T. M .U.

Language Related Research
E-ISSN: 2383-0816
Vol.11, No.4 (Tome 58),
September, October & November 2020

The Validity of Speaking Scoring Rubric in Ferdowsi Persian Proficiency Test

Mohsen Roudmajani[1*], Ehsan Ghabool[2]

1. PhD Candidate in Persian Language and Literature, Ferdowsi University, Mashhad, Iran.
2. Assistant Professor of Persian Language and Literature, Ferdowsi University, Mashhad, Iran.

Received: 1/09/2019
Accepted: 11/01/2020

* Corresponding Author's E-mail: mroodmajani@gmail.com

Abstract

The ability to speak is an important part of every body's language proficiency. This ability plays an important role in the academic life of students. But scoring and assessing speaking is not easy. In this research, we try to study the validity of Ferdowsi University's Persian proficiency test. We know that every test has a certain amount of error; but in scoring speaking ability if the scoring rubric is designed in a scientific way, the score attributed to the speakers' speech ability is likely to be very similar to their actual language ability. In other words, the appropriate scoring rubric can have a significant effect on reducing the error rate of the test. In norm-reference tests, this can be achieved only when test designers can say what scoring constructs they intend to measure and how successful they are in achieving that goal. Also, it should be clear whether the scoring scale can distinguish weak, medium, and strong test takers. On the other hand, in applying the scoring rubric , the level of consensus of the scorers should be clear. In order to see how successful is the scoring rubric in Ferdowsi Persian proficiency test, in measuring the test taker's speaking ability, the authors analyzed the result of one of the proficiency tests administered at Ferdowsi University with Rasch model and factor analysis. The result showed that scorer reliability is 0.97 which is so high. It showed that scorers have the same understanding of the scoring rubric. This means that the scorers have given the test takers a relatively stable score, which is a strong point for the test. Also, the scores have used the scoring rubric properly because the cut score goes up in an organized way as the ability of test-takers increase. Each of the four thresholds obtained by the Rash statistical model differs by approximately 5 degrees, respectively. A

T. M .U.

Language Related Research
E-ISSN: 2383-0816
Vol.11, No.4 (Tome 58),
September, October & November 2020

regular increase in thresholds is commensurate with the ability of the test takers. This indicates a correct understanding of the scorers of the 5 grades specified in the scoring rubric; in other words, scorers have a good understanding of the level of competence of test takers and its relationship with the grades in the scoring rubric. The Wright map shoes that the scoring rubric can differentiate basic, intermediate and advanced test-takers well. Although on the top of the map there are 8 test-takers which there is no score for them that means the needs some higher scores for them. On the other hand, factor load for three constructs, delivery, language use and topic development are 0.74, 0.78 and 0.76. This shows that dividing speaking ability into these three constructs is proper while language use has the highest factor load and topic development has the lowest factor load.

Keywords: Language assessment, Speaking skill, Scoring rubric, Rasch, Factor analysis, Ferdowsi Persian proficiency test, Validity.

دوماهنامهٔ علمی- پژوهشی
د۱۱، ش٤ (پیاپی ۵۸)، مهر و آبان ۱۳۹۹، صص۱۸۳-۲۱۰

اعتبارسنجی معیارنمره‌دهی مهارت صحبت کردن در آزمون جامع زبان فارسی فردوسی

محسن رودمعجنی[۱*]، احسان قبول[۲]

۱. دانش‌آموختهٔ دکتری گروه زبان و ادبیات فارسی دانشگاه فردوسی، مشهد، ایران.
۲. استادیار گروه زبان و ادبیات فارسی دانشگاه فردوسی، مشهد، ایران.

دریافت: ۹۸/۰۶/۱۰ پذیرش: ۹۸/۱۰/۲۱

چکیده

مهارت صحبت کردن بخش بسیار مهمی از توانایی زبانی افراد را دربر می‌گیرد. بهره‌مندی از این مهارت در محیط دانشگاه نیز اهمیت بسزایی دارد؛ اما سنجش صحبت کردن کار چندان ساده‌ای نیست و با مشکلاتی مانند دشواری در نمره‌دهی روبه‌رو است. در این پژوهش تلاش شده است تا میزان اعتبارِ معیار نمره‌دهی مهارت صحبت کردن در آزمون جامع زبان فارسی مرکز بین‌المللی دانشگاه فردوسی مشهد مطالعه شود. به همین منظور، نتایج به‌دست آمده از یکی از آزمون‌های برگزارشده در این مرکز به‌وسیلهٔ مدل‌های آماری راش و تحلیل عاملی بررسی شد. نتایج نشان داد که پایایی آزمون‌گیرنده ۹۷ درصد است. این عدد بیانگر درک نسبتاً یکسان آزمون‌گیرندگان از معیار نمره‌دهی است. همچنین، در این آزمون، آزمون‌گیرندگان توانسته‌اند به‌شکل مناسبی مقیاس نمره‌دهی را برای آزمون‌دهندگان با توانایی‌های مختلف به‌کار گیرند؛ زیرا آستانه‌های به‌دست‌آمده بر اساس مدل راش، سیر صعودی منظمی داشته‌اند. نقشهٔ آزمون‌دهنده – پرسش نیز نشان می‌دهد که معیار نمره‌دهی توانایی تمییز زبان‌آموزان ضعیف، متوسط و قوی از یکدیگر را داشته است. با این حال، در بالای طیف توانمندی آزمون‌دهندگان، هشت آزمون‌دهنده قرار گرفته است که هیچ نمره‌ای متناسب با سطح توانمندی‌شان دیده نمی‌شود؛ یعنی معیار نمره‌دهی در تمییز آن‌ها کارآمد نبوده است. از سوی دیگر، بار عاملی به‌دست آمده برای سه سازهٔ شیوهٔ بیان، کیفیت زبان و بسط موضوع به‌ترتیب ۷۶، ۷۸ و ۷٤ درصد بوده است. این امر نشان می‌دهد تقسیم توانایی صحبت کردن به سه عامل یاد-شده متناسب و دقیق است و هر کدام از این سازه‌ها توانمندی متفاوتی را سنجش می‌کنند. از این میان کیفیت زبان بیشترین و سازهٔ بسط موضوع، کم‌ترین میزان بار عاملی را داشته‌اند.

---

* نویسندهٔ مسئول مقاله: E-mail:mroodmajani@gmail.com

واژه‌های کلیدی: اعتبارسنجی، مهارت صحبت کردن، معیار نمره‌دهی، راش، تحلیل عاملی، آزمون جامع زبان فارسی فردوسی.

۱. مقدمه

بخش بسیار مهمی از مهارت زبانی افراد را توانایی صحبت کردن آن‌ها تشکیل می‌دهد. اهمیت این مهارت تا حدی است که برای بسیاری از افراد دانستن یک زبان، معادل صحبت کردن به آن زبان است. بسیاری از فعالیت‌های محیط دانشگاه نیز با توانمندی صحبت کردن درهم تنیده است. از آن‌جمله، می‌توان به ارائه و بحث‌های کلاسی، گفت‌وگو در اتاق استادان، گفت‌وگو در بخش اداری دانشگاه، سالن غذاخوری و یا خوابگاه اشاره کرد. از همین رو، هرگونه سنجش بسندگی زبان دانشجویان بدون درنظر گرفتن مهارت صحبت کردن آن‌ها ناقص خواهد بود؛ اما سنجش این مهارت کار چندان ساده‌ای نیست و با مشکلاتی مانند پایایی در نمره‌دهی به عملکرد آزمون-دهندگان روبه‌رو است.

می‌دانیم که هر آزمونی دارای میزانی از خطاست؛ اما در سنجش مهارت صحبت کردن اگر معیار نمره‌دهی به‌شکلی علمی طراحی شده باشد، احتمالاً نمره‌ای که به توانمندی گفتاری آزمون-دهندگان نسبت داده می‌شود، بسیار شبیه به توانمندی حقیقی زبانی آن‌ها خواهد بود. به عبارتی دیگر، معیار نمره‌دهی مناسب می‌تواند در کاهش میزان خطای آزمون تأثیر چشمگیری داشته باشد. در آزمون‌های هنجارمحور[۱]، رسیدن به این مهم زمانی می‌تواند امکان‌پذیر باشد که طراحان معیار بتوانند برای سه پرسشی که آورده می‌شود، پاسخ‌هایی قانع‌کننده ارائه کنند:

۱. معیار نمره‌دهی قصد دارد چه سازه‌هایی را اندازه بگیرد و تا چه میزان در رسیدن به این هدف خود موفق عمل کرده است؟

۲. آیا مقیاس نمره‌دهی[۲] به‌کار رفته، توانایی تمییز آزمون‌دهندگان ضعیف، متوسط و قوی از یکدیگر را دارد؟

۳. تا چه میزان نمره‌دهندگان در به‌کارگیری معیار نمره‌دهی، با یکدیگر اتفاق نظر دارند؟

سه پرسش ذکرشده مربوط به میزانِ اعتبارِ معیارِ نمره‌دهی مهارت صحبت کردن در آزمون جامع زبان فارسی فردوسی می‌شود. این آزمون هر ساله در دو نوبت، یک‌بار در اواسط تابستان و بار دیگر در اوایل فصل زمستان در دانشگاه فردوسی مشهد و چند شهر کشور عراق برگزار

می‌شود. هدف از این آزمون، سنجش میزان توانمندی متقاضیان ورود به دانشگاه فردوسی در انجام فعالیت‌های زبانی دانشگاهی است. از این رو، نتایج به‌دست آمده از یکی از آزمون‌های برگزارشده در مرکز بین‌المللی زبان فارسی فردوسی از سوی مدل‌های آماری راش[۳] و تحلیل عاملی بررسی خواهد شد. می‌دانیم که مدل آماری راش ـ که ساختاری پیچیده‌تر از مدل‌های کلاسیک دارد ـ یک مزیت اساسی دارد و آن وابسته نبودن به نمونهٔ آماری پژوهش است (Bachman, 2004: 139).

