در زندگی واقعی اتفاق می‌افتد و این آزمون بیشتر در سنجش دانش

زبانی مانند دستور و مهارت‌های دریافتی^{۱۰} مانند خواندن سودمند است (براون و هادسن، ۱۹۹۸). انتقاد دیگر به احتمال حدس‌زدن آزمون‌دهندگان هنگام پاسخگویی و دشواری طراحی و صرف وقت مضاعف برای طراحی سؤال‌های چندگزینه‌ای خوب مربوط بود (پورپورا، ۲۰۰۴).

در دهه گذشته، بسیاری از طراحان آزمون زبان رویکردی میانه را پی گرفته‌اند. این گروه معتقدند که صورت خاصی از سؤال به تنهایی قابلیت این را ندارد که تمام ابعاد سازه آزمون را به صورت کامل بسنجند. ضمن تأیید این نظر، باک (۲۰۰۱: ۱۵۳) می‌گوید: «تمام صورت‌های سؤال دارای ضعف‌ها و قوت‌های خاص خود است و مهارت‌های گوناگونی برای پاسخ به آن‌ها نیاز است: برخی یک دسته از مهارت‌ها را ارزیابی می‌کنند و برخی دسته‌ای دیگر را؛ برخی به نفع گروه خاصی از آزمون‌دهنده‌هاست و برخی به نفع گروهی دیگر. با استفاده از صورت‌های متنوع، آزمون قابلیت بیشتری پیدا می‌کند تا مبنای ارزیابی متوازی قرار بگیرد.» براون و هادسن (۱۹۹۸) نیز رأی مشابهی دارند. آن‌ها ضمن بررسی سه صورت سؤال که هر کدام مبتنی بر یکی از دو رویکرد روان‌سنجی و اصل‌محوری بود، به این نتیجه رسیدند که هر یک از این صورت‌های طرح سؤال برای مقاصد خاصی از ارزیابی کارایی دارد و برای سنجش بخشی از دانش و مهارت که ممکن است با صورت دیگر سنجیده نشود، استفاده از صورت‌های مختلف سؤال را به معلمان توصیه کردند. الدرسون، کلپهام و وال (۱۹۹۵) نیز معتقدند تا زمانی که اطلاعات کافی در مورد کارایی صورت سؤال برای سنجش جوانب مختلف توانایی‌های افراد وجود ندارد، بهتر است از چند صورت سؤال مختلف استفاده کرد.

لازمه ارزیابی متوازن، در رویکرد نظام‌مند، بررسی مشخصات آزمون^{۱۱} است. در دهه‌های اخیر کتب مرجع آزمون‌سازی زبان به موضوع مشخصات آزمون پرداخته‌اند (از جمله، باکمن و پالمر، ۱۹۹۶؛ باکمن و پالمر، ۲۰۱۰؛ فالچر و دیویدسن، ۲۰۰۷؛ دیویدسن و لینچ، ۲۰۰۲). مشخصات آزمون در واقع نقشه مهندسی آن به شمار می‌رود که به کمک آن سؤال‌های یکسان و همگونی را می‌توان تولید کرد (دیویدسن و لینچ، ۲۰۰۲). تهیه سند مشخصات آزمون قدمی است در راستای رویکردی نظام‌مند نسبت به طراحی سؤال‌ها، زیرا در این رویکرد اصل بر آن

10. skills receptive
11. test specifications

است که به نمای وسیع‌تر سیر تولید آزمون توجه و نوشتن سؤال‌ها نظام‌مند شود (شین، ۲۰۱۲). در واقع، سند مشخصات آزمون ابزاری مهم در استانداردسازی و کنترل کیفیت آزمون در بسیاری از شرکت‌ها و سازمان‌هایی است که در امر سنجش آموزش فعالیت می‌کنند (دیویدسون، ۲۰۱۲).

