Journal of Teaching Persian to Speakers of Other Languages
Vol. 9, No. 1 (Tome 19), Spring & Summer2020
Print Issn:5394-2322
Electronic Issn:2676-3354

The Construct Validity of Writing Skill's Scoring Rubric in the Persian Proficiency Test of Ferdowsi University of Mashhad

Mohsen Roudmajani[1]

Corresponding author, Phd graduate in Persian Language and Literature, Ferdowsi University, Mashhad, Iran

Ehsan Ghabol[2]

Assisstent professor, Departement of Persian language and literature, Ferdowsi University of Mashhad, Iran

Behzad Ghonsoli[3]

Full Professor, Derpartement of Teaching English language, Ferdowsi University of Mashhad, Iran

Abstract:

Language assessment is one of the most important part of any language educational system.Much of the effectiveness of Persian language centers depends on the use of precise assessment and evaluation techniques.In fact, Persian language institutions face the fundamental question of how to convert abstract concepts of linguistic knowledge and communication ability into numbers.Understanding learners' progress, identifying their weaknesses, and making accurate decisions about them requires accurate and scientific methods of assessment and evaluation. The present study attempts to evaluate the construct validity of the Writing Skill's Assessment rubric in the Ferdowsi Persian Language Examination.The test, which is approved by the Ministry of Science, is held twice a year at Ferdowsi University of Mashhad, Iraq, and in some other countries.This research attempts to answer the following three questions:

1. To what extent do the constructs defined in the scoring rubric measure distinct components of the writing skill?
2. To what extent can the six-point scale distinguish intermediate, weak, and strong test takers?
3. To what extent do the scorers agree on the use of the scoring criteria?

So far, several scoring rubrics have been presented in Persian language in order to assess non-native Persian language speaker's writing ability but none of them have been validated by quantitative research methods. The aim of current study was to investigate

Received on: 29/07/2019 Accepted on: 02/02/2020

[1]. Email: Mroodmajani@gmail.com
[2]. Email: Ghabool@Ferdowsi.um.ac.ir
[3]. Email: Ghonsooly@yahoo.com

DOI: 10.30479/jtpsol.2020.11294.1451
pp.25-46

the construct validity of writing section of Ferdowsi University's Persian language proficiency test. This test is designed based on the TOEFL theoretical underpinnings. The writing scoring rubric in Ferdowsi University's Persian language proficiency test consists of three components namely language quality, cohesion, and topic development. Obviously, this scoring rubric is derived from a communicative view of the concept of language, which not only measures test taker's linguistic knowledge at the level of words, sentence and discourse, but also assesses the ability to perform language tasks.Based on this rubric, in addition to the quality of the language and the cohesion of the text, the evaluators examine to what extent the written text is consistent with the purpose of the language task.

In order to evaluate the construct validity of the scoring rubric of the test, the results of one of the tests held in Ferdowsi International Persian Language Center were analyzed by Rash statistical model and factor analysis.The test was held on July 8, 1397 at the International Center for Persian Language at Ferdowsi University of Mashhad and Strasbourg University in France.The writing section of the test includes two tasks.In the first task, an audio file was first played for the test takers, and then they were asked to write a summary of it.In the second task, the test takers were given a topic to write about it in 200 words.

The participants in this study were 106 students consisting of 30 women and 76 men.Iraq, with 50 participants, and Pakistan with 30, respectively, had the first and second highest participants.The other participants were from India, Indonesia, Lebanon, Syria, and Italy, each with 13, 2, 2, 4 and 5 participants, respectively.In terms of educational background, the Humanities, with 68 members, constitutes the most participants.Engineering with 23 and medicine with 11 were in the next ranks in terms of number of participants.

Factor analysis results showed that the three identified constructs all have high validity.That is, writing skill can be divided into language quality, cohesion, and topic development and can be scored separately.Cohesion with 0.98 had the highest factor load and the other two constructs each with 0.97 had the second highest factor loadings.Based on these statistics, it can be said that writing proficiency can be divided into language quality, cohesion and topic development constructs and each of these components measures a separate construct.

The Ferdowsi Persian Language Proficiency test use a six-point scoring scale. Rash's statistical model showed that each of the scorers was able to use this criterion fairly correctly because the order of thresholds were in accordance with the order of the scores and did not change. On the other hand, the write map indicated that this scoring scale has the ability to distinguish between weak, intermediate, and strong test takers, and the 0 to 5 criteria cover all range of test takers so this scale can measure all candidates with any level of writing ability.On the other hand, the reliability of the scoring was 0.96, which is very well.This result indicates that test evaluators have used scoring criteria in the same way so the test has reasonable scorer reliability.

Keywords: Language assessment, Constrcut validity, Wrtitng skill, AZFA

پژوهش‌نامه‌ی آموزش زبان فارسی به غیر فارسی‌زبانان(علمی-پژوهشی)
سال نهم، شماره‌ی اول (پیاپی ۱۹)، بهار و تابستان ۱۳۹۹
شاپا چاپی: ۵۳۹۴-۲۳۲۲
شاپا الکترونیکی: ۲۳۵۴-۲۶۷۶

اعتبارِ سازه‌ی معیارنمره‌دهیِ مهارت نوشتن در آزمون جامع فارسی دانشگاه فردوسی برای غیرفارسی‌زبانان (پژوهشی)

محسن رودمعجنی[۱]
نویسنده‌ی مسئول، دکتری زبان و ادبیات فارسی، دانشگاه فردوسی مشهد
احسان قبول[۲]
استادیار گروه زبان وادبیات فارسی، دانشگاه فردوسی مشهد
بهزاد قنسولی[۳]
استادیار گروه آموزش زبان انگلیسی، دانشگاه فردوسی مشهد