۲. پیشینهٔ پژوهش

در زبان فارسی پژوهش‌های چندانی در ارتباط با معیارهای نمره‌دهی مهارت صحبت کردن صورت نگرفته است و این شاخه از علم سنجش زبان تقریباً ناکاویده باقی مانده است؛ ازجمله معدود پژوهش‌های این حوزه می‌توان به جلیلی (۱۳۹۰) و گلپور (۱۳۹٤) اشاره کرد که معیارهایی برای نمره‌دهی مهارت صحبت کرن غیرفارسی‌زبانان ارائه کرده‌اند. معیار جلیلی (۱۳۹۰: ۱٤۷) از پنج مؤلفهٔ تلفظ، دستور زبان، دایرهٔ لغات، روانی و درک تشکیل شده است. هر کدام از آن مؤلفه‌ها نیز خود به بخش‌های دیگری تقسیم می‌شود و درمجموع، ۲۵ نمره را دربر می‌گیرند. گلپور نیز که مدل براون (2001: 406)را اساس کار خود قرار داده است، مؤلفه‌هایی همانند جلیلی دارد. دربارهٔ این دو پژوهش، سه نکته مهم است: نخست اینکه میزان کارایی هیچ‌کدام از آن‌ها در زبان فارسی در عمل مشخص نشده و پژوهش کمّی بر روی آن‌ها صورت نگرفته است. به‌عبارتی، ما نمی‌دانیم آیا این معیارها همان سازه‌هایی را که ادعا کرده‌اند اندازه‌گیری می‌کنند یا خیر. از سوی دیگر، نمی‌دانیم که این معیارها تا چه اندازه می‌توانند آزمون‌دهندگان ضعیف، متوسط و قوی را از یکدیگر متمایز کنند. همچنین، هیچ‌کدام از این پژوهش‌ها نشان نداده‌اند که تا چه میزان ارزیاب‌ها می‌توانند معیارهای معرفی‌شده را به‌شکلی پایا به‌کار گیرند.

درمقابل، در زبان انگلیسی سنجش مهارت صحبت کردن حوزهٔ گسترده‌ای از پژوهش‌ها را دربر می‌گیرد که از آن جمله می‌توان به تلاش‌های لولت[٤](1989)، دبات[۵] (1992) و دن داگلاس[٦] (1997) در توصیف فرایند تولید گفتار اشاره کرد. درحقیقت، مدل‌های دبات و دن داگلاس بازبینی‌هایی از همان مدل اولیهٔ لولت هستند که مراحل تولید گفتار را به‌شکلی مرحله به مرحله و

موازی به تصویر می‌کشد. از آنجا که مدل لولت برای افراد تک‌زبانه طراحی شده است. دبات با اندک تغییرات نشان می‌دهد که چگونه می‌توان از آن برای توصیف گفتار دوزبانه‌ها استفاده کرد. دن داگلاس نیز از آن برای شرح مبانی نظری سنجش زبان انگلیسی به‌منزلهٔ زبان دوم در محیط دانشگاه استفاده می‌کند. این مدل‌ها که محصول سال‌ها پژوهش هستند، تصویر بسیار راهگشایی از سازه‌های تشکیل‌دهندهٔ فرایند تولید گفتار ارائه می‌دهند. از آنجا که که مدل دن داگلاس مبنای طراحی آزمون تافل و نیز آزمون بسندگی فارسی فردوسی بوده است، در بخش مبانی نظری به تفصیل به‌شرح آن خواهیم پرداخت.

از سوی دیگر، در زبان انگلیسی برخی از پژوهش‌ها تلاش کرده‌اند تا بر اساس مبانی نظری شناخته‌شده به تعریف سازه‌های تشکیل‌دهندهٔ مهارت صحبت کردن بپردازند. شاید جدی‌ترین پژوهش در این حوزه اثر گلن فولچر[۷] (2014) با عنوان ***سنجش صحبت کردن در زبان دوم***[۸] باشد. او توانش زبانی، توانش راهبردی، دانش متنی، دانش کاربردی و دانش جامعه‌شناختی را به‌منزلهٔ سازه‌های مهارت صحبت کردن معرفی می‌کند. پیشقدم[۹] (2013) نیز تلاش می‌کند تا با بررسی معیار نمره‌دهی آزمون آیلتس تعریفی جدید از سازه‌های مهارت صحبت کردن ارائه دهد. او در کنار سازه‌های متداول دو سازهٔ هوش روایی و هوش بیانی را نیز برمی‌شمرد.

پژوهش‌های متعدد دیگری نیز در زبان انگلیسی به بررسی انواع تکلیف‌های زبانی و ارزیاب‌ها و تأثیر آن‌ها در سنجش مهارت صحبت کردن پرداخته‌اند. تارون[۱۰](1983)، الیس[۱۱] (1985, 1987, 1990) و همچنین، لارسن فریمن[۱۲] و لانگ[۱۳] (1991) به این نتیجهٔ رسیده‌اند که نوع تکلیف در نمرهٔ مهارت صحبت کردن اثرگذار است. از همین رو، توصیهٔ این پژوهشگران بر این است که تا حد امکان در سنجش مهارت صحبت کردن از تکلیف‌های متنوع استفاده شود. همچنین، هدن[۱۴] (1991) و لودویگ[۱۵] (1982) نشان داده‌اند که ارزیاب‌های بی‌تجربه (بدون تجربهٔ تدریس) عموماً بر محتوا تمرکز دارند و درمقابل، ارزیاب‌های باتجربه بر ساخت‌های زبانی مانند دستور و تلفظ.

شاید یکی از اثرگذارترین پژوهش‌ها در ارتباط با اعتبارسنجی آزمون‌های زبانی را بتوان اثر چپل[۱۶] و همکاران (2008) دانست. در این پژوهش که خود متشکل از مقاله‌ها و پژوهش‌های متعددی است، ساخت اعتباری آزمون تافل بررسی شده است. بخشی از آن نیز به مهارت صحبت

کردن اختصاص دارد. در بخش مبانی نظری به معرفی ساختار اعتباری این پژوهش خواهیم پرداخت.

همان‌طور که گفته شد، در زبان فارسی تحقیقات چندانی در زمینهٔ زبان فارسی صورت نگرفته است. پژوهش‌های موجود نیز نمی‌تواند جوابگوی نیاز این شاخهٔ مهم از علم زبان‌شناسی کاربردی باشد. مشخص کردن نیاز زبانی فارسی‌آموزان در سطوح مختلف، رابطهٔ انواع تکلیف‌های زبانی و ارزیاب‌ها و تأثیر آن در نمره‌دهی مهارت صحبت کردن و نیز طراحی علمی معیارهای نمره‌دهی ازجمله نیازمندی‌های این حوزه است که پژوهشگران برای رسیدن به سنجش دقیق باید به آن توجه کنند.

۳. مبانی نظری

در این بخش از پژوهش سعی خواهیم کرد در وهلهٔ نخست، تعریفی از مفهوم اعتبار در سنجش زبان ارائه دهیم و سپس به بررسی سازه‌های تشکیل مهارت صحبت بپردازیم. همان‌طور که گفته شد، هدف از این پژوهش، اعتبارسنجی معیار نمره‌دهی مهارت صحبت کردن در آزمون بسندگی زبان فارسی در دانشگاه فردوسی مشهد است. از آنجا که تعاریف مختلفی از مفهوم اعتبار وجود دارد، لازم است که منظور از این مفهوم در پژوهش حاضر مشخص شود. همچنین، تمامی پرسش‌های پژوهش به‌نوعی با مفهوم سازه‌های مهارت صحبت کردن در ارتباط هستند. از همین رو، ارائهٔ تعریفی از این سازه نیز الزامی به‌نظر می‌رسد.

۳ـ ۱. مفهوم اعتبار

اعتبار یکی از محوری‌ترین و بنیادی‌ترین مفاهیم در حوزهٔ سنجش زبان است (Fulcher & Davidson, 2007: 3; Lissitz, 2009: 1) که بر تمامی مراحل طراحی آزمون سایه می‌افکند. بدون داشتن درک کاملی از مفهوم اعتبار، طراحی علمی یک آزمون ناممکن است. فرایند اعتبار همانند پلی است که عملکرد آزمون‌دهندگان در جلسهٔ امتحان را به بخش پایانی طراحی آزمون، یعنی تفسیر نمره‌ها و تصمیم‌گیری پیوند می‌دهد. اعتبار قضاوتی است ارزش‌گذارانه و یکپارچه از اینکه تا چه میزان مستندات تجربی و مبانی نظری، بسنده بودن و مناسب بودن استنباط‌ها و اقدامات صورت‌گرفته را بر مبنای نمره‌های آزمون تأیید می‌کنند (Messick, 1987: 1).

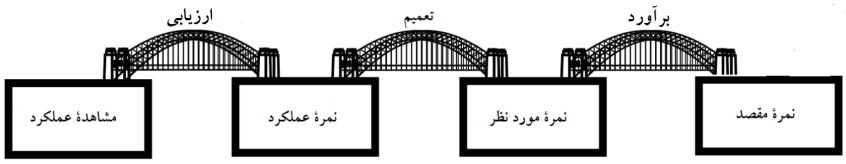
آزمون‌گرها به‌طور مستقیم به توانایی زبانی و دانش آزمون‌دهندگان دسترسی ندارند و تنها از طریق مشاهدهٔ عملکرد آن‌ها در امتحان می‌توانند از آن آگاه شوند. نمی‌توان مغز یک فرد را شکافت و دانش او را مشاهده کرد. از همین رو، تنها وسیله برای اطلاع از توانایی و دانش آزمون‌دهنده، مطالب موجود در برگهٔ کاغذ و یا آزمون شفاهی است؛ اما می‌دانیم که عملکرد آزمون‌دهندگان در جلسهٔ امتحان محدود است و شامل تمامی توانایی آن‌ها نمی‌شود. بنابراین، سنجش زبان همواره نیازمند تعمیم و تفسیر است. آزمونی دارای میزان اعتبار بالایی است که برای تفاسیر خود دلایل و مستندات قوی ارائه می‌دهد.