به علاوه، فرایند تهیه این سند زمینه‌ای را ایجاد می‌کند تا دست‌اندرکاران تهیه آزمون، مانند گروهی از مدرسان مؤسسه آموزشی، در مورد نحوه طراحی، اهداف آزمون، و تناسب آن با اهداف آموزشی به گفت‌وگو و تبادل نظر بپردازند، دیدگاه‌های خود را به صورت شفاف بیان کنند، و به اجماع برسند. لی (۲۰۰۶) رابطه اجماع نظر در تنظیم سند مشخصات آزمون و روایی را این‌گونه بیان می‌کند که در ابتدا نسخه‌ای اولیه از سند مشخصات آزمون تهیه می‌شود و به طور طبیعی دارای کاستی‌هایی است. سپس، این سند را گروهی دیگر از مدرسان مؤسسه در معرض نقد و بررسی قرار می‌دهند. پس از جمع‌آوری بازخوردهای این افراد، نسخه‌ای جدید و کامل‌تر از سند تنظیم و بار دیگر به بوتۀ نقد گذاشته می‌شود. این چرخه آنقدر ادامه پیدا می‌کند که افراد به اجماع نظر برسند که سند مشخصات آزمون توانایی تولید سؤال‌های همگون با توجه به سازه آزمون را دارد، و نیز در راستای اهداف کلی سندهای بالادستی مانند برنامه آموزشی است. بدین ترتیب، مجموعه این اقدام‌ها کمک می‌کند تا بعضی ملزومات روایی آزمون، حتی پیش از آنکه آزمونی برگزار شود، مد نظر تهیه‌کنندگان آزمون قرار بگیرد. به این دسته از اقدام‌ها، که با استدلال به آن‌ها بتوان در مورد روایی آزمونی پیش از برگزاری آن بحث کرد، روایی سنجی ماقبل^{۱۲} می‌گویند که مکمل روایی سنجی متداول مابعد^{۱۳} برگزاری آزمون است (ور، ۲۰۰۵). در نتیجه، در رویکرد نظام‌مند، استناد به سند مشخصات آزمون را چند کارشناس نقد و تصحیح کردند. بررسی انطباق سؤال‌های آزمون با سند، بخشی از استدلال برای روایی آزمون را تشکیل می‌دهد.

تحقیق حاضر قصد دارد با استفاده از رویکرد نظام‌مند در طراحی سؤال، ویژگی‌های پایایی و دشواری طراحی سؤال را در صورت‌های چندگزینه‌ای و ترجمه‌ای بررسی کند. این پژوهش از آن جهت مفید است که سؤال‌های چندگزینه‌ای در آزمون‌های زبان بسیار رایج است ولی

12. priori validation

13. posteriori

کمتر به ظرفیت و امتیازهای سؤال‌های ترجمه‌ای با استفاده از رویکرد نظام‌مند در ایران پرداخته شده است. سؤال‌های این پژوهش عبارت‌اند از:

- ۱- آیا از نظر پایایی بین آزمون سؤال‌های چندگزینه‌ای و آزمون سؤال‌های ترجمه‌ای دانش دستوری تفاوت معناداری وجود دارد؟
- ۲- آیا از نظر دشواری بین آزمون سؤال‌های چندگزینه‌ای و آزمون سؤال‌های ترجمه‌ای دانش دستوری تفاوت معناداری وجود دارد؟

روش تحقیق

ابزار تحقیق

یکی از مراحل مهم این پژوهش تهیه سند مشخصات آزمون بود. برای این منظور سندی با ویژگی‌های توصیه شده دیویدسن و لینچ (۲۰۰۲) برگرفته از نظرهای پاپ‌هام (۱۹۷۸) تهیه شد. این سند شامل بخش‌های توضیحات کلی در مورد آزمون، ویژگی‌های سؤال‌ها، ویژگی‌های جواب مطلوب، و نمونه سؤال بود که هشت معلم زبان انگلیسی که خود دانشجوی دوره کارشناسی ارشد یا دکتری یا دکترای رشته آموزش یا سنجش زبان بودند، طی نشست‌های کارگاهی متعدد بررسی کردند. پس از دسته‌بندی بازخوردها و با بهره بردن از نظرهای کارشناسی این افراد، سند آزمون بازنویسی و اصلاح شد. از جمله دستورات عمل‌هایی که برای طراحی سؤال در سند نهایی درج شده بود، به موارد زیر اشاره می‌شود.

مشخصات آزمون

این آزمون، که با هدف سنجش دانش دستوری فراگیران زبان انگلیسی از زمان‌ها، جملات شرطی، و عبارات موصولی طراحی شده بود، شامل دو قسمت ترجمه‌ای با ۲۸ سؤال و چندگزینه‌ای با ۳۲ سؤال^{۱۴}. باید توجه کرد که گرچه صورت سؤال‌های بخش ترجمه‌ای شبیه به سؤال‌های ترجمه بود، در واقع، این سؤال‌ها، دانش دستوری زبان‌آموزان را می‌سنجید. به این

۱۴. چهار سؤال ترجمه‌ای به طور هم‌زمان دو نکته دستوری را می‌سنجید. در نتیجه، هر یک با دو سؤال در بخش دوم متناظر بود.

دلیل، این صورت سؤال ترجمه‌ای است. بخش سؤال‌های ترجمه‌ای آزمون پیش از بخش چندگزینه‌ای برگزار شد. هر سؤال بخش ترجمه متناظر با سؤالی در بخش چندگزینه‌ای بود؛ برای مثال، در هر دو بخش آزمون سؤالی برای سنجش دانش فرد در زمینه منفی کردن زمان حال ساده وجود داشت تا هر دو بخش آزمون از نظر محتوا شبیه هم باشند. در تمامی سؤال‌های بخش ترجمه، هدف سنجش توانایی فرد برای ارائه معادل جملات در زبان مقصد بود. در توضیحات آزمون قید شد که افراد باید تا جایی که می‌توانند جملات را به انگلیسی ترجمه کنند و نگران غلط‌های املائی خود نباشند. در ضمن نمونه‌ای از پاسخ صحیح در اختیار آن‌ها قرار داده شد. هر سؤال حداکثر به طول یک سطر بود. سؤال‌های ترجمه‌ای به گونه‌ای طراحی شده بود که تنها یک معادل صحیح داشت. در نتیجه، از متون ادبی و پیچیده پرهیز شده بود.