چکیده

سنجش زبان یکی از ارکان اساسی هر نظام آموزش زبان به شمار می‌آید. بخش عمده‌ای از کارآمدی مراکز آموزشی در گرو بهره‌گیری از شیوه‌های معتبر سنجش است. در پژوهش حاضر تلاش شده است تا به بررسی اعتبارسازه‌ایِ معیار نمره‌دهیِ مهارت نوشتن در آزمون رسمی پایان دوره‌ی مرکز زبان فارسی دانشگاه فردوسی پرداخته شود. به این منظور نتایج به‌دست آمده از یکی از آزمون‌های برگزار شده در این دانشگاه توسط مدل‌های آماری راش و تحلیل عاملی مورد بررسی قرار گرفت. نتایج تحلیل عاملی نشان داد که سه سازه‌ی کیفیت زبان، انسجام و ارتباط با موضوع برای سنجش مهارت نوشتن از میزان اعتبار بالایی برخوردار هستند. در این میان سازه‌ی انسجام با ۰,۹۸ بیشترین میزان و دو سازه‌ی دیگر هر کدام با ۰,۹۷ دومین میزان بار عاملی را داشتند. همچنین در معیار نمره‌دهی این آزمون بسندگی از یک مقیاس شش درجه‌ای برای نمره‌دهی هر یک از سازه‌ها استفاده شده است. مدل آماری راش نشان داد که هر یک از ارزیاب‌ها توانسته‌اند به شکل نسبتاً صحیحی از این مقیاس برای نمره‌دهی استفاده کنند، زیرا ترتیب آستانه‌ها مطابق ترتیب نمرات است و بهم‌ریختگی ندارد. از سویی دیگر نقشه‌ی آزمون‌دهنده پرسش گویای این امر بود که این مقیاس نمره‌دهی توانایی تمیز آزمون‌دهندگان ضعیف، متوسط و قوی را از یکدیگر دارد و مؤلفه‌ها و درجه‌های نمره‌گذاری (۰ تا ۵) همه‌ی گستره‌ی توانایی آزمون‌دهندگان را دربرمی‌گیرند و با این مقیاس می‌توان تمام داوطلبان با هر میزان توانایی نوشتن را اندازه گرفت. همچنین میزان پایایی نمره‌دهنده در این آزمون ۰,۹۶ برآورد شد که رقم بسیار مناسبی محسوب می‌شود.

واژگان کلیدی: سنجش زبان، اعتبار سازه‌ای، مهارت نوشتن، آزفا

تاریخ دریافت مقاله: ۱۳۹۸/۰۵/۰۷ تاریخ پذیرش نهایی مقاله: ۱۳۹۸/۱۱/۱۳

۱. رایانامه: mroodmajani@gmail.com
۲. رایانامه: Ghabool@Ferdowsi.um.ac.ir
۳. رایانامه: Ghonsooly@yahoo.com

شناسه دیجیتال (DOI):10.30479/jtpsol.2020.11294.1451
صص: ۴۶-۲۵

۱. مقدمه

در چند سال اخیر تعداد متقاضیان غیرفارسی‌زبان ورود به دانشگاه‌های ایران به شکل چشمگیری افزایش یافته‌است. تأسیس و راه‌اندازی مراکز بین‌المللی آموزش زبان فارسی در دانشگاه‌های مختلف، نمودی از این پدیده‌ی اجتماعی است. بدیهی است که رونق گرفتن این مراکز نه تنها حوزه‌ی زبان‌شناسی کاربردی را در زبان فارسی با پرسش‌های نوینی روبه‌رو می‌کند، بلکه شرایط بازاندیشی در مسائلی را فراهم می‌آورد که پیشتر تلاش‌هایی برای پاسخ‌گویی به آن‌ها صورت گرفته است. از آن جمله می‌توان به حوزه‌ی سنجش زبان فارسی اشاره کرد.

سنجش زبان[1] یکی از ارکان اساسی هر نظام آموزش زبان به شمار می‌آید. بخشی عمده‌ای از کارآمدی مراکز آموزشی در گرو بهره‌گیری از شیوه‌های صحیح سنجش و ارزشیابی است. در حقیقت نهادهای آموزش زبان فارسی با این پرسش اساسی روبه‌رو هستند که چگونه می‌توان مفاهیم انتزاعی دانش زبانی و توانش ارتباطی[2] را به عدد تبدیل کرد. آگاهی از میزان پیشرفت زبان‌آموزان، شناسایی نقاط ضعف آن‌ها و همچنین گرفتن تصمیمات دقیق درباره‌ی آن‌ها، نیازمند برخورداری از شیوه‌های صحیح و علمی سنجش و ارزشیابی[3] است.

در پژوهش حاضر تلاش می‌شود تا به بررسی اعتبارسازه‌ای معیار نمره‌دهیِ مهارت نوشتن در آزمون رسمی پایان دوره‌ی مرکز زبان فارسی دانشگاه فردوسی پرداخته شود. این آزمون که مورد تأیید وزارت علوم است، سالانه در دو نوبت در دانشگاه فردوسی مشهد، کشور عراق و در برخی از موارد درکشورهای دیگر برگزار می‌شود. این پژوهش سعی می‌کند تا به سه پرسش زیر پاسخ دهد:

۱. تا چه میزان سازه‌های تعریف شده در طراحی معیار نمره‌دهی، مؤلفه‌های متمایزی از مهارت نوشتاری را می‌سنجند؟

۲. مقیاس شش درجه‌ای برای نمره‌دهی سازه‌های شناسایی شده تا چه میزان می‌تواند زبان‌آموزان متوسط، ضعیف و قوی را از یکدیگر تمییز دهد؟

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

به منظور پاسخ‌دهی به پرسش‌های فوق نتایج به‌دست آمده از آزمون برگزار شده در تاریخ هشتم تیرماه ۱۳۹۸ در مرکز بین‌المللی زبان فارسی دانشگاه فردوسی، توسط مدل‌های آماری راش[4] و تحلیل عاملی مورد بررسی قرار گرفت. می‌دانیم که مدل آماری راش که ساختاری پیچیده‌تر از مدل‌های کلاسیک دارد، از یک

---

1. Language assessment
2. Communicative competence
3. Evaluation
4. Rasch

مزیت اساسی برخوردار است و آن وابسته نبودن به نمونه‌ی آماری پژوهش است (Bachman, 2004,139; Mousavi, 2012).