کین (1999) نگاه دقیق‌تری به فرایند تفسیر آزمون دارد. او معتقد است آنچه مشاهدهٔ عملکرد آزمون‌دهندگان در جلسهٔ امتحان را به نتیجه‌گیری نهایی بر مبنای نمرات آزمون پیوند می‌دهد، زنجیره‌ای به‌هم پیوسته متشکل از سه تفسیر و یا استنباط[۱۷] است که آن‌ها را به‌ترتیب ارزیابی[۱۸]، تعمیم[۱۹] و برآورد[۲۰] می‌نامد. استنباط نخست زمانی صورت می‌گیرد که عملکرد مشاهده‌شده را به عدد تبدیل می‌کنیم. پرسش اساسی در این بخش این است که آیا فرایند تبدیل عملکرد به عدد به‌شکل دقیق و سازمان‌یافته‌ای صورت گرفته است. همان‌طور که گفته شد، این استنباط ارزیابی نامیده می‌شود.

از سوی دیگر، ازجمله شرط‌های اساسی گرفتن تصمیمات عادلانه بر مبنای نمره‌های آزمون این است که بتوانیم عملکرد آزمون‌دهندگان در جلسهٔ امتحان را به عملکرد آن‌ها در رویارویی با فعالیت‌های مشابه تعمیم دهیم (Kane, 1999: 9). اگر آزمون توانمندی حقیقی داوطلبان را سنجیده است، پس باید بتوان این عملکرد را به سایر تکالیف[۲۱] مشابه بسط داد. بنابراین، سومین استنباطِ فرایند اعتبار آزمون، تعمیم عملکرد آزمون‌دهندگان به فعالیت‌های زبانی مشابه است. طراحان یک آزمون در این بخش با مسائلی مربوط به پایایی آزمون روبه‌رو هستند.

بر مبنای الگوی کین، مرحلهٔ پایانی فرایند اعتبار آزمون، برآورد نام دارد. برآورد به این معناست که طراحان می‌توانند بر اساس عملکرد آزمون‌دهندگان در جلسهٔ امتحان، عملکرد آن‌ها را در محیط مقصد پیش‌بینی کنند (*ibid*). برای مثال، با توجه به عملکرد یک فرد در آزمون تافل می‌توان عملکرد او در محیط دانشگاه‌های امریکای شمالی را پیش‌بینی کرد؛ اما تا چه میزان نمرهٔ تافل گویای توانمندی حقیقی آزمون‌دهنده است. هر چه نمرهٔ به‌دست‌آمده به توانمندی

آزمون‌دهنده در محیط دانشگاه شبیه‌تر باشد، آزمون دارای میزان برآورد بالاتری است. وی مجموعهٔ این تفاسیر را به‌شکل زنجیره‌ای از سه پلِ به‌هم پیوسته، به‌تصویر کشیده است. هر آزمون زمانی دارای میزان اعتبار بالایی است که هر یک از این پل‌ها استوار باشد؛ یعنی بتوان دلایل و مستندات کافی برای دفاع از استنباط‌های آزمون ارائه داد.



شکل ۱: تمثیل پل(*ibid*)

Figure 1: Bridge allegory

با این حال، طراحان تافل بخش‌های دیگری به مدل کین اضافه کرده‌اند. یکی از آن‌ها استنباطی با نام تبیین[۲۲] است. می‌دانیم که آزمون‌های بسندگی بر مبنای یک نظریهٔ زبانی طراحی می‌شوند. بنابراین، ازجمله نشانه‌های دیگر اعتبار یک آزمون این است که بتوان عملکرد آزمون‌دهندگان را بر مبنای سازه‌های تعریف‌شده، تبیین کرد (Chapelle et al., 2008: 12). این مرحله از مفهوم اعتبار یک آزمون مربوط به پیوند مبانی نظری و عملکرد آزمون‌دهندگان در جلسهٔ امتحان می‌شود. در این بخش عملکرد داوطلبان بر مبنای مفهوم بسندگیِ تعریف‌شده، تفسیر می‌شود. در این بخش طراحان آزمون با مسائل مربوط به اعتبار سازهٔ آزمون روبه‌رو هستند.

طبیعی است که بررسی کردن مجموعهٔ استنباط‌های هر آزمون نیاز به پژوهش‌های بسیار گسترده‌ای دارد. در پژوهش حاضر، هدف ما بررسی برخی از پیش‌فرض‌ها مربوط به استنباط‌های ارزیابی، تعمیم و تبیین می‌شود.

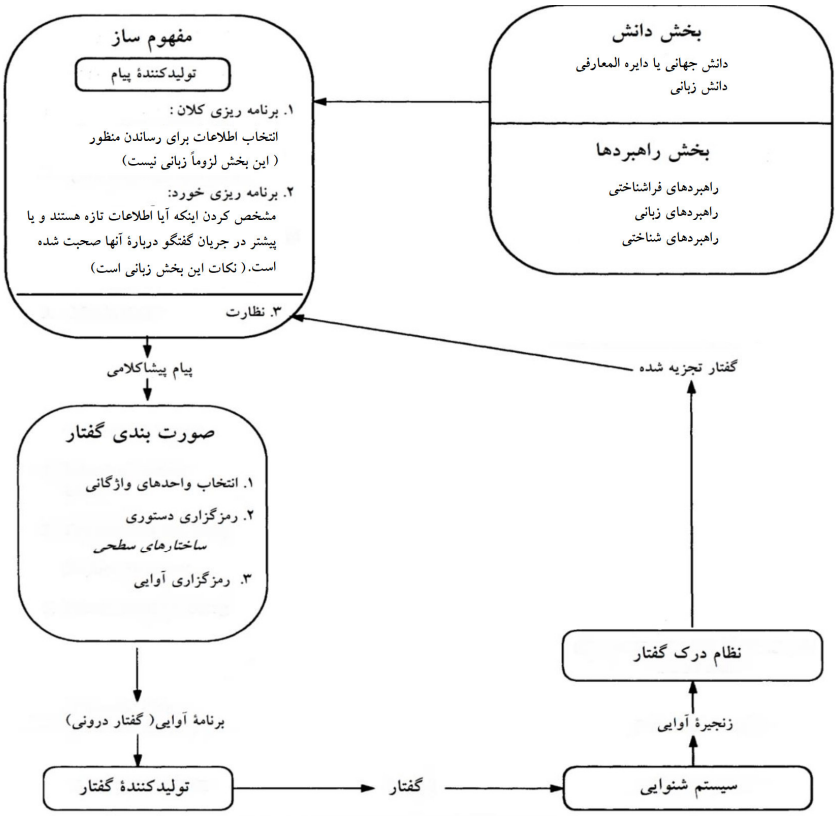
۳ـ۲. سازه‌های تشکیل‌دهندهٔ مهارت صحبت کردن

برای دستیابی به سنجش دقیق از مهارت صحبت کردن، باید بدانیم که این مهارت چیست و از چه سازه‌هایی تشکیل شده است. دن داگلاس[۲۳] (1997) تصویر روشنی از فرایند تولید گفتار ارائه می‌دهد که می‌تواند در شناخت سازه‌های این مهارت و ارتباط آن‌ها با یکدیگر بسیار راهگشا

باشد. بر اساس این مدل ـ که خود برگرفته از مدل لولت (1989) و دبات است ـ ابتدا در بخشی با نام مفهوم‌ساز[۲۴] برنامه‌ریزی اولیه برای گفتار صورت می‌گیرد و دانش مورد نیاز برای انجام تکلیف شناسایی می‌شود. دانش جهانی[۲۵]، یعنی دانش فرد از اشیا، اشخاص، مکان‌ها و نیز نحوهٔ انجام امور، در کنار دانش زبانی، منابع اطلاعاتی مورد نیاز برای شکل‌گیری برنامهٔ اولیه را فراهم می‌کنند. توانش راهبردی که شامل شناسایی موقعیت، هدف‌گذاری، برنامه‌ریزی و نظارت می‌شود، نقش حیاتی در شکل‌گیری پیام پیشاکلامی[۲۶] دارد.

محصول بخش مفهوم‌ساز هرچند تمام اطلاعات مورد نیاز برای تبدیل معنا را به زبان دارد، خود زبانی نیست (Debot, 1992: 4). تنها در فرایند صورت‌بندی[۲۷] است که پیام تبدیل به طرح آوایی[۲۸] می‌شود؛ یعنی واژگان مناسب انتخاب می‌شوند و قوانین دستوری و آوایی نیز به‌کار گرفته می‌شوند تا یک طرح آوایی شکل بگیرد که به آن گفتار درونی می‌گویند. نظام تولیدکنندهٔ گفتار طرح آوایی را به گفتار حقیقی تبدیل می‌کند، سپس این گفتار از سوی سیستم شنوایی تجزیه و تحلیل می‌شود و مجدد به بخش مفهوم‌ساز بازمی‌گردد و فرایند تولید گفتار دوباره آغاز می‌شود.