برای آنکه کاستی در دانش واژگانی آزمون‌دهنده مشکلی در فرایند ترجمه ایجاد نکند، فرهنگ لغت کوچک حاوی تمامی لغات اصلی در اختیار افراد قرار گرفت. هنگام تصحیح سؤال‌ها تنها به بخشی از ترجمه نمره داده شد که هدف اصلی سنجش بود. در نتیجه، سیاق، سبک و املا در نمره‌دهی تأثیری نداشت. برای مثال، در سنجش توانایی منفی کردن زمان حال ساده تنها به بخش گروه فعلی توجه شد. در نتیجه، نمره ترجمه عبارت «علی معمولاً ریاضیات درس نمی‌دهد.» در ترجمه‌های ۱ و ۲ یکسان است ولی ترجمه ۳ نمره‌ای ندارد.

1. math teach not dose usually Ali.
2. math teach usually not dose Ali.
3. Ali do not usually teach math.

هدف این سؤال خاص تنها تشخیص دانش منفی کردن است و نه کاربرد قیود تکرار. همچنین، تفاوت در ترجمه «ریاضیات» مربوط به دانش واژگانی است و در نمره‌دهی به دانش دستوری نقشی ندارد. اما ترجمه ۳ غلط است، زیرا بیانگر نقص در دانش دستوری زمان حال ساده است.

در سؤال‌های چندگزینه‌ای نیز تمامی کلمات به کار رفته در متن سؤال جزو ۲۰۰۰ کلمه پربسامد پیکره (BNC) بود. تمامی سؤال‌ها تقریباً هم‌اندازه، کاملاً روان، و از نظر زبانی ساده بود. بعد از هر سؤال سه گزینه غلط و یک گزینه صحیح وجود داشت.

پیش مطالعه

تمامی سؤال‌ها را کارشناسان بررسی کردند. در این مطالعه آزمودنی‌ها چهل دانشجوی دختر و پسر بودند که در رشته‌ای غیر از زبان انگلیسی تحصیل می‌کردند. تمامی سؤال‌ها از لحاظ میزان آسانی، دشواری، همبستگی سری دو نقطه‌ای^{۱۵}، آلفای کرونباخ، و تحلیل گزینه‌ها بررسی آماری و سؤال‌های مشکل‌دار بازنویسی شد. از جمله اشکال‌های عمده در سؤال‌های ترجمه آن بود که گاهی برخی از آن‌ها را می‌شد به دو گونه ترجمه کرد. این دسته از سؤال‌ها به گونه‌ای اصلاح شد که تنها یک حالت ترجمه ممکن بود. از جمله اشکال‌ها سؤال‌های چندگزینه‌ای، ضعف و نقص در طراحی گزینه‌ها بود که باید برطرف می‌شد.

پس از بازنویسی سؤال‌های مشکل‌دار، سؤال‌ها را بار دیگر کارشناسان بازبینی کردند و برای مطالعه‌ای دیگر آماده شد. در این مرحله ۳۳ دختر و پسر مشابه گروه آزمودنی مطالعه نخست به سؤال‌ها پاسخ دادند و تمامی تحلیل‌های آماری مطالعه پیشین از جمله میزان آسانی، دشواری، همبستگی سری دو نقطه‌ای^{۱۶}، آلفای کرونباخ، و تحلیل گزینه‌ها روی سؤال‌ها اعمال شد. این بار تعداد کمتری از سؤال‌های مشکل‌دار بود که بار دیگر بازنویسی و برای آزمون اصلی آماده شد.

مطالعه اصلی

دویست و پنجاه دختر و پسر ایرانی، با توانش‌های متفاوت زبانی و در رشته تحصیلی غیر از زبان انگلیسی در این آزمون شرکت کردند. پس از بررسی‌های اولیه، ۹۲ نفر از آن‌ها به دلیل پاسخگویی ناقص به سؤال‌ها از تحلیل نهایی کنار گذاشته شدند. سپس، از نرم‌افزارهای اکسل، اسپیس اس‌اس برای تحلیل‌های نظریه کلاسیک آزمون و کانستراکت مپ (کندی و همکاران، ۲۰۰۶) برای تحلیل‌های مدل راش استفاده شد.