۲. مبانی نظری

به طور کلی طراحی هر معیار نمره‌دهی مهارت نوشتن می‌تواند دو مبنای اصلی داشته باشد. از یک سو باید تعریفی از این مهارت زبانی ارائه شود و سازه‌های تشکیل‌دهنده‌ی آن مشخص گردد؛ یعنی به کمک یک نظریه‌ی زبانی اولویت‌های اصلی نمره‌دهی تعیین شوند. به عنوان مثال اگر شخصی همانند لادو (Lado, 1961) و هریس (Harris, 1969) تعریف ساخت‌گرایانه از زبان ارائه دهد، اولویت‌های نمره‌دهی برای او اجزاء سازنده‌ی زبان یعنی واژه‌ها، ساختارهای دستوری و یا گفتمانی خواهد بود؛ اما اگر تعریف بسندگی زبان بر اساس نظریه‌های ارتباطی صورت گیرد، آنگاه توانایی انجام فعالیت‌های زبانی نیز از اولویت‌های نمره‌دهی خواهد بود (Canale & Swain, 1980; Canale, 1983; Bachman, 1990; Bachman & Palmer 1996; Celce-Murcia et al., 1995; Chappelle, 1997; Bachman & Palmer, 2010)

۲. ۱. مفهوم بسندگی

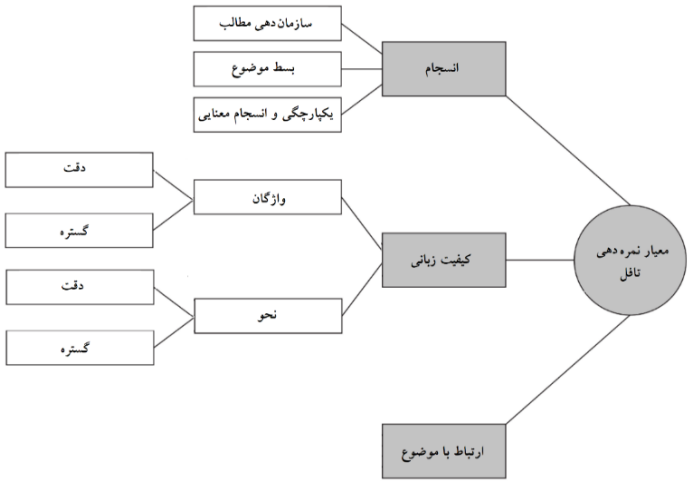
نخستین مدل سنجش زبان به اوایل دهه‌ی ۶۰ قرن بیستم برمی‌گردد. در آن دوره، آزمون‌سازی زبان تحت تأثیر دو جریان ساخت‌گرایی در زبان‌شناسی و رفتارگرایی در روان‌شناسی قرار داشت و باور بر این بود که می‌توان زبان را به عناصرکوچک‌ترسازنده‌ی آن، یعنی واج، تکیه، آهنگ، ساخت‌های نحوی و واژگانی تقسیم کرد و آن‌ها را به‌صورت جداگانه در قالب سؤال‌های صحیح و غلط، تست‌های چندگزینه‌ای و تصاویر، اندازه گرفت (Buck, 2001, p. 62). امروزه از این رویکرد با نام سنجش نکات مجزّا[1] یاد می‌کنند. اما از دهه‌ی هفتاد میلادی مدل‌های زبانی تحت تأثیر اندیشه‌های افرادی مانند هالیدی[2]، هایمز[3]، سرل[4] و آستین[5] دچار تغییرات عمده‌ای شدند و تعاریف ساخت‌گرایانه جای خود را به مدل‌های ارتباطی دادند (Richards & Rodgers, 2014, p. 84). در باور ارتباطی آنچه اهمّیت دارد، درک مطلوب پیام است. یعنی شنونده نه تنها باید اطّلاعات زبانی پیام را درک کند و از ساخت‌های زبانی آگاه باشد، بلکه باید بتواند اطّلاعات درک شده را با بافت ارتباطی کلام پیوند دهد.

---

1. Discrete point assessment
2. Halliday
3. Hymes
4. Searle
5. Austin

هدف اصلی زبان، برقراری ارتباط در یک موقعیت خاص و با یک محیط خاص است. آنچه اهمّیت دارد، دانش زبانی افراد و یا صحّت دستوری جملات آن‌ها نیست، بلکه توانایی آن‌ها در استفاده از زبان به‌منظور برقراری ارتباط در محیط زبانی مقصد است (Buck, 2001, p. 83).

توانایی به‌کارگیری زبان در محیط‌های مرتبط که هدف اصلی سنجش ارتباطی است. مفهومی بسیار فراتر از دانش آوایی، واژگانی و نحوی افراد را دربرمی‌گیرد. اگر هدف از سنجش پی بردن به این نکته است که داوطلبان تا چه میزان توانایی انجام فعالیت‌های زبانی موجود در محیط دانشگاه را دارند، بنابراین آنچه باید مورد سنجش قرار گیرد توانش ارتباطی آزمون‌دهندگان است.