بر اساس این مدل، صحبت کردن فرایندی چرخه‌ای دارد. در وهلهٔ نخست فرد به بررسی موقعیت و بافتی که در آن قرار گرفته است، می‌پردازد و مدلی ذهنی از آن موقعیت برای خود ترسیم می‌کند. سپس بر مبنای ارزیابی صورت‌گرفته اهدافی را برای خود مشخص می‌کند و برنامه‌ای برای اجرای آن می‌ریزد و درپایان نیز بر اجرای برنامه نظارت می‌شود تا امور آن‌گونه که برنامه‌ریزی شده‌اند، انجام گیرند. مدل دن داگلاس به ما نشان می‌دهد که مهارت صحبت کردن از سازه‌هایی مانند دانش جهانی، دانش زبانی و توانش راهبردی تشکیل شده است. علاوه بر این، سخن‌گو باید توانایی تولید گفتار سلیس، روان و قابل فهم را داشته باشد. آنچه در گفتار اهمیت بسیاری دارد، تشخیص درست بافت و موقعیت و تولید زبانیِ متناسب با آن است.



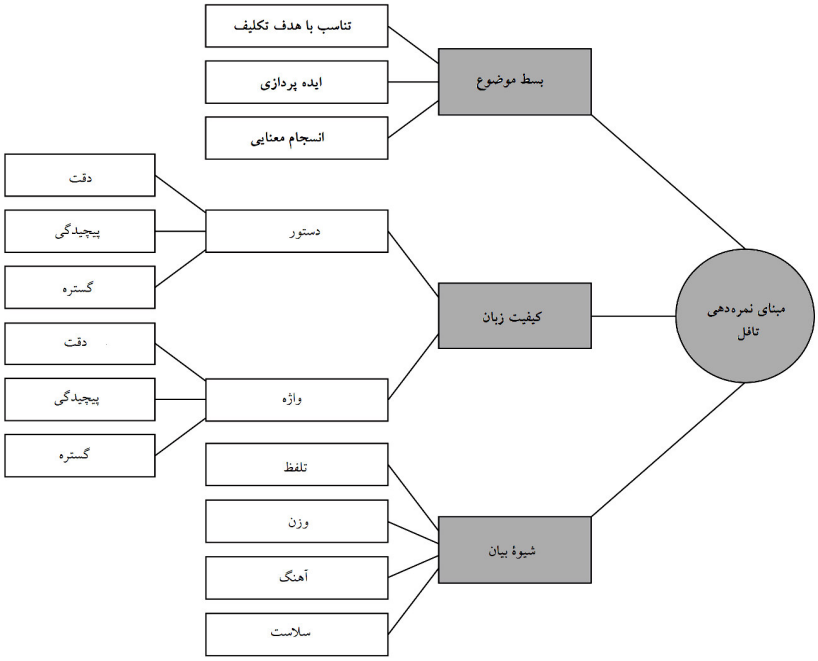
شکل ۲: فرایند تولید گفتار برگرفته از داگلاس (Douglas, 1997: 4)

Figure 2: The process of speech production

۳ـ۳. معیاری برای نمره‌دهی مهارت صحبت کردن

مدل دن داگلاس نشان داد که مهارت صحبت کردن از سازه‌هایی مانند توانش راهبردی، دانش زبانی و توانایی تولید آوایی تشکیل شده است. همان‌طور که در شکل سه دیده می‌شود، طراحان تافل تحت تأثیر این مدل، معیاری طراحی کرده‌اند که از سه مؤلفهٔ بسط موضوع، کیفیت زبان و شیوهٔ بیان تشکیل شده است. در بسط موضوع توانایی آزمون‌دهنده در تولید گفتار متناسب با بافت، ایده‌پردازی و انسجام معنایی بررسی می‌شود. برای مثال، اگر از فردی بخواهیم که دربارهٔ

معایب و مزایای پدیدهٔ مهاجرت برای ما صحبت کند، تا چه میزان او می‌تواند به‌شکلی منسجم و هدفمند به بسط این موضوع بپردازد. منظور از کیفیت زبان نیز صحت، گستره و پیچیدگی ساختارهای واژگانی و نحوی است. کلام تولیدشده دارای چه گسترهٔ واژگانی و نحوی است. آیا آزمون‌دهنده توانسته است از ساختارهای پیچیدهٔ زبانی نیز استفاده کند؟ درپایان نیز مؤلفهٔ شیوهٔ بیان قرار دارد که تمرکز آن بر سلاست گفتار، تلفظ، لحن و آهنگ کلام است. در این بخش تمرکز بر سرعت گفتار، تعداد مکث‌ها و تلفظ صحیح واژگان است. طراحان تافل مؤلفه‌های شناخته‌شده را در یک مقیاس پنج درجه‌ای تنظیم کرده‌اند که در جدول شمارهٔ یک آمده است. یادآوری می‌شود که طراحان آزمون بسندگی فارسی فردوسی نیز از این معیار برای نمره‌دهی مهارت گفتاری فارسی‌آموزان به‌صورت تحلیلی استفاده کرده‌اند.



شکل ۳: مبانی نمره‌دهی مهارت صحبت کردن در آزمون تافل

(Xi, Higgins, Zechner & Williamson, 2008: 29 cited in Hughes, 2011: 99)

Figure 3: The constructs of speaking scoring rubric in TOEF

جدول ۱: معیار نمره‌دهی صحبت کردن برای تکلیف مستقل برگرفته از تافل (ETS, 2012: 188 - 189)

Table 1: Toefl's speaking scoring rubric for independent task

نمره	شیوهٔ بیان	کیفیت زبان	بسط موضوع
۴	به‌طور کلی گفتار روان است و دارای سرعت خوبی است. سخن روشن است. ممکن است اشتباهات مختصری در تلفظ و یا آهنگ کلام وجود داشته باشد؛ اما این مسئله تأثیری در فهم کلام ندارد.	دستور و واژگان به‌شکل کارآمدی به‌کار گرفته شده‌اند. استفاده از نحو و واژگان به‌شکل خودکار صورت می‌گیرد و حکایت از تسلط خوبی بر ساختارهای ساده و پیچیده (مناسب) دارد. هر چند گاهی اوقات اشتباهاتی دیده می‌شود؛ اما معنا دچار اختلال نمی‌شود.	پاسخ داده‌شده کافی و پیوسته است. به‌طورکلی خوب بسط داده شده و منسجم است. ارتباط بین ایده‌ها روشن است.
۳	سخن به‌طور کلی روشن است و تقریباً روان؛ اما اشتباهات مختصری در تلفظ و آهنگ کلام وجود دارد که ممکن است شنونده را برای فهم به زحمت بیندازد (هر چند فهم کلام به‌طور جدی دچار مشکل نمی‌شود).	دستور و واژگان به‌شکل کارآمد و خودکار به‌کار گرفته می‌شود. ایده‌های مرتبط به‌شکل منسجمی بیان می‌شوند. ممکن است در پاسخ برخی ساختارهای نحوی و واژگانی به‌صورت اشتباه به‌کار گرفته شود و این امر بر سلاست کلی گفتار تأثیر بگذارد؛ اما انتقال پیام را با مشکل روبه‌رو نمی‌کند.	پاسخ داده‌شده پیوسته است و اطلاعات مرتبط را منتقل می‌کند. به‌طورکلی مطلب به‌شکل کامل بسط نیافته و به اندازهٔ کافی از مثال‌ها و جزئیات استفاده نشده است. گاهی ارتباط بین اطلاعات و ایده‌ها روشن نیست.
۲	زبان عمدتاً قابل فهم است، هر چند به‌دلیل بیان غیرشفاف، آهنگ کلام نامناسب و وزن گفتارِ مقطع مقطع، شنونده به زحمت می‌افتد. معنا نیز در برخی موارد مبهم است.	پاسخ بیانگر محدود بودن گسترهٔ واژگان و دستور است. این محدودیت معمولاً مانع بیان کامل ایده‌ها می‌شود. در بیشتر موارد تنها ساختارهای مقدماتی به‌طور کامل و سلیس بیان می‌شوند. تنها گزاره‌هایی با ساختار نحوی و واژگانی ساده استفاده می‌شود که به‌شکلی ابتدایی و یا مبهمی به یکدیگر پیوند خورده‌اند.	پاسخ مرتبط با تکلیف است. هر چند بسط اطلاعات و ایده‌ها و یا تعداد آن‌ها محدود است، بیشتر ایده‌های مقدماتی بیان می‌شوند با جزئیات محدود؛ گاهی محتوای مطرح‌شده مبهم است و یا تکراری. ارتباط بین ایده‌ها ممکن است روشن نباشد.
۱	اشتباهات در تلفظ، تکیه و آهنگ کلام، شنونده را برای فهم کلام بسیار به زحمت می‌اندازد. شیوهٔ بیان مقطع مقطع و تلگرافی است. مکث‌های بسیاری در کلام دیده می‌شود.	گسترهٔ محدود واژگان و دستور به‌طور جدی انتقال و پیوند ایده‌ها را با مشکل روبه‌رو می‌کند؛ برخی از پاسخ‌ها تنها شامل ساختارهایی از پیش حفظ‌شده هستند.	محتوای مرتبطِ کمی ارائه شده است. پاسخ به‌ندرت از حد محتوای بسیار مقدماتی فراتر می‌رود؛ سخن‌گو ممکن است توان تداوم سخن برای انجام تکلیف را نداشته باشد و بر تکرار صورت پرسش تکیه کند.
۰	داوطلب هیچ تلاشی برای پاسخ‌دهی نمی‌کند و یا پاسخ او به موضوع مرتبط نیست.		

۴. روش پژوهش

هدف از این پژوهش، بررسی اعتبار معیار نمره‌دهی مهارت صحبت کردن در آزمون زبان فارسی فردوسی برای غیرفارسی‌زبانان است. به همین منظور، نتایج به‌دست آمده از بخش گفتاری یکی از آزمون‌های برگزارشده در مرکز بین‌المللی زبان فارسی فردوسی از سوی مدل‌های آماری راش و تحلیل عاملی بررسی شد. در این آزمون تعداد ۱۰۶ آزمون‌دهنده از ملیت‌های مختلف شرکت داشتند.