15. biserial point

16. biserial point

تجزیه و تحلیل داده‌ها

به منظور بررسی پایایی و دشواری آزمون سؤال‌های چندگزینه‌ای و آزمون سؤال‌های ترجمه‌ای دانش دستوری، داده‌ها با نرم‌افزار کانستراکت مپ تحلیل شد تا درجه‌بندی^{۱۷} دشواری سؤال‌ها و توانایی آزمودنی‌ها مشخص شود. مدل راش این قابلیت را دارد که نتایج حاصل وابسته به گروه خاصی از آزمودنی‌ها نیست و می‌توان آن‌ها را به گروه‌های آزمودنی دیگر نیز تعمیم داد. به نظر می‌رسد در متون تحقیق معیار واحدی برای حجم نمونه آماری در مدل راش مشخص نشده است (موریزوت و همکاران، ۲۰۰۷). با وجود این، دیمارس (۲۰۱۰) پیشنهاد می‌کند نمونه‌هایی با حجم ۱۰۰ یا ۲۰۰ در مدل راش تحلیل شود و حجم نمونه‌هایی با حدود ۱۵۰ نفر، درجه‌بندی سؤال‌ها را با ثبات بالا در لاجیت‌ها به دست می‌دهد که تا ۹۹ درصد درجه اطمینان دارد (لینکر، ۱۹۹۴). بررسی فرض تک‌بعدی بودن کل آزمون حاکی از آن است که بعد تبیین شده با مدل راش ۳۷/۶ درصد واریانس داده‌ها را دربرمی‌گیرد و بزرگ‌ترین بعد ثانویه تنها ۴/۱ درصد واریانس را تبیین می‌کند. در نتیجه، می‌توان گفت هر دو بخش آزمون یک بعد، یعنی دانش دستوری زبان‌آموزان، را می‌سنجد.

مکنامارا (۱۹۹۶) پیشنهاد می‌کند که آمار برازشی^{۱۸} در حدود تقریبی ۰/۷۵ تا ۱/۳ پذیرفتنی است. لینکر (۲۰۱۰) نیز حدود ۰/۷۰ تا ۱/۳ را پیشنهاد می‌دهد، اما معتقد است که آمار برازشی بین ۰/۵ تا ۱/۵ برای مقاصد سنجش قابل استفاده است (لینکر، ۲۰۰۲). تحلیل آمار برازشی که در پیوست آمده است، بیانگر آن است که درون برازش^{۱۹} سؤال‌ها به طور متوسط ۰/۸۵ است و هیچ سؤالی بیش‌برازش^{۲۰} ندارد؛ یعنی، درون‌برازش بالای ۱/۳ ندارد. به عبارت دیگر، تمام سؤال‌های سازه آزمون، یعنی دانش دستوری، را می‌سنجد. فقط در هشت سؤال مقدار درون‌برازش کمتر از حد ۰/۷۰ لینکر (۲۰۱۰) بود. اما، با توجه به پیشنهاد لینکر (۲۰۰۲) این کم‌برازشی^{۲۱} قابل اغماض و بالای ۰/۵ است.

17. calibration

18. fit

19. infit

20. overfit

21. underfit

مقادیر برون‌برازش^{۲۲} بیانگر میزان حساسیت سؤال‌ها به توانایی آزمودنی‌هاست، به گونه‌ای که هرچه تفاوت بین میزان توانایی آزمودنی و دشواری سؤال بیشتر باشد، به همان میزان مقدار برون‌برازش بیش از عدد ۱ می‌شود. لیکر (۲۰۰۲) معتقد است شاخص برون‌برازش در مقایسه با درون‌برازش مشکل مهمی را برای سنجش ایجاد نمی‌کند. از آنجا که تمامی سؤال‌های آزمون تحقیق حاضر دارای برون‌برازش بین ۰/۷۵ تا ۱/۳ و به طور متوسط ۰/۹۱ بود، برای مقاصد سنجش کارایی لازم را داشت.

دشواری آزمون ترجمه‌ای و آزمون چندگزینه‌ای با استفاده از مدل راش و بر مبنای لاجیت به شرح جدول ۱ است. در مبنای لاجیت، دشواری یک سؤال ممکن است عددی منفی (آسان‌تر از حد متوسط)، مثبت (دشواری‌تر از حد متوسط) یا صفر (با دشواری متوسط) باشد. در این آزمون دشواری سؤال‌ها بین حدود ۴- و ۴+ بود. همان‌گونه که در جدول ۱ مشاهده می‌شود، میانگین دشواری دو صورت سؤال نزدیک است. علاوه بر آن، میانگین خطای استاندارد سنجش هر دو بخش آزمون یکسان و در حد قابل قبول است. میانگین خطای استاندارد سنجش در مدل راش کاربردی همانند مفهوم پایایی در نظریه کلاسیک آزمون دارد، زیرا از مقدار پایایی برای محاسبه خطای استاندارد سنجش استفاده می‌شود. با استفاده از نظریه کلاسیک، آزمون پایایی آزمون بخش ترجمه (۰/۸۸۴ = آلفای کرونباخ) و بخش چندگزینه‌ای (۰/۸۴۵ = آلفای کرونباخ) و کل آزمون (۰/۹۲۴ = آلفای کرونباخ) با یافته‌های تحلیل راش هم‌خوانی دارد.