شکل ۱. معیار نمره‌دهی آزمون تافل برگرفته از ای تی اس[۱](ETS, 2014, pp. 209-210)

اولین تأثیری که این شیوه‌ی نگرش به مفهوم بسندگی در نمره‌دهی آن می‌گذارد، قرار گرفتن بخشی با نام انجام تکلیف در معیارهای نمره‌دهی است. هدف از سنجش زبان تنها اندازه‌گیری دانش زبانی افراد نیست، بلکه باید توانایی به‌کارگیری ساختارهای زبانی متناسب با بافت مورد نظر نیز مورد بررسی قرار گیرد. از همین رو در معیارهای نمره‌دهی مهارت نوشتاری آزمون‌های شناخته شده‌ای مانند آیلتس[۲]، تافل[۳] بخشی با نام توانایی انجام تکلیف و یا ارتباط با موضوع، قرار گرفته است که میزان قابل ملاحظه‌ای از نمره را به خود اختصاص می‌دهد.

[1]. ETS: English Testing System
[۲]. IELTS
[۳]. TOEFL

از سوی دیگر همان‌طور که در شکل یک مشخص است، دانش زبانی به ساختارهای واژگانی و نحوی در سطح کلمه و جمله محدود نمی‌شود، بلکه ارتباط بین جمله‌ها و بندها را نیز دربرمی‌گیرد. فرد نه تنها باید بتواند واژگان و ساختارهای نحوی را به شکلی مناسب، دقیق و گسترده به‌کار گیرد، بلکه باید جمله‌ها و بندهای او از انسجام لازم برخوردار باشد. دانش افراد در سطحی فراتر از جمله توانش گفتمانی[۱] نام دارد و در سنجش زبان از اهمیت بسیاری برخوردار است.

بر اساس این مدل مهارت نوشتن به سه سازه‌ی کلی انسجام، کیفیت زبانی و ارتباط با موضوع تقسیم می‌شود. در بخش کیفیت زبانی توانایی نوشتاری فرد در سطح فروجمله مورد بررسی قرار می‌گیرد. تا چه میزان آزمون‌دهنده توانسته است واژگان و ساختارهای نحوی صحیح و متنوعی به‌کارگیرد. برای نمره‌دهنده در این بخش نه تنها صحت و دقت واژگان و ساختارهای دستوری متن نوشته شده از اهمیت برخودار است، بلکه تنوع و گستردگی ساختارهای زبانی نیز حائز اهمیت است. از سوی دیگر در بخش انسجام متن، زیرسازه‌هایی مانند سازمان‌دهی مطالب، بسط موضوع و یکپارچگی معنایی متن مورد توجه قرار می‌گیرد. منظور از سازمان‌دهی مطالب، رعایت اصول مقاله نویسی است. آیا متن نوشته شده، مقدمه، بدنه و نتیجه‌ی مناسبی دارد؟ در بخش مقدمه فرد باید موضوع نوشته‌ی خود را به روشنی معرفی کرده باشد، در بدنه به بررسی آن پرداخته و در پایان نیز با نتیجه‌گیری مناسبی نوشته‌ی خود را به اتمام رساند. علاوه بر این ارتباط معنایی بین بخش‌های مختلف متن، سازه‌ای با عنوان یکپارچگی متن را تشکیل می‌دهد. بخش‌های مختلف متن نباید یکدیگر را نقض‌کنند و نویسنده باید توانسته باشد انسجام معنایی مناسبی بین آن‌ها برقرار کند. سازه‌ی بسط موضوع نیز توانایی نویسنده از شیوه‌هایی مانند مثال‌آوری، تجربه‌ی شخصی، حکایت و حقایق علمی در توسعه‌ی ایده‌های اصلی متن است. آزمون‌دهنده باید بتواند نوشته‌ی خود را به‌خوبی بسط دهد.

آخرین سازه‌ی مورد نظر در سنجش مهارت نوشتاری، ارتباط متن با موضوع فعالیت زبانی است. متنی را تصور کنید که قرار بوده است درباره‌ی دلایل گسترش فقر در جامعه نوشته شود. ساختارهای خرد و کلان زبانی صحیح و مناسبی دارد. از انسجام معنایی خوبی نیز برخوردار است و نویسنده موضوع را به خوبی بسط داده است. اما این نویسنده به جای پرداختن به دلایل گسترش فقر در جامعه، به بحث درباره‌ی تأثیرات این پدیده پرداخته است. بنابراین مشکل این متن نه در ساختارهای زبانی، بلکه در ارتباط آن با موضوع است. در نظر گرفتن سازه‌ای با عنوان ارتباط متن با موضوع، حکایت از نگرشی ارتباطی به سنجش مهارت نوشتن دارد.

---

۱. Discourse competence

۲.۲. مفهوم اعتبار در سنجش زبان

اعتبار[1] یکی از محوری‌ترین و بنیادی‌ترین مفاهیم در حوزه‌ی‌سنجش زبان است (Fulcher & Davidson, 2007, p. 3; Lissitz, 2009, p. 1) که بر تمامی مراحل طراحی آزمون سایه می‌افکند. بدون داشتن درک کاملی از مفهوم اعتبار، طراحی علمی یک آزمون ناممکن است. فرآیند اعتبار همانند پلی است که عملکرد آزمون‌دهندگان در جلسه‌ی امتحان را به بخش پایانی طراحی آزمون؛ یعنی تفسیر نمره‌ها و تصمیم‌گیری پیوند می‌دهد.

در تعریف سنتی آزمونی معتبر است که آن چیزی را اندازه‌گیری کند که قصد داشته است اندازه بگیرد (Kelly, 1927, p. 14). همان‌طور که می‌دانیم آزمون ابزاری برای اندازه‌گیری دانش، رفتار و یا توانمندی افراد در یک حوزه‌ی خاص است. بنابراین هرگاه ما آزمونی طراحی می‌کنیم، تصمیم داریم چیزی را اندازه بگیریم و تحقیق درباره‌ی اعتبار آزمون، درحقیقت پاسخ‌گویی به این پرسش است که آیا آزمون ما آن چیزی را که قصد داشته، اندازه گرفته است یاخیر. تعریف فوق از مفهوم اعتبار بسیار مبهم است و باید به شکل دقیق‌تری بیان شود. از همین رو، پژوهشگران تلاش کردند تعاریف دیگری از این مفهوم بنیادین ارائه دهند.