۴ ـ ۱. ابزار پژوهش

ابزار اصلی این پژوهش برای گردآوری اطلاعات، آزمون بسندگی زبان فارسی فردوسی است که سالانه در دو نوبت، در اواسط زمستان و تابستان، در مرکز بین‌الملل دانشگاه فردوسی مشهد برگزار می‌شود. سنجش مهارت صحبت کردن در این آزمون به‌صورت مصاحبه انجام می‌شود و از دو تکلیف تشکیل شده است. در تکلیف نخست که خود چهار پرسش را دربر می‌گیرد، از داوطلب خواسته می‌شود تا یک فرد، مکان و یا شی را توصیف کند. همچنین، در این بخش پرسش‌هایی دربارهٔ نظر و عقیدهٔ شخصی داوطلب دربارهٔ یک موضوع عمومی اجتماعی پرسیده می‌شود. در تکلیف شمارهٔ دو، بخش صحبت کردن ـ که ترکیبی از مهارت گوش کردن، خواندن و صحبت کردن است ـ ابتدا داوطلب مطلبی را می‌خواند. سپس به گفت‌وگویی مرتبط با آن گوش می‌دهد و درنهایت، از او در ارتباط با آنچه خوانده و شنیده، پرسشی پرسیده می‌شود (نمونه‌سؤالات این آزمون در بخش پیوست آمده است).

۴ ـ ۲. شرکت‌کنندگان در پژوهش

شرکت‌کنندگان در این پژوهش ۱۰۶ تن متشکل از ۳۰ زن و ۷۶ مرد بوده‌اند. عراق با ۵۰ شرکت‌کننده و پاکستان با ۳۰ تن، به ترتیب اولین و دومین تعداد شرکت‌کننده را در این آزمون داشته‌اند. سایر شرکت‌کنندگان از کشورهای هندوستان، اندونزی، لبنان، سوریه و ایتالیا بوده‌اند که هر کدام به ترتیب ،۱۳، ۲، ۲، ٤ و ۵ شرکت‌کننده داشته‌اند. از نظر رشتهٔ تحصیلی، گروه علوم انسانی با ٦٨ نفر، پُرمخاطب‌ترین حوزهٔ دانشگاهی را تشکیل می‌دهد. علوم مهندسی با ۲۳ و پزشکی با ۱۱ نفر در رتبه‌های بعدی از نظر تعداد شرکت‌کننده بوده‌اند.

جدول ۲: شرکت‌کنندگان در آزمون بسندگی فارسی

Table 2: Participnts in Persian proficiency test

جنسیت		ملیت				رشتهٔ تحصیلی		
زن	مرد	عراقی	هندی	پاکستانی	غیره	علوم انسانی	علوم پزشکی	علوم مهندسی
۲۹	۷۶	۵۰	۱۳	۳۰	۱۳	۶۸	۱۱	۲۳

۴ـ۳. مراحل انجام پژوهش

آزمون بسندگی زبان فارسی در تاریخ ۸ تیر ۱۳۹۷ در مرکز بین‌المللی زبان فارسی دانشگاه فردوسی مشهد و در تاریخ ۲۵ آبان ۱۳۹۷ در دانشگاه استراسبورگ فرانسه برگزار شد. پاسخ‌گویی به پرسش‌ها در ساعت ۱۵: ۸ آغاز شد و داوطلبان در مدت‌زمان حدود سه ساعت به‌ترتیب به پرسش‌های بخش‌های گوش کردن، خواندن و نوشتنِ آزمون پاسخ دادند. بخش صحبت کردن نیز پس از استراحت یک ساعته برگزار شد. مراحل برگزاری این بخش ـ که برای هر فرد حدود ۱۰ تا ۱۲ دقیقه است ـ در جدول شمارهٔ ۳ آمده است.

جدول ۳: مراحل برگزاری بخش صحبت کردن در آزمون جامع زبان فارسی فردوسی

Table 3: The administration process in the speaking part of Persian proficiency test

زمان	مراحل
۱ ـ ۲ دقیقه	**آماده‌سازی** ابتدا داوطلب در صندلی خود قرار می‌گیرد و ممتحن با او احوال‌پرسی می‌کند. از زبان‌آموز خواسته می‌شود که خود را معرفی کند. هدف از این مرحله، آماده کردن داوطلب برای آزمون و کم کردن استرس او است.
۳ ـ ٤ دقیقه	**تکلیف شمارهٔ یک**: پرسش‌های شخصی و بیان و نظر و عقیده این بخش از آزمون خود دو نوع پرسش را دربر می‌گیرد. پرسش‌های نوع اول به پرسش‌های شخصی دربارهٔ توصیف یک فرد، مکان و یا شئ می‌پردازد که از آن میان ممتحن باید دو پرسش را انتخاب کند و از داوطلب بپرسد. پرسش‌های نوع دوم مربوط به بیان نظر و عقیدهٔ شخصی آزمون‌دهنده دربارهٔ یک موضوع اجتماعی می‌شود. ممتحن از میان پرسش‌های مربوط به این بخش نیز دو پرسش را انتخاب کرده و از آزمون‌دهنده می‌پرسد.

زمان	مراحل
۳ ـ ٤ دقیقه	**تکلیف شمارهٔ دو**: ترکیبی در این بخش از آزمون، داوطلب متنی را می‌خواند و سپس به گفت‌وگویی مرتبط با آن گوش می‌دهد و درنهایت، بر اساس آنچه خوانده و شنیده است، پرسشی از او می‌شود. داوطلب باید بتواند پاسخ آن را به‌صورت شفاهی بیان کند. **پایان مصاحبه** به داوطلب اعلام می‌شود که آزمون به پایان رسیده است. دستگاه ضبط صدا خاموش شده است و نحوهٔ اعلام نتیجهٔ آزمون و زمان آن به اطلاعش رسانده می‌شود.

۴ ـ ۴. مدل‌های آماری به‌کار رفته برای تحلیل داده‌ها

برای بررسی میزان کارآمدی معیار نمره‌دهی و نیز پایایی آن از مدل آماری راش استفاده شد. این مدل که ساختاری پیچیده‌تر از مدل‌های کلاسیک دارد، دارای یک مزیت اساسی است و آن وابسته نبودن به نمونهٔ آماری پژوهش است (Bachman, 2004: 139). مدل‌های آماری کلاسیک مانند کودر ریچاردسون[۲۹]، روش دونیمه[۳۰]، آلفا کرونباک[۳۱] و آزمون بازآزمون[۳۲] بر این فرض اساسی استوارند که نمرهٔ به‌دست‌آمده از یک آزمون حاصل جمع نمرهٔ حقیقی فرد و میزانی از خطا در محاسبه است. مبنی قرار دادن چنین فرضی سبب شده است که این مدل‌ها دارای دو ضعف اساسی باشند. نخست اینکه توانایی تمییز خطاهای نظام‌مند از خطاهای تصادفی را نداشته باشند (*ibid*, 1990: 186)، دوم اینکه با تغییر نمونهٔ پژوهش نتایج به‌دست آمده نیز دچار تغییر شود. درمقابل، مدل‌های راش این قابلیت را دارند که عملکرد آزمون‌دهندگان را در آزمون با توجه به توانمندی زبانی آن‌ها بررسی کنند. همین رویکرد سبب می‌شود که دیگر به نمونهٔ آماری پژوهش وابسته نباشند.

به منظور بررسی اعتبار سازه‌ای آزمون از مدل تحلیل عاملی تأییدی استفاده شده است. این مدل ابزاری است برای بررسی رابطهٔ مفروض بین عامل‌های مکنون و متغیرهای قابل مشاهده (*ibid*, 2004: 113). روش تحلیل عاملی تأییدی تعیین می‌کند که آیا داده‌ها با ساختار عاملی معین هماهنگ هستند یا نه. می‌توان این مدل آماری را شیوه‌ای دقیق و علمی برای بررسی مبانی

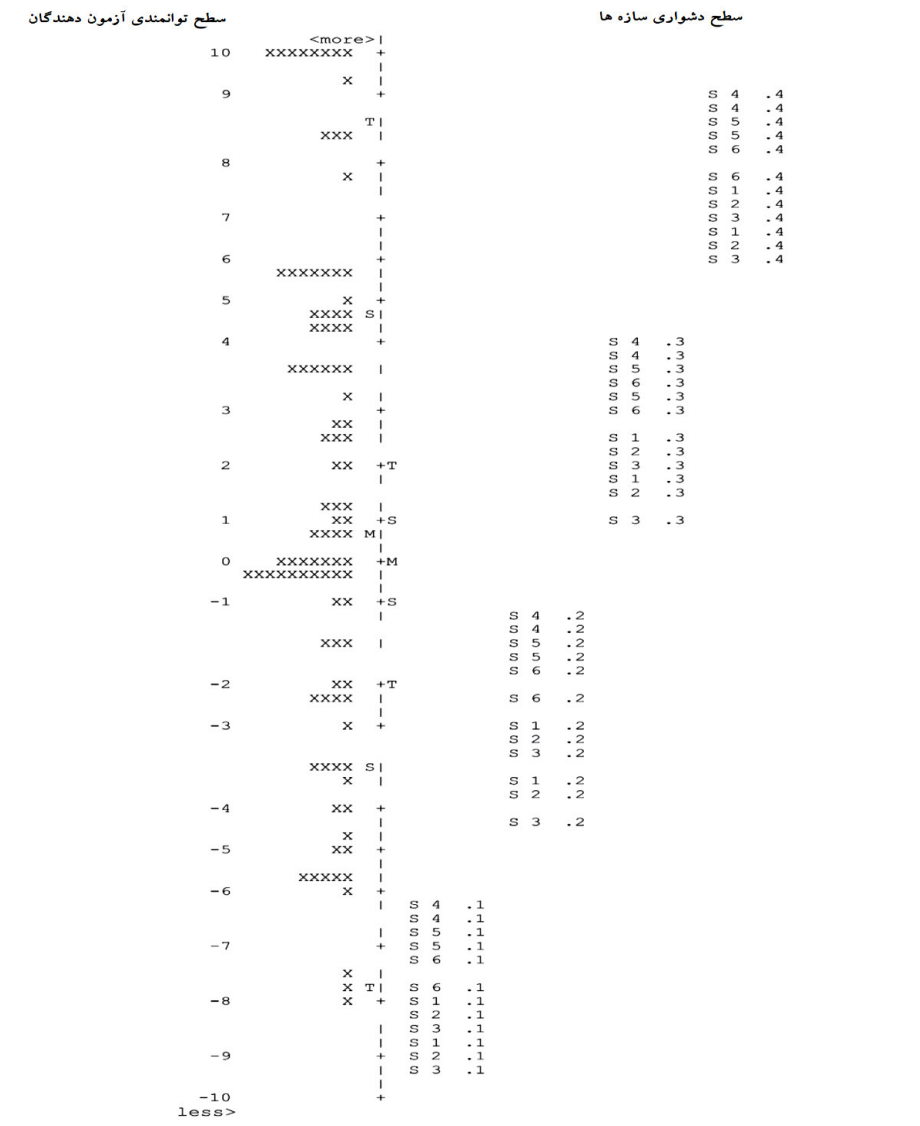
نظری طراحی آزمون دانست. در این پژوهش از نرم‌افزار آموس[۳۳] ـ که یکی از مشهورترین نرم‌افزارها برای اجرای این گونه مدل‌هاست ـ به منظور آزمون فرضیه‌ها استفاده شد.