جدول ۱. آمار توصیفی آزمون ترجمه و آزمون چندگزینه‌ای با استفاده از مدل راش و بر مبنای لاجیت

بیشینه	کمینه	میانگین خطای استاندارد سنجش	میانگین دشواری	صورت سؤال
۳/۲۷	-۴/۳۳	۰/۱۵۱	۰/۱۹۳	ترجمه‌ای
۵۰/۲	-۲/۲۲	۰/۱۵۱	-۰/۴۵۷	چندگزینه‌ای

پس از بررسی فرض نرمال بودن داده‌ها با آزمون کلموگروف-اسمیرنوف تک‌نمونه‌ای و روی داده‌های برازش سؤال‌ها مشخص شد که داده‌های آماری سؤال‌های ترجمه‌ای و چندگزینه‌ای نرمال است: برای دشواری سؤال‌های ترجمه‌ای $P > ۰/۰۵$ ، $D(۲۸) = ۰/۱۰$ و برای

دشواری سؤال‌های چندگزینه‌ای $P > 0/05$ ، $D(32) = 0/12$. بنابراین، از آزمون پارامتریک تی مستقل برای مقایسه دشواری این دو صورت سؤال ترجمه‌ای و چندگزینه‌ای استفاده شد.

جدول ۲. آمار توصیفی دشواری دو صورت سؤال ترجمه‌ای و چهارگزینه‌ای مورد استفاده برای آزمون تی مستقل

گروه‌ها	تعداد سؤال‌ها	انحراف معیار	خطای میانگین استاندارد
دشواری سؤال ترجمه‌ای	۲۸	۱/۷۵۴۷۸	۰/۳۳۱۶۲
چندگزینه‌ای	۳۲	۱/۱۸۸۵۳	۰/۲۱۰۱۰

جدول ۲ آمار توصیفی آزمون تی و جدول ۳ نتایج آزمون تی را نشان می‌دهد. بر اساس نتایج جدول ۲ می‌توان گفت بین دو صورت سؤال از لحاظ دشواری تفاوت معناداری وجود ندارد ($t=696/1$ ، $df=58$ ، $P=0/095$ ، آزمون دو دامنه). در نتیجه می‌توان نتیجه گرفت میزان دشواری هر دو صورت سؤال با اندازه اثر کم ($f=0/22$) یکسان است.

جدول ۳. نتایج آزمون تی مستقل برای مقایسه دشواری دو صورت سؤال ترجمه‌ای و چهارگزینه‌ای

آزمون لونز برای برابری واریانس		آزمون تی برای برابری میانگین‌ها			
اف	سطح معناداری	تی	درجه آزادی	سطح معناداری (دو دامنه)	تفاوت میانگین استاندارد
۳/۲۸۶	۰/۰۷۵۰	۱/۶۹۶	۵۸	۰/۰۹۵۰	۰/۶۴۹۳۲۰
صورت با فرض سؤال برابری واریانس					

بحث و نتیجه‌گیری

هدف اصلی پژوهش حاضر بررسی تفاوت پایایی سؤال‌های چندگزینه‌ای و آزمون سؤال‌های ترجمه‌ای دانش دستوری بود. محاسبه پایایی این دو آزمون به روش آلفای کرونباخ حاکی از آن است که اگرچه پایایی آزمون چندگزینه‌ای حدود ۰/۰۴ کمتر از آزمون ترجمه‌ای است، این تفاوت چندان چشمگیر نیست و می‌توان هر دو آزمون را از نظر پایایی تقریباً یکسان تلقی کرد. به علاوه، تحلیل مدل راش (جدول ۱) نیز بیانگر این است که میانگین

خطای استاندارد سنجش برای آزمون ترجمه‌ای ۰/۱۵ و برای آزمون چندگزینه‌ای نیز ۰/۱۵ است. به عبارت دیگر، احتمال خطا در سنجش برای هر دو صورت سؤال بر مبنای لاجیت با ۶۸ درصد درجه اطمینان بین ۰/۱۵- و ۰/۱۵+، با ۹۵ درصد درجه اطمینان بین ۰/۳- و ۰/۳+ لاجیت، و با ۹۹ درصد درجه اطمینان بین ۰/۴۵- و ۰/۴۵+ است. در نتیجه، به نظر می‌رسد از این لحاظ تفاوتی میان این دو صورت سؤال وجود نداشته باشد. به علاوه، برآزش مطلوب و خطای پایین تحلیل مدل راش دلیلی مضاعف بر ویژگی پایایی سؤال‌های صورت ترجمه‌ای و قابل قیاس بودن آن با نوع چندگزینه‌ای است. در نتیجه، مقادیر پایایی محاسبه شده بر مبنای نظریه کلاسیک آزمون مؤید یافته‌های مدل راش مبنی بر پایایی آزمون طراحی شده برای پژوهش از یک سو و از سویی دیگر مشابهت بسیار زیاد دو صورت طراحی سؤال از نظر پایایی است.