در نگاه ساموئل مسیک[2] اعتبار، قضاوتی است ارزش‌گذارانه و یکپارچه از این‌که تا چه میزان مستندات تجربی و مبانی نظری، بسنده‌بودن و مناسب‌بودن استنباط‌ها و اقدامات صورت‌گرفته بر مبنای نمره‌های آزمون را تأیید می‌کنند (Messick, 1987: p. 1). در این تعریف از مفهوم اعتبار، این نکته حائز اهمیت است که اعتبار مفهومی یکپارچه و چندوجهی است. برخلاف نگاه سنتی که اعتبار را به سه دسته‌ی محتوا، سازه و معیار تقسیم می‌کرد، در باور مسیک، اعتبار مفهومی یکپارچه و منسجم است. او انواع مختلف اعتبار را زیرمجموعه‌ای از مفهوم اعتبار سازه می‌داند. مسیک درباره‌ی اعتبار محتوا می‌گوید این نوع اعتبار، قضاوت افراد متخصص درباره‌ی رابطه‌ی بین محتوای آزمون و محتوای حوزه‌ی رفتاری مورد نظر است. در اعتبار محتوا به بررسی این می‌پردازیم که تا چه میزان اقلام آزمون، حوزه‌ی مورد نظر را پوشش می‌دهند. بدیهی است در این نوع اعتبار پاسخ‌های آزمون‌دهندگان و نمره‌های آزمون مورد بررسی قرار نمی‌گیرند. بنابراین در نگاه مسیک اعتبار محتوا نمی‌تواند واجد شرایط مفهوم اعتبار باشد (Messick, 1987, p. 9).

باید این نکته را در نظر داشت که برخلاف نگرش سنتی که اعتبار را ویژگی خود آزمون می‌دانست، اعتبار در نگاه مسیک ویژگی تفسیرهای برآمده از آزمون و یا تصمیمات گرفته‌شده براساس نمره‌های آزمون است. بنابراین، به‌عنوان نمونه نمی‌توان از اعتبار آزمون تافل صحبت کرد؛ بلکه باید از اعتبار تفاسیر نمره‌های آزمون تافل و یا تصمیمات گرفته‌شده برمبنای آن سخن گفت. اگر نمره‌های آزمون تافل برای انتخاب دانشجویان واجد

---

1. Validity
2. Samuel Messick

شرایط ورود به دانشگاه‌های آمریکا مورد استفاده قرار گرفته باشد، معتبر و اگر برای تصمیم‌گیری درباره‌ی انتخاب افراد در یک شرکت تجاری به کار روند، ممکن است نامعتبر باشد. همان‌طور که گفتیم، ازآنجایی‌که در اعتبار محتوا، نمره‌های آزمون بررسی نمی‌شود، در نگاه مسیک این نوع اعتبار واجد شرایط مفهوم اعتبار نیست.

اعتبار معیار نیز برمبنای رابطه‌ی بین نمره‌های آزمون و نمره‌های معیار است و معمولاً در قالب همبستگی یا رگرسیون ارائه می‌شود. اگر فرض را بر این بگیریم که معیارهای بیرونی متعددی وجود داشته باشند، بنابراین ما با چندین اعتبار معیار روبه‌رو خواهیم بود. پس اعتبار معیار نمی‌تواند مفهوم ثابت و مشخصی باشد. از سوی دیگر اگر نمره‌های آزمونی با یک معیار، رابطه‌ی همبستگی بالایی داشته باشد؛ یعنی از اعتبار معیار بالایی برخوردار باشد، این امر نمی‌تواند تضمینی برای اعتبار محتوای آن آزمون باشد (Messick, 1987, p. 9). بنابراین، لزوماً رابطه‌ی مستقیمی بین اعتبار معیار یک آزمون و اعتبار محتوای یک آزمون و نیز اعتبار تفسیرها و تصمیمات برآمده از نمره‌های آزمون وجود ندارد. مسیک اعتبار معیار را نیز واجد شرایط مفهوم اعتبار نمی‌داند.

اعتبار سازه، بررسی این موضوع است که آزمون ما چه چیزی را اندازه می‌گیرد و چه مفاهیمی و چه سازه‌هایی مبنای عملکرد آزمون‌دهندگان در آزمون هستند (Messick, 1987, p. 8). اساس اعتبار سازه، گردآوری هرگونه مستندات و ادله‌ای است که بر معنابخشیدن و تفسیر نمره‌های آزمون تأثیر می‌گذارد (Messick, 1987, p. 10). بنابراین، اعتبار محتوا و اعتبار معیار می‌توانند ابزاری باشند که به درک اعتبار سازه کمک کنند. در نگاه مسیک اعتبار سازه، تقریباً تمام انواع دیگر اعتبار را دربرمی‌گیرد و این همان مفهوم یکپارچه‌ی چندوجهی مسیک در تعریف اعتبار آزمون است که ما در این پژوهش در نظر داریم.