۵. یافته‌های پژوهش برای مهارت صحبت کردن

۵ ـ ۱. نقشهٔ آزمون‌دهنده ـ پرسش

شکل ٤ نقشهٔ رایت[۳٤] یا آزمون‌دهنده ـ پرسش است. در این نقشه، سطح توانایی زبانی آزمون‌دهندگان و سطح دشواری سازه‌های موجود در معیار نمره‌دهی ـ که از سوی هر یک از نمره‌دهندگان به آزمون‌دهندگان اختصاص داده شده است ـ به‌طور هم‌زمان نمایش داده می‌شود و درنتیجه، مستقیماً قابل‌مقایسه هستند. در سمت راست نقشه، سطح سختی سازه‌ها و در سمت چپ، سطح توانایی آزمون‌دهندگان قرار دارند. مؤلفه‌های پایین نقشه نشان‌دهندهٔ سازه‌های

ساده‌تر و آزمون‌دهندگان ضعیف‌تر و مؤلفه‌های بالای نقشه، نشان‌دهندهٔ سازه‌های دشوارتر و آزمون‌دهندگان قوی‌تر است. S1 سازهٔ بسط موضوع برای نمره‌دهندهٔ اول است. همچنین، S2 سازهٔ کیفیت زبان و S3 شیوهٔ بیان برای این ارزیاب است. S1.1 نشان‌دهندهٔ نمرهٔ یک برای سازهٔ بسط موضوع است که از سوی نمره‌دهندهٔ اول اختصاص یافته است. به همین ترتیب، S1.2 نشان‌دهندهٔ نمرهٔ دو برای این سازه است. از سوی دیگر S4، S5، S6 به‌ترتیب نمره‌هایی هستند که نمره‌دهندهٔ شمارهٔ دو به هر یک از سازه‌های بسط موضوع، کیفیت زبان و شیوهٔ بیان داده است. بنابراین، S4.3، یعنی نمرهٔ ۳ برای سازهٔ بسط موضوع که از سوی نمره‌دهندهٔ شمارهٔ دوم اختصاص داده شده است. انتظار می‌رود که دشواری سازه‌ها متناسب با سطوحِ زبانیِ مختلفِ آزمون‌دهندگان بالا روند تا توانایی تمییز داوطلبان ضعیف، متوسط و قوی را داشته باشند. اگر در نقشه‌ای در مقابل پرسش و مجموعه‌ای از پرسش‌ها، هیچ داوطلبی قرار نگیرد و یا تعداد کمی داوطلب دیده شود. این نکته نشان می‌دهد که آن معیار کارکرد چندانی در تمییز توانایی داوطلبان از یکدیگر ندارد و باید بازبینی شود.



شکل ۴: نقشهٔ آزمون‌دهنده ـ پرسش برای مهارت صحبت کردن

Figure 4: The Wright map for speaking test

نقشه نشان می‌دهد که مؤلفه‌ها و درجه‌های نمره‌گذاری (۰ ـ ٤) همهٔ گسترهٔ توانایی آزمون‌دهندگان را در برمی‌گیرد و با این مقیاس می‌توان همهٔ آزمون‌دهندگان با هر میزان توانایی صحبت کردن را اندازه گرفت.

۵ ـ ۲. مقیاس نمره‌گذاری[۳۵]

در آزمون صحبت کردن از یک مقیاس پنج‌درجه‌ای برای نمره‌دهی استفاده شد. بررسی عملکرد مقیاس و یا استفادهٔ درست آزمون‌گیرندگان از مقیاس، بخش مهمی از تحلیل آزمون‌های عملکرد-محور[۳۶] است. جدول شمارهٔ ٤ ویژگی‌های مقیاس استفاده‌شده در این آزمون برای نمره‌دهی را نشان می‌دهد. اولین ستون از راست، بیانگر نمرهٔ هر درجه از مقیاس است که باید بین ۰ ـ ٤ باشد. ستون دوم تعداد دفعاتی است که هر نمره داده شده است. برای مثال، نمرهٔ صفر ۳٤ بار و نمرهٔ چهار ۱۵۸ بار داده شده است. ستون سوم میانگین توانایی (در مقیاس راش) افرادی است که آن نمره به آن‌ها داده شده است. انتظار می‌رود با بالا رفتن نمره، میانگین توانایی نیز بالاتر رود؛ چون افراد تواناتر قاعدتاً نمرهٔ بیشتری گرفته‌اند. ستون چهارم نیز دشواری آستانه‌ها را نشان می‌دهد که باید به‌ترتیب نمره اضافه شود. به‌هم‌ریختگی ترتیب بزرگی آستانه‌ها نشانهٔ استفادهٔ نادرست از مقیاس خواهد بود (Linacre, 2009).

آستانه‌ها نقاطی بر روی طیف توانایی‌اند که احتمال دو نمره کنار هم در آن نقاط پنجاه درصد است. نمرهٔ اول، یعنی ۰ آستانه‌ای ندارد؛ چون قبل از آن نمرهٔ دیگری نیست. آستانهٔ بین ۰ و ۱ دشواری آن ۷/۷۳- است؛ یعنی اگر فردی توانایی او در صحبت کردن برابر این مقدار باشد، به احتمال پنجاه درصد، آزمون‌گیرنده به او نمره‌ای بین ۰ و یا ۱ می‌دهد. به عبارت دیگر، برای اینکه یک آزمون‌دهنده نمرهٔ ۰ یا ۱ بگیرد (در هر تکلیف و در هر مؤلفه) باید توانایی‌اش ۷/۷۳- باشد؛ اما اگر فردی توانایی او ۲/۵۱- باشد، ۵۰ درصد احتمال دارد ۱ و یا ۲ بگیرد.

اگر از مقیاس نمره‌گذاری درست استفاده شده باشد، انتظار می‌رود که دشواری آستانه‌ها به‌ترتیب با اندازهٔ نمره زیاد شود که در اینجا همین‌گونه است. نمره‌های بین ۰ تا ٤ هر یک معرف میزان متفاوتی از توانایی صحبت کردن هستند و جای متفاوتی بر روی مقیاس دارند. آزمون‌گیرندگان باید بتوانند تفاوت‌های بین این مقادیر را تشخیص دهند و از مقیاس برای بیان تفاوت‌های بین آزمون‌دهندگان در میزان سازهٔ مورد اندازه‌گیری، درست استفاده کنند. همان‌طور

که جدول شمارهٔ ٤ نشان می‌دهد، ترتیب آستانه‌ها مطابق ترتیب نمره‌هاست و به‌هم‌ریختگی ندارد که این خود گواهی بر استفادهٔ درست از مقیاس و نمره‌گذاری صحیح آزمون است.

جدول ۴: آماره‌های مقیاس نمره‌گذاری مهارت صحبت کردن

Table 4: Speaking skills scale statistics

مقیاس نمره‌گذاری	تعداد دفعات	میانگین توانایی برای کسب نمره	آستانه
۰	۳٤	-۷/۲۵	بدون آستانه
۱	۲۵۰	-٤/۰۰	-۷/۷۳
۲	٤۵۹	۰/۱٦	-۲/۵۱
۳	۲۹۹	٤/۱۹	۲/٦٦
٤	۱۵۸	۷/۸۹	۷/۵۸

۵ـ۳. تحلیل عاملی تأییدی[۳۷] بخش صحبت کردن

به‌منظور بررسی اعتبار سازه‌ای آزمون از مدل تحلیل عاملی تأییدی استفاده می‌شود. این مدل ابزاری است برای بررسی رابطهٔ مفروض بین عامل‌های مکنون و متغیرهای قابل مشاهده (Bachman, 2004: 113). روش تحلیل عاملی تأییدی تعیین می‌کند که آیا داده‌ها با یک ساختار عاملی معین هماهنگ هستند یا نه. می‌توان این مدل آماری را شیوه‌ای دقیق و علمی برای بررسی مبانی نظری طراحی آزمون دانست. در این پژوهش از نرم‌افزار ایموس[۳۸] ۲٤ ـ که یکی از مشهورترین نرم‌افزارها برای اجرای این‌گونه مدل‌هاست ـ به‌منظور آزمون فرضیه‌ها استفاده خواهد شد.