هدف پرسش دوم پژوهش یافتن تفاوت دشواری بین آزمون سؤال‌های چندگزینه‌ای و آزمون سؤال‌های ترجمه‌ای دانش دستوری بود. طبق تحلیل مدل راش میانگین دشواری آزمون ترجمه حدود ۰/۶۵ لاجیت از آزمون چندگزینه‌ای بالاتر است. با توجه به اینکه میانگین خطای استاندارد سنجش برای هر دو صورت سؤال بر مبنای لاجیت با ۹۹ درصد درجه اطمینان بین ۰/۴۵- و ۰/۴۵+ و این مقدار از میزان ۰/۶۵ کمتر است، این امکان وجود دارد که از لحاظ دشواری میان این دو سؤال در این پژوهش تفاوت وجود داشته باشد (لینکر، ۲۰۱۰). برای اطمینان، از آزمون تی مستقل برای بررسی دشواری این دو آزمون دانش دستوری استفاده شد، ولی برخلاف انتظار، تفاوت این دو بخش آزمون از لحاظ دشواری معنادار نبود. از آنجا که این دو بخش آزمون کاملاً از لحاظ محتوا و سازه مورد سنجش شبیه هم بودند، تا جایی که سؤال‌ها در هر دو آزمون با هم متناظر و تفاوت آن‌ها فقط در صورت سؤال بود، می‌توان نتیجه گرفت که این دو صورت سؤال نه از لحاظ پایایی و نه از لحاظ دشواری برای سنجش دانش دستوری باهم تفاوت چندانی ندارد.

از جمله یافته‌های جنبی این پژوهش سهولت طراحی صورت ترجمه‌ای در مقایسه با صورت چندگزینه‌ای بود. پس از انجام پیش‌مطالعه، چهار سؤال ترجمه و نوزده سؤال چندگزینه‌ای اشکال‌دار شناسایی شد. حتی پس از سعی در رفع این اشکال‌ها، در مطالعه دوم فقط یک سؤال ترجمه‌ای در مقایسه با یازده سؤال چندگزینه‌ای ایراددار شناخته شد. این امر

خود بیانگر آن است که طراحی سؤال‌های ترجمه‌ای به مراتب آسان‌تر از طرح صورت چندگزینه‌ای است.

بر اساس یافته‌های این پژوهش، به نظر می‌رسد سؤال‌های ترجمه‌ای از نظر ویژگی‌های آماری پایایی و دشواری با سؤال‌های چندگزینه‌ای برابری می‌کند و حتی سریع‌تر و راحت‌تر از صورت چندگزینه‌ای طراحی می‌شود. با وجود این مزایا جای سؤال است که چرا از این صورت بهره‌چندانی در آموزش و سنجش زبان برده نمی‌شود. در پاسخ باید گفت در دوره قبل از قرن بیست از صورت ترجمه در آزمون‌های زبان استفاده می‌شد (اسپالسکی، ۲۰۱۰). ولی چون نمره‌دهی به ترجمه معیاری عینی نداشت، جایگاه خود را در فضای رویکرد روان‌سنجی از دست داد. برخلاف تفکر رایج که سؤال‌های عینی را چندگزینه‌ای می‌داند، باید گفت سؤال‌های عینی لزوماً به صورت چندگزینه‌ای نیست. به نظر الدرسون، کلپهام، و وال (۱۹۹۵) سؤال‌هایی که پاسخ آن‌ها عبارت، کلمه، یا عددی خاص باشد، یا در نحوه نمره‌دهی آن‌ها دستورالعمل مشخصی وجود داشته باشد، به گونه‌ای که پاسخ افراد را بتوان به صورت درست یا غلط ارزش‌گذاری کرد، در دایره سؤال‌های عینی می‌گنجد. بنابراین، آزمون ترجمه با توجه به ملاک‌های معین به صورت عینی ارزیابی می‌شود. در نتیجه، می‌توان با تدوین دستورالعمل‌های در اختیار، به مانند دستورالعملی که جزو سند مشخصات آزمون این پژوهش بود، یا با تربیت مصحح، نمره‌دهی به سؤال‌های ترجمه‌ای را عینی‌تر کرد.