۳. پیشینه‌ی پژوهش

ارزیابی مهارت نوشتاری غیرفارسی‌زبانان موضوع برخی از پژوهش‌ها مانند گل‌پور (Golpour, 2018)، جلیلی(Jalili, 2011)، گل‌پور(Golpour, 2014) و جلیلی(Jalili, 2017)، بوده است. هرچند هر کدام از این تحقیقات تلاش کرده‌اند تا معیاری برای نمره‌دهی این مهارت ارائه دهند؛ اما به طور کلی آن‌ها را می‌توان دارای سه ضعف اساسی دانست. در وهله‌ی نخست مشخص نیست که معیار نمره‌دهی معرفی شده در این تحقیقات براساس چه مبانی‌ای انتخاب شده‌اند. به عنوان مثال جلیلی به معیارهای محتوا، سازماندهی، انسجام، دایره‌ی لغات، دستور زبان، دقت مکانیکی اشاره می‌کند (Jalili, 2011, p. 146)؛ اما در این پژوهش معلوم نمی‌شود که این معیارها بر بنیان چه پژوهش‌هایی برگزیده شده‌اند. به نظر می‌رسد ذوق و سلیقه‌ی شخصی پژوهشگر تنها ملاک انتخاب معیارهای ارزیابی بوده است. سایر پژوهش‌های یادشده نیز شرایط مشابهی دارند. بنابراین بزرگترین ضعف مبانی نمره‌دهی مهارت نوشتاری غیرفارسی‌زبانان را می‌توان نقص در روش پژوهشی طراحی آن‌ها دانست. معیارهای معرفی شده بر اساس شیوه‌ای علمی انتخاب نشده‌اند.

علاوه بر این، میزان کارآمدی این مبانی نمره‌دهی در عمل نیز مشخص نشده است. آیا مؤلفه‌های متعدد تعیین شده در این پژوهش‌ها حقیقتاً سازه‌های متفاوتی را اندازه می‌گیرند؟ پرسشی که ماهیت این مبانی نمره‌دهی را با چالش جدی روبه‌رو می‌کند و در هیچکدام از پژوهش‌های صورت گرفته پاسخی به آن داده نشده است. همچنین نقطه‌ی ضعف دیگر این مبانی نمره‌دهی مشخص نبودن میزان پایایی آن‌هاست. نمره‌دهندگان تا چه میزان می‌توانند این معیارها را به شیوه‌ای پایا به‌کار گیرند؟ یک معیار نمره‌دهی زمانی از اعتبار علمی برخوردار است که بتواند برای پرسش‌های فوق، پاسخ‌های قانع‌کننده‌ای ارائه دهد. یعنی نه تنها بر اساس مبانی علمی طراحی شده باشد، بلکه اعتبار و پایایی آن نیز مورد بررسی قرار گرفته باشد. اتفاقی که تا به امروز در سنجش مهارت نوشتن غیرفارسی‌زبانان رخ نداده است.

در مقابل در زبان انگلیسی پژوهش‌های گسترده‌ای در ارتباط با معیارهای نمره‌دهی مهارت نوشتاری انجام شده است که در یک تقسیم‌بندی کلی می‌توان هر یک از آن‌ها را در سه حوزه‌ی مختلف ویژگی‌های متنی، فرآیند تصمیم‌گیری ارزیاب‌ها و نیز ارزیابی‌های الکترونیکی جای داد. دسته‌ی نخست، یعنی بررسی ویژگی‌های متنی متون نوشتاری که خود طیف بسیار گسترده‌ای از پژوهش‌ها را دربرمی‌گیرد، مشخصات خرد و کلان متون نوشتاری در سطوح مختلف زبانی را نشان می‌دهد.[۱] این پژوهش‌ها می‌تواند مبنایی برای طراحی معیارهای نمره‌دهی باشند. باید این نکته را در نظر داشت که آنچه در زبان فارسی تا به امروز مورد توجه بوده است عمدتاً دسته‌بندی و توصیف خطاهای زبان‌آموزان است (Motavallian & Malekian, 2013; Motavallian & Abarghouyie, 2013). تعیین خطاهای زبانی زمانی می‌تواند از اهمیت برخودار باشد که بتواند اطلاعاتی درباره‌ی سطوح زبانی در اختیار آزمون‌گران بگذارد. به عنوان مثال بتوان گفت که در سطح پیشرفته چه نوع خطاهایی را می‌توان انتظار داشت. از سویی دیگر منظور از ویژگی‌های متنی لزوماً خطاهای زبانی نیست، بلکه ویژگی‌های واژگانی، نحوی و انسجامی متن را نیز دربرمی‌گیرد. یعنی آزمونگران باید بدانند که یک متن پیشرفته یا متوسط از چه خصوصیت‌هایی برخوردار است تا بتوانند بر اساس آن‌ها نمره‌دهی کنند. کامینگ و همکاران (Cumming et al., 2000)در زبان انگلیسی به این ویژگی‌ها اشاره می‌کنند. به عنوان مثال در متون دانشگاهی معمولاً ایده‌ها به شکل مناسبی دسته‌بندی شده‌اند به طوری‌که رابطه‌ی بین ایده‌های اصلی و فرعی، نکات اصلی و زیر مجموعه‌های آن‌ها و همچنین حقایق بیان شده و مستندات، مشخص هستند. در بخش ابتدایی متن مقدمه‌ای سازمان‌یافته، برجسته و منسجم قرار دارد به عنوان مثال نوشته‌های استدلالی با بیان نظر و عقیده آغاز شوند. همچنین تعداد کلمات به‌کاررفته در متن، گستردگی و کیفیت آن‌ها، با توجه به زمان داده شده از سطح بالایی برخوردار است.

---

[۱]. رجوع کنید به کامینگ (۲۰۰۰).