برای بررسی نرمال بودن از ضریب چولگی[۳۹] و ضریب کشیدگی[٤۰] استفاده می‌شود. با کمک ضرایب چولگی و کشیدگی می‌توان نرمال بودن توزیع متغیر را بررسی کرد. زمانی که توزیع متغیرها نرمال است، برای آزمون فرضیه‌ها از آزمون‌های پارامتریک استفاده می‌شود و در غیر این صورت آزمون‌های ناپارامتریک استفاده می‌شود. اگر قدر مطلق ضریب چولگی و کشیدگی بزرگ‌تر از ۲ باشد، تخطی از نرمال بودن داده‌ها را نشان می‌دهد و در این صورت فرض نرمال بودن توزیع متغیر رد می‌شود. در کل تحقیق سطح اطمینان ۹۵ درصد (سطح خطای ۵ درصد) در نظر گرفته شده است.

می‌دانیم که به‌منظور کاهش متغیرها و درنظر گرفتن آن‌ها به‌منزلهٔ یک متغیر مکنون، بار عاملی[41] به‌دست‌آمده باید بیشتر از ۳ درصد باشد. در تحلیل عاملی تأییدی محقق می‌داند چه پرسشی مربوط به چه بعدی است. به‌عبارتی، مدل مفهومی برای هر یک از مفاهیم یا متغیرهای تحقیق وجود دارد. در بررسی هر کدام از مدل‌ها پرسش اساسی این است که آیا این مدل‌های اندازه‌گیری مناسب هستند؟ به عبارت دیگر، آیا داده‌های تحقیق با مدل مفهومی هم‌خوانی دارد یا نه؟ به‌طور کلی دو نوع شاخص برای آزمودن برازش مدل وجود دارد: ۱. شاخص‌های خوب بودن، ۲. شاخص‌های بد بودن.

شاخص‌های خوب بودن شاخص نیکویی برازش[42] و نیکویی برازش تعدیل‌شده[43] هستند که هر چه مقدار آن‌ها بیشتر باشد، بهتر است. مقدار پیشنهادی برای چنین شاخص‌هایی ۸ درصد است. همچنین، شاخص‌های بد بودن نیز شامل کای‌دو بر درجهٔ آزادی[44] و میانگین مجذور خطاهای مدل[45] است که هر چه مقدار آن‌ها کم‌تر باشد، مدل دارای برازش بهتری است. حد مجاز کای‌دو بر درجهٔ آزادی عدد ۳ و حد مجاز میانگین مجذور خطاهای مدل عدد ۸ درصد است. برای پاسخ به پرسش برازش مدل باید شاخص‌های خوب بودن و بد بودن بررسی شوند.

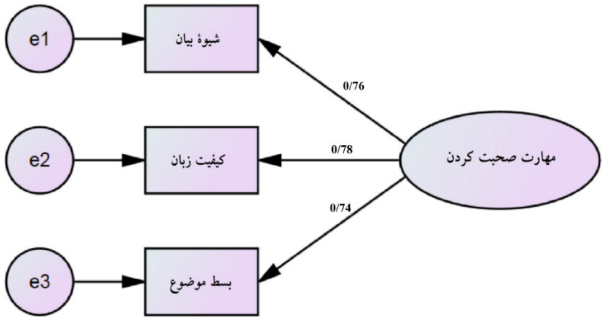
شاخص کای‌دو نشان‌دهندهٔ میزان آمارهٔ کای‌دو برای مدل است. درواقع، این شاخص اختلاف بین مدل و داده‌ها را نشان می‌دهد و معیاری برای بد بودن مدل است. هرقدر که میزان آن کم‌تر باشد، نشان‌دهندهٔ اختلاف کم‌تر بین ماتریس واریانس ـ کوواریانس نمونهٔ اتخاذشده و ماتریس واریانس ـ کوواریانس حاصل از مدل اتخاذشده بوده و بد بودن مدل را نشان می‌دهد. البته، میزان این شاخص تحت تأثیر تعداد نمونهٔ اتخاذشده قرار می‌گیرد. درواقع، چنانچه حجم نمونه بیشتر از ۲۰۰ بشود، این شاخص تمایل زیادی به افزایش دارد. تحلیل برازندگی مدل با این شاخص، معمولاً در نمونه‌های بین ۱۰۰ تا ۲۰۰ قابل اتکا است و بهتر است که این شاخص، با درنظر گرفتن درجهٔ آزادی تفسیر شود.

درجهٔ آزادی نیز نباید کوچک‌تر از صفر باشد. یکی از بهترین شاخص‌های بررسی نیکویی برازش مدل، بررسی نسبت آمارهٔ کای‌دو بر درجهٔ آزادی است. البته، حد استانداردی برای مناسب بودن میزان این شاخص وجود ندارد؛ اما بسیاری از صاحب‌نظران در این زمینه بر این عقیده‌اند که این شاخص باید کم‌تر از ۳ باشد.

شاخص میانگین مجذور خطاهای مدل ازجمله معیارها برای بد بودن مدل است. برخی بر این

عقیده‌اند که این شاخص باید کم‌تر از ۵ درصد باشد، همچنین، برخی دیگر میزان کم‌تر از ۵ درصد را مناسب می‌دانند. از سوی دیگر، شاخص نیکویی برازش معیاری برای سنجش میزان خوب بودن مدل است و میزانی بالاتر از ۸ درصد، نشان‌دهندهٔ مناسب بودن مدل استخراج‌شده با توجه به داده‌هاست. همچنین، نیکویی برازش تطبیق داده‌شده درواقع، حالت تطبیق داده‌شدهٔ شاخص نیکویی برازش با درنظر گرفتن میزان درجهٔ آزادی است و معیار دیگری برای خوب بودن مدل است. چنانچه میزان این شاخص نیز بالاتر از ۸ درصد باشد، نشان‌دهندهٔ مناسب بودن مدل استخراجی با توجه به داده‌هاست.

قدرت رابطهٔ بین عامل (متغیر پنهان) و متغیر قابل مشاهده به‌وسیلهٔ بار عاملی نشان داده می‌شود. بار عاملی مقداری بزرگ‌تر از صفر است. اگر بار عاملی کم‌تر از ۳ درصد باشد، رابطهٔ ضعیف در نظر گرفته شده و از آن صرف‌نظر می‌شود. بار عاملی بین ۳ درصد تا ٦ درصد قابل قبول است و اگر بزرگ‌تر از ٦ درصد باشد، خیلی مطلوب است. بارهای عاملی میزان تأثیر هر کدام از مؤلفه‌ها را در توضیح و تبیین واریانس نمرات متغیر نشان می‌دهد. به عبارت دیگر، بار عاملی نشان‌دهندهٔ میزان هم‌بستگی هر متغیر مشاهده‌شده (ابعاد مربوط به هر یک از بخش‌های آزمون) با متغیرهای پنهان (چهار بخش آزمون) است. برای بررسی معنادار بودن رابطهٔ بین متغیرها از آمارهٔ تی[٤٦] استفاده می‌شود؛ چون معناداری در سطح خطای ۵ درصد بررسی می‌شود، بنابراین، اگر مقدار آمارهٔ تی از ۱/۹٦ بزرگ‌تر شود، رابطه معنادار است. مدل اندازه‌گیری بخش صحبت کردن آزمون در حالت تخمین استاندارد در شکل شمارهٔ ۵ نشان داده شده است.



شکل ۵: تحلیل عاملی تأییدی بخش صحبت کردن

Figure 5: Confirmatory factor analysis for speaking test

در این مدل، مسئلهٔ مهم در درجهٔ اول، مقادیر بار عاملی بین عوامل و پرسش‌های مربوط است که تمامی مقادیر بالاتر از ۳۰ درصد است. با توجه به بارهای عاملی می‌توان میزان اهمیت و تأثیر هر یک از ابعاد برای متغیر مورد نظر را مشخص کرد. بعد بسط موضوع مهم‌ترین عامل در بخش صحبت کردن (۰.۹۹۵) است. در جدول شمارهٔ ۵ مقادیر بارهای عاملی، برآورد پارامترها و معنی‌داری آن‌ها از سوی آزمون تی (T) گزارش می‌شود.

جدول ۵: معنی‌داری برآورد پارامترهای بخش صحبت کردن

Table 5: The estimation of the meaningfulness of the speaking test's parameters

متغیر اول	بار عاملی	ضریب رگرسیون	خطای برآورد	مقدار آمارهٔ T
بسط موضوع	۰/۷٦	۱/۰۵۱	۰/۰۲۸	۳۷/۵۷۷
کیفیت زبان	۰/۷۸	۱/۰٤٦	۰/۰۲۹	۳۵/۸۲۳
شیوهٔ بیان	۰/۷٤	۱		

نتایج جدول شمارهٔ ۵ نشان می‌دهد که هر سه بعد بخش صحبت کردن تأثیر معناداری در سطح خطای ۵ درصد دارند. اعتبار این مدل از سوی شاخص‌های نیکویی برازش در جدول شمارهٔ ٦ گزارش شده است.

جدول ۶: شاخص‌های نیکویی برازش تحلیل عاملی تأییدی بخش صحبت کردن

Table 6: The goodness of fit in the speaking test's confirmatory factor analysis

کای دو بر درجهٔ آزادی	میانگین مجذور خطاهای مدل	شاخص نیکویی برازش	کای دو بر درجهٔ آزادی
مقدار	۱/۸۸٤	۰/۰٦۷	۰/۹۹٤
حالت مطلوب	$1 \leq \& \leq 3$	$0 \leq \& \leq 0.08$	$0.9 \leq \& \leq 1$

با توجه به خروجی نرم‌افزار که در جدول شمارهٔ ٦ ارائه شده است، شاخص‌های نیکویی برازش مدل در حد مطلوب قرار دارند. با توجه به شاخص‌ها و خروجی‌های نرم‌افزار ایموس می‌توان گفت که داده‌ها با مدل منطبق هستند و شاخص‌های ارائه‌شده نشان‌دهندهٔ این موضوع هستند که درمجموع، مدل ارائه‌شده مدل مناسبی است و داده‌های تجربی دراصطلاح به‌خوبی با

آن منطبق هستند.