به نظر می‌رسد، افول استفاده از ترجمه در آموزش زبان دلایل دیگری داشته است. از جمله اینکه بنا به نظر کوک (۲۰۱۰) و ویدوسن (۲۰۰۳) روش‌های تدریس تک‌زبانی بیشتر توجیه اقتصادی و سیاسی دارند تا توجیه علمی؛ برای مثال، اشتغال‌زایی برای معلمان اهل کشور زبان مبدأ و کاهش هزینه‌های تولید کتاب‌های آموزشی با تولید کتاب برای فراگیران با زبان‌های مادری متفاوت. یکی دیگر از دلایل افول استفاده از ترجمه در آموزش زبان، موضع‌گیری کسانی است که آموزش تک‌زبانی را برتر می‌دانند. ویدوسن (۲۰۰۳) بار د برتری آموزش تک‌زبانی معتقد است در فرایند یادگیری زبان دوم باید زبان مادری و زبان دوم را در ذهن فراگیر به هم نزدیک ساخت و رابطه و تفاوت‌های آن‌ها را تبیین کرد. وی با رد فرض ایجاد اختلال زبان مادری در یادگیری زبان خارجی، بیان می‌کند هر قدر هم روش‌های آموزش

زبان سعی کنند افراد طی فرایند دوزبانگی هم‌پایه^{۲۳} زبان دوم را فراگیرند، یعنی فرایندی که دو نظام معنایی متفاوت در ذهن افراد به صورت هم‌پایه شکل می‌گیرد، زبان‌آموزان در ذهن خود فرایند دوزبانگی ترکیبی^{۲۴} را با گذر از دوره‌های مختلف زبان‌های میانی^{۲۵} فرامی‌گیرند. البته، ممکن است در نهایت به مرحله دوزبانگی هم‌پایه نیز برسند، ولی اساس یادگیری آن‌ها دوزبانگی ترکیبی است. این دیدگاه با نظر اسپالسکی (۱۹۹۸: ۴۹) نیز هم‌خوانی دارد جایی که می‌گوید: «پدیده دوزبانگی مثال بارزی از تماس زبانی است، که دو زبان در فرد دو زبانی در تماس‌اند.»

کوک (۲۰۱۰) نیز معتقد است روش‌های تدریس دوزبانی و استفاده از ترجمه نقش مهمی در یادگیری زبان بر عهده دارد و باعث افزایش آگاهی زبانی و عامل بالابردن کیفیت تدریس می‌شود. اگر بپذیریم که بنا به نظر کوک (۲۰۱۰) و ویدوسن (۲۰۰۳) ترجمه نقش بی‌بدیلی در افزایش آگاهی زبانی افراد دارد، می‌توان از آزمون‌های ترجمه‌ای برای سنجش این آگاهی بهره جست، به‌خصوص در رویکرد ارزیابی برای یادگیری^{۲۶}. در این رویکرد نسبتاً جدید سنجش که به صورت روزافزون مورد توجه محققان است (ترنر، ۲۰۱۲)، به آن نوع ارزیابی بها داده می‌شود که گامی در راستای آموزش باشد. این امر زمانی ممکن است که سنجش، اطلاعاتی را در اختیار معلمان و فراگیران بگذارد تا بتوانند خود را ارزیابی و فعالیت‌های بعدی کلاس را بر اساس آن تنظیم کنند (بلک و همکاران، ۲۰۰۴). کارگروه اصلاحات در ارزیابی (۲۰۰۲: ۲-۳) نیز ارزیابی برای یادگیری را این گونه تعریف می‌کنند: «فرایند جست‌وجو و تفسیر دلایل و شواهد برای استفاده فراگیران و آموزگاران‌شان تا تصمیم بگیرند که فراگیران در چه مرحله‌ای از یادگیری خودند، به کجا باید برسند و بهترین حالت ممکن برای رسیدن به آن چیست.» بنابراین، در رویکردهای جدید سنجش کلاسی که در آن تشخیص نقاط ضعف و قوت فراگیران اقدامی اساسی است (ویلیام، ۲۰۱۱)، می‌توان از صورت سؤال ترجمه‌ای برای طراحی آزمون‌های تشخیصی استفاده شایانی کرد، زیرا این نوع سؤال‌ها دریچه‌ای است تا معلم بتواند وضعیت تطابق ساختارهای زبان اول و زبان خارجی در ذهن زبان‌آموز را دقیق‌تر رصد کند و ایرادات وی را بهتر تشخیص دهد.