دسته‌ی دیگر پژوهش‌های مربوط به معیار نمره‌دهی مهارت نوشتن به بررسی فرآیند تصمیم‌گیری ارزیاب‌ها و نیز عوامل تأثیرگذار در نمره‌دهی آن‌ها پرداخته‌اند. به عنوان مثال ویگل (Weigle, 1999)نحوه‌ی نمره‌دهی ارزیاب‌های باتجربه را با ارزیاب‌های کم‌تجربه مقایسه کرده است. او به این نتیجه رسیده است که در برخی مواقع ارزیاب‌های کم‌تجربه سختگیرانه‌تر نمره‌دهی می‌کنند. اردوسی (Erdosi, 2004) نیز به بررسی ارتباط بین نحوه‌ی نمره‌دهی ارزیاب‌ها و تجربیات پیشین فرهنگی و زبانی آن‌ها پرداخته است. پژوهش او تأثیر عمده‌ی این عوامل را در نمره‌دهی مهارت نوشتاری نشان می‌دهد. از سوی دیگر کامینگ و دیگران (Cumming et al., 2001) و کامینگ و دیگران (Cumming et al., 2002) نیز تلاش کرده‌اند تا با بررسی فرآیند تصمیم‌گیری ارزیاب‌ها، انگاره‌ی توصیفی از رفتارهای آن‌ها ارائه دهند. پژوهش آن‌ها مبنای اصلی طراحی معیار نمره‌دهی آزمون تافل بوده است.

دسته‌ی سوم پژوهش‌ها، به بررسی ارزیاب‌های الکترونیکی پرداختند. به‌کارگیری تکنولوژی در نمره‌دهی زبانی تحولی ارزشمند در حوزه‌ی زبانشناسی کاربردی محسوب می‌شود. از همین رو، در چند ساله‌ی اخیر پژوهش‌های بسیاری بر این موضوع تمرکز داشته‌اند. هرچند شباهت‌های بسیاری بین نمره‌دهی نرم‌افزارهای کامپیوتری و انسانی وجود دارد (Zhang, Breyer & Lorenz, 2013)، با این حال، اتالی (Attali, 2007)، نشان داده است که ارزیاب‌های الکترونیک در مقایسه با ارزیاب‌های انسانی پایایی بیشتری دارند. اما نکته‌ی حائز اهمیت این است که امکان به چالش کشیدن این تکنولوژی‌های نوین نیز وجود دارد (Powers et al., 2001).

۴. روش‌شناسی پژوهش

به منظور بررسی میزان اعتبارسازه‌ای بخش نوشتاری آزمون جامع زبان فارسی در دانشگاه فردوسی، نتایج به‌دست آمده از بخش نوشتاری یکی از آزمون‌های برگزار شده در مرکز بین‌المللی زبان فارسی دانشگاه فردوسی، توسط مدل‌های آماری راش و تحلیل عاملی مورد بررسی قرار گرفت. این آزمون در تاریخ ۸ تیر ۱۳۹۷در دانشگاه فردوسی مشهد و استراسبورگ فرانسه برگزار شد. پاسخ‌گویی به سؤال‌ها درساعت ۸:۱۵ آغاز گردید و داوطلبان در مدّت زمان حدود سه ساعت به ترتیب به سؤال‌های بخش‌های گوش کردن، خواندن و نوشتنِ آزمون پاسخ دادند. بخش صحبت‌کردن نیز پس از استراحت یک ساعته برگزار گردید. مهارت نوشتن بخش سوم آزمون را به خود اختصاص می‌داد و داوطلبان در مدت زمان حدود ۶۰ دقیقه به دو تکلیف آن پاسخ دادند. در تکلیف نخست، ابتدا یک فایل شنیداری برای داوطلبان پخش گردید سپس از داوطلبان خواسته شد تا خلاصه‌ای از آن بنویسند. در تکلیف دوم به داوطلبان موضوعی داده شد تا نظر خود را درباره‌ی آن در حدود

۲۰۰ کلمه بنویسند. نمونه‌ی سؤال‌های این آزمون را می‌توان در وبگاه رسمی مرکز بین‌المللی آموزش زبان فارسی دانشگاه فردوسی مشاهده کرد.

۴. ۱. 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 هرکدام به ترتیب، ۱۳، ۲، ۲، ۴ و ۵ شرکت‌کننده داشته‌اند. از نظر رشته‌ی تحصیلی، گروه علوم انسانی با ۶۸ نفر، پُرمخاطب‌ترین حوزه‌ی دانشگاهی را تشکیل می‌دهد. علوم مهندسی با ۲۳ و پزشکی با ۱۱ نفر در رتبه‌های بعدی از نظر تعداد شرکت‌کننده بوده‌اند.

جدول۱. شرکت‌کنندگان در آزمون بسندگی فارسی

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

۴. ۲. ابزار پژوهش

ابزار اصلی این پژوهش برای گردآوری اطلاعات، آزمون جامع زبان فارسی دانشگاه فردوسی است که سالانه در دو نوبت، یک‌بار در اواسط زمستان و بار دیگر در اواسط تابستان، در مرکز بین‌الملل دانشگاه فردوسی مشهد برگزار می‌شود. این آزمون چهار مهارت زبانی را دربرمی‌گیرد و نتیجه‌ی آن مورد قبول وزارت علوم است. بخش نوشتاری این آزمون شامل دو تکلیف[۱] می‌شود. تکلیف نخست، از نوع سنجش ترکیبی[۲] است و ابتدا آزمون‌دهندگان یک متن سخنرانی را می‌شنوند، سپس خلاصه‌ای از آن می‌نویسند. در تکلیف دوم زبان‌آموزان باید درباره‌ی یک موضوع اجتماعی، متنی در حدود ۲۰۰ کلمه تولید کنند. بخش نوشتن آزمون زبان دانشگاه فردوسی ۲۵ نمره از مجموع ۱۰۰ نمره‌ی کل آزمون را شامل می‌شود و مدت زمان پاسخ‌گویی آن ۶۰ دقیقه است. دو نمره‌دهنده، متن نوشته شده توسط آزمون‌دهندگان را با استفاده از معیار نمره‌دهی که در جدول شماره‌ی (۲) آمده است تصحیح می‌کنند.