۶. نتیجه

طراحان تافل بر مبنای تصویری که لولت (1989)، دبات (1992) و دن داگلاس (1997) از تولید گفتار ارائه داده‌اند، معیاری برای نمره‌دهی مهارت گفتاری طراحی کرده‌اند که از سه سازهٔ بسط موضوع، شیوهٔ بیان و کیفیت زبان تشکیل شده است. طراحان آزمون بسندگی زبان فارسی فردوسی نیز این معیار را برای سنجش داوطلبان غیرفارسی‌زبان ورود به دانشگاه‌های ایران استفاده کرده‌اند. در این پژوهش اعتبار این معیار بررسی شد. نتایج نشان داد که پایایی آزمون‌گیرنده برای آزمون بسندگی زبان فارسی ۹۷ درصد است. این عدد بیانگر درک نسبتاً یکسان هر یک از آزمون‌گیرندگان از معیار نمره‌دهی است؛ یعنی آزمون‌گیرندگان به آزمودنی‌ها نمرهٔ نسبتاً ثابتی داده‌اند که این خود یک نقطهٔ قوت برای آزمون محسوب می‌شود. باید این نکته را در نظر داشت که در آزمون برگزارشده تنها دو ارزیاب تمامی آزمون‌دهندگان را نمره‌دهی کرده‌اند. اگر قرار باشد افراد بیشتری به نمره‌دهی بپردازند، به‌طور قطع پایایی آزمون کاهش خواهد یافت.

همچنین، در این آزمون آزمون‌گیرندگان توانسته‌اند به‌شکل مناسبی مقیاس نمره‌دهی را برای آزمون‌دهندگان با سطوح توانمندی متفاوت به‌کار گیرند. هر کدام از چهار آستانهٔ به‌دست‌آمده از سوی مدل آماری راش به‌ترتیب حدود ۵ درجه با یکدیگر تفاوت دارند. افزایش منظم درجهٔ آستانه‌ها متناسب با توانمندی آزمون‌دهندگان است. این امر حکایت از درک صحیح آزمون‌گیرندگان از ۵ درجهٔ مشخص‌شده در معیار نمره‌دهی دارد؛ یعنی آزمون‌گیرندگان درک مناسبی از سطح توانمندی آزمون‌دهندگان و ارتباط آن با درجات معیار نمره‌دهی داشته‌اند.

نقشهٔ آزمون‌دهنده ـ پرسش نیز نشان داد که معیار نمره‌دهی توانایی تمییز زبان‌آموزان ضعیف، متوسط و قوی از یکدیگر را دارد؛ زیرا نمره‌های آزمون تمامی گسترهٔ توانمندی آزمون‌دهندگان از ضعیف تا قوی را دربر می‌گیرند. با این حال، در بالای طیف توانمندی آزمون‌دهنده هشت نفر قرار گرفته که هیچ نمره‌ای متناسب با سطح توانمندی آن‌ها دیده نمی‌شود؛ یعنی معیار نمره‌دهی توانایی تمییز این گروه را نداشته است. شاید بهتر باشد درجه‌ای متناسب با آن‌ها نیز تعریف شود.

ازجمله نقاط قوت آزمون بسندگی فارسی استفاده از دو تکلیف مستقل و ترکیبی است که هم‌راستا با توصیهٔ لارسن فریمن و لانگ (1991) در ارتباط با استفاده از تکلیف‌های متنوع در سنجش مهارت گفتاری است. در این آزمون به منظور نمره‌دهی به دو تکلیف مشخص‌شده سه سازهٔ بسط موضوع، کیفیت زبان و شیوهٔ بیان در نظر گرفته شده است. قرار گرفتن سازه‌ای با عنوان بسط موضوع حکایت از هم‌راستا بودن معیار نمره‌دهی با تصویری است که لولت (1989)، دبات (1992)، دن داگلاس (1997) و پیشقدم (2013) از مهارت گفتاری ارائه می‌دهند. توانایی صحبت کردن نه تنها دانش واژگانی، نحوی و آوایی را دربر می‌گیرد؛ بلکه فرد باید توانایی برنامه‌ریزی کردن برای تولید گفتار مرتبط با موضوع و بسط آن را نیز داشته باشد. به عبارتی، همان چیزی که پیشقدم (*ibid*) با نام هوش روایی معرفی می‌کند. بار عاملی به‌دست‌آمده برای سه سازهٔ بسط موضوع، کیفیت زبان و شیوهٔ بیان به‌ترتیب ۷٦ درصد، ۷۸ درصد و ۷٤ درصد بوده است. این امر نشان می‌دهد تقسیم توانایی صحبت کردن به سه عامل یادشده متناسب و دقیق است و هر کدام از این سازه‌ها توانمندی متفاوتی را سنجش می‌کنند که از این میان سازهٔ کیفیت زبان بیشترین و شیوهٔ بیان کم‌ترین میزان بار عاملی را داشته‌اند.

۷. پی‌نوشت‌ها

1. Norm reference
2. Scoring Scale
3. Rasch
4. Levelt
5. DeBot
6. Douglas
7. Fulcher Glenn
8. Testing second language speaking
9. Pishghadam
10. Tarone
11. Ellis
12. Larsen-Freeman
13. Long
14. Hadden
15. Ludwig
16. Chapelle
17. Inferences
18. Evaluation

19. Generalization
20. Extrapolation
21. Tasks
22. Explanation
23. Douglas
24. Conceptualizer
25. World Knowledge
26. Preverbal Message
27. Formulation
28. Phonetic plan
29. Kuder-Richardson
30. Split half Experiment
31. Cronbach's Alpha
32. Test-Retest experiment
33. Amos
34. wright map
35. Rating scale analysis
36. Performance-based exams
37. Confirmatory Factor Analysis
38. Amos
39. Skewness
40. Kurtosis
41. Factor load
42. GFI: Goodness of fit
43. AGFI: Ajusted Goodness of fit
44. χ^2/df
45. RMSEA
46. T-value

۸. منابع

- جلیلی، سید اکبر (۱۳۹۰). ***آزمون مهارتی فارسی (آمفا) بر پایهٔ چهار مهارت اصلی زبانی***. پایان‌نامهٔ کارشناسی ارشد. دانشگاه علامه طباطبائی تهران.
- گلپور، لیلا (۱۳۹٤). ***طراحی و اعتباربخشی آزمون بسندگی زبان فارسی بر پایهٔ چهار مهارت زبانی***. پایان‌نامهٔ دکتری. دانشگاه پیام‌نور مرکز، تهران.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. (2004). *Statistical analysis for language assessment*. New York: Cambridge University Press.

- Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy*. Englewood Cliff, NJ: Prentice Hall Regents.
- Chapelle, C. A.; Enright M. K. & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.
- Debot, K. (1992). "A bilingual production model: Levelt's speaking model adapted". *Applied Linguistics*.Pp: 13: 1-24.
- Douglas, D. (1997). *Testing speaking ability in academic context: theoretical considerations* (TOEFL No. 8). Princeton, NJ: ETS.
- Ellis, R. (1985). *Understanding second language acquisition.* Oxford: Oxford University Press.
- Ellis, R. (Ed.). (1987). *Second language acquisition in context.* Englewood Cliffs, NJ: Prentice-Hall International.
- Ellis, R. (1990). *Instructed second language acquisition.* Cambridge, MA: Basil Blackwell.
- ETS: Educational Testing System, (2012), *The official Guide to the TOEFL test*. (4th ed). New York: Mc Graw Hill.
- Fulcher, G. (2014). *Testing second language speaking*. NY: Routledge.
- Fulcher, G. & Davidson, F. (2007). *Language testing and assessment and advanced resource Book.* New York: Routledge.
- Golpour, L. (2014). *Designing and Validating Persian Proficiency Test based on Four Language Skills.* (PhD. Dissertation). Pyame Nour University-Central Branch Iran. [In Persian].
- Hadden, B. L. (1991). "Teacher and non-teacher perceptions of second-language communication". *Language Learning*. 41. Pp: 1-24.
- Hughes, Arthur (2003). *Testing for language teachers*. Cambridge: Cambridge University press.
- Jalili, A**.** (2011). *Persian Language Proficiency Test based on Four Main Langage Skills.* (MA. Thesis). Allameh Tabatabaei University, Iran. [In Persian].

- Kane M. T. (1999). "Validating measures of performance". *Educational Measurement. 18* (2).Pp: 5-17
- Larsen-Freeman, D. & Long, M. (1991). *An introduction to second language acquisition research.* New York: Longman.
- Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge: MA: The MIT Press.
- Linacre, J. M. (2009). *WINSTEPS Rasch Measurement* [Computer program]. Chicago, IL: Winsteps.
- Lissitz, R. W. (ed.), (2009). *The concept of validity: Revisions new directions and applications*. Charlotte, NC: Information Age Publishing, INC.
- Ludwig, J. (1982). "Native-speaker judgments of second language learners' efforts at communication: A review". *Modern Language Journal.* 66. Pp: 274-283.
- Messick, S. (1987). *Validity* (Report no. RR-87-40). Princeton: ETS.
- Pishghadam, R., Shams, M.A. (2013). "A new look into construct validity of the IELTS Speaking module". *The journal of teaching language skills*. 5(1). Pp: 71-99.
- Tarone, E. (1983). "On the variability of interlanguage systems". *Applied Linguistics. 4.*Pp: 143-63.