23. bilingualism coordinate

24. bilingualism compound

25. interlanguage

26. assessment for learning

نتیجه‌گیری

به نظر می‌رسد صورت سؤال ترجمه‌ای قابلیت لازم برای استفاده به منزله ابزار سنجش دانش دستوری به‌خصوص در رویکرد سنجش برای آموزش را داشته باشد، زیرا از یک طرف از نظر مزایا و ویژگی‌های مهم در سنجش همچون پایایی در حد سؤال چندگزینه‌ای است؛ و از طرفی دیگر، به دلیل مطابق بودن با فرایند یادگیری زبان به صورت دوزبانگی ترکیبی فرصتی را فراهم می‌آورد که زبان‌آموزان در ذهن خود دانش دستوری خویش از دو زبان را با هم مقایسه کنند و از این راه آگاهی زبانی خود را ارتقا دهند. بنابراین، آن‌ها و استادانشان می‌توانند به ضعف و قوت خود پی ببرند و برای آن چاره‌اندیشی کنند. از لحاظ رویکرد نظام‌مند در طراحی سؤال نیز صورت ترجمه‌ای در کنار صورت‌های دیگر ممکن است به چند بعدی شدن آزمون کمک کند. برای مثال، سؤال‌های چندگزینه‌ای تک‌زبانی دانش دستوری صرفاً آن بخش از دانش فرد را می‌سنجد که مربوط به زبان دوم می‌شود و سؤال‌های ترجمه‌ای آگاهی فرد از تفاوت‌ها و شباهت‌های زبان مادری و زبان دوم را می‌سنجد. از آنجا که هر دو صورت سنجش مزیت‌های خاص خود را دارد، توصیه می‌شود معلمان در رویکردی نظام‌مند در طراحی سؤال به تناسب نیاز فراگیران از صورت‌های گوناگون سؤال استفاده کنند. در پایان، باید متذکر شد صورت سؤال ترجمه‌ای و چندگزینه‌ای در این پژوهش برای سنجش دانش دستوری به‌کار گرفته شده است و بررسی مزایا یا ضعف‌های صورت ترجمه‌ای در سنجش دیگر دانش‌ها و مهارت‌های زبانی، نیازمند تحقیقات بیشتر است.

References

- Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press: Cambridge.
- Assessment Reform Group (2002). *Assessment for learning: 10 principles*. Retrieved November 2012 from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf. Cambridge, UK: University of Cambridge. School of Education.
- Bachman, L. F., and Palmer, A. S. (2010). *Language Testing in Practice*: Oxford: Oxford University Press.
- (1996). *Language Testing in Practice*: Oxford: Oxford University Press.

- Black, P., Harrison, C., Lee, C., Marshall, B., and William, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, 86(1), 8–21.
- Brown J.D. and Hudson, T.D. (1998). The alternatives in language assessment. *TEOSL Quarterly*, 32 653-75.
- Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.
- Cook, G. (2010). *Translation in Language Teaching*. Oxford: Oxford University Press.
- Davidson, F. (2012). Releasability of Language Test Specifications. *Japan Language Testing Association (JLTA) Journal*, 15, 1-23.
- Davidson, F. and Lynch, B.K. (2002). *Testcraft: A teacher's guide to writing and using language test specification*. New Haven: Yale University Press.
- DeMars, C. (2010). *Item Response Theory (Understanding statistics: measurement)*. Oxford University Press: Oxford.
- Farhady, H. (2007). Teaching and testing EFL in Iran: global trends and local dilemmas. *TELL*, 2 75-98.
- Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. Oxford: Routledge.
- Johnson, K. and Johnson, H. (1999). *Encyclopedic dictionary of applied linguistics*. Oxford: Blackwell Publishers.
- Kennedy, C.A., Wilson, M., Draney, K., Tutunciyani S., and Vorp, R. (2006). *ConstructMap software*. Berkeley Evaluation and Assessment Research (BEAR) Center. Berkeley, CA.
- Li, J. (2006). *Introducing audit trails to the world of language testing*. Unpublished master's thesis. University of Illinois at Urbana-Champaign, USA: Division of English as an International Language.
- Linacre, J.M. (2010). *A User's Guide to WINSTEPS®*. Retrieved May 2, 2010 from [http://www.winsteps.com/Linacre, J.M. \(2010b\) Winsteps® \(Version 3.70.0\) \[Computer Software\]. Beaverton, Oregon: Winsteps.com](http://www.winsteps.com/Linacre, J.M. (2010b) Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com).
- (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16:2 p.878.
- (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* 1994 7:4.
- McNamara, T.F. (1996). *Measuring second language performance*. Longman: Harlow.

- McNamara, T. and Roever, C. (2006). *Language testing: the social dimension*. London, UK: Blackwell Publishing.
- Morizot, J., Ainsworth, A.T. and Reise, S.P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.
- Shin, D. (2012). Item writing and writers. In G, Fulcher and F, Davidson (Eds.): *Routledge handbook of language testing*. London: Routledge.
- Spolsky, B. (2010). Language testing in historical and future perspective. In E., Shohamy and N., Hornberger (eds). *Encyclopedia of language and education* (2nd edition) Volume 7: Language testing and assessment: Springer: New York.
- (1998). *Sociolinguistics*. Oxford: Oxford university press.
- Turner, C.E. (2012). Classroom assessment. In G, Fulcher and F, Davidson (Eds.): *Routledge handbook of language testing*. London: Routledge.
- Widdowson, H.G. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation* (37), 3-14.