---

۱. Task

۲. Integrated assessment

جدول ۲. معیارنمره‌دهی مهارت نوشتن در آزمون زبان فارسی دانشگاه فردوسی

نمره	کیفیت زبان	انسجام	ارتباط متن با موضوع تکلیف
۵	داوطلب زبان را به راحتی به کار می‌گیرد. متن او از تنوع نحوی برخوردار است. واژگان و اصطلاحات به شکلی مناسب و صحیح به‌کار رفته‌اند. هر چند ممکن است اشکالات واژگانی و نحوی مختصری دیده شود.	یکپارچه و منسجم است و توالی مطالب در آن رعایت شده است. سازمان‌یافته است و با کمک به‌کارگیری مثال‌ها و توضیحات و یا جزئیات به خوبی بسط یافته است.	متن نوشته شده به موضوع و هدف تکلیف مرتبط است.
۴	زبان به سهولت به کار گرفته می‌شود و از تنوع نحوی برخوردار است، واژگان و اصطلاحات به شکلی مناسب و صحیح به‌کار رفته‌اند. هر چند گاهی ممکن است اشکالات واضحی دیده شود، اما اختلالی در معنا ایجاد نمی‌کند.	یکپارچه و منسجم است و توالی مطالب رعایت شده است هر چند گاهی دچار حشو شده و یا از موضوع منحرف می‌شود و یا ارتباط مطالب واضح نیست. به طور کلی سازمان یافته و بسط یافته است. از توضیحات، مثال‌ها و جزئیات مناسبی استفاده شده است.	متن نوشته شده به موضوع و هدف تکلیف مرتبط است، هر چند برخی نکات ممکن است به خوبی توضیح داده نشده باشند.
۳	ممکن است ساختار جمله‌ها و انتخاب واژگان با مهارت صورت نگرفته باشد که این امر موجب ابهام و نامشخص بودن معنا می‌شود. ساختارهای نحوی و واژگانی دقیق، اما محدود هستند.	یکپارچه و منسجم است و توالی مطالب رعایت شده، هر چند ارتباط بین ایده‌ها ممکن است گاهی مبهم باشد. تا حدودی توانسته در بسط مطالب از مثال‌ها، توضیحات و جزئیات مرتبط با موضوع و هدف تکلیف استفاده کند.	در برخی مواقع متن نوشته شده به موضوع و هدف تکلیف مرتبط نیست.
۲	انتخاب واژگان و یا ساخت‌های واژگانی مشخصاً نامناسب هستند. اشکالات فراوان در ساخت‌های جملات و کاربرد آنها دیده می‌شود.	ارتباط بین ایده‌ها و سازمان‌دهی آنها نامناسب است. ارتباط بین ایده‌ها و سازمان‌دهی آن‌ها نامناسب است.	بسط موضوع محدود است. مثال‌ها، توضیحات و جزئیات ارائه شده برای پاسخ‌گویی به تکلیف مناسب و یا کافی نیستند.
۱	اشکالات جدی و بسیاری در ساختارهای جملات و کاربرد آن‌ها دیده می‌شود.	به طور جدی سازمان نیافته و بسط نیافته است.	جزئیات کم هستند و یا بدون جزئیات است، توضیحات نامرتبط هستند. پاسخ‌گویی روشنی به تکلیف داده نشده است.
۰	مقاله در این سطح کپی از صورت سؤال است. به موضوع نپرداخته و یا مرتبط با آن نیست. از زبان های دیگر برای پاسخ گویی استفاده کرده و یا سفید است.		

۴. ۳. 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، از یک مزیت اساسی برخوردار است و آن وابسته‌نبودن به نمونه‌ی

آماری پژوهش است (Bachman, 2004, p. 139). مدل‌های آماری کلاسیک مانند کودر ریچاردسون[۱]، روش دونیمه[۲]، آلفا کرونباخ[۳] و آزمون بازآزمون[۴] بر این فرض اساسی استوار هستند که نمره‌ی به‌دست آمده از یک آزمون حاصل جمع نمره‌ی حقیقی فرد و میزانی از خطا در محاسبه است. مبنا قراردادن چنین فرضی باعث شده است که این مدل‌ها دارای دو ضعف اساسی باشند. نخست این‌که توانایی تمییز خطاهای نظام‌مند از خطاهای تصادفی را نداشته باشند (Bachman, 1990, p. 186)، دوم این‌که با تغییر نمونه‌ی پژوهش نتایج به‌دست آمده نیز دچار تغییر شود. در مقابل مدل‌های راش این قابلیت را دارند که عملکرد آزمون‌دهندگان را در آزمون با توجه به توانمندی زبانی آن‌ها مورد بررسی قرار دهند؛ همین رویکرد باعث می‌شود که دیگر به نمونه‌ی آماری پژوهش وابسته نباشند.

به منظور بررسی اعتبار سازه‌ای آزمون از مدل تحلیل عاملی تأییدی استفاده می‌شود. این مدل ابزاری است برای بررسی رابطه‌ی مفروض بین عامل‌های مکنون و متغیرهای قابل مشاهده(Bachman, 2004, p. 113). روش‌تحلیل 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 به منظور آزمون فرضیه‌ها، از نرم‌افزار آموس[۵] که یکی از مشهورترین نرم افزارها جهت اجرای این گونه مدل‌ها است، استفاده شد.

۵. یافته‌های پژوهش

نرم‌افزار وینستپس[۶](Linacre, 2009)، نتایج به‌دست آمده از معیار نمره‌دهی نوشتن را توسط مدل آماری راش تحلیل کرد. باید این نکته را در نظر داشت که استفاده از مدل راش برای تحلیل داده‌ها دو پیش‌شرط دارد (Hamblton,1991, p. 10). نخست این‌که داده‌ها باید با مدل راش هم‌خوان بوده و از میزان برازش قابل قبولی برخوردار باشند. ستون‌های درون برازش[۷] و برون‌برازش[۸]، برازش پرسش‌ها به مدل راش را نشان می‌دهند که باید بین۰,۷۰تا ۱,۳۰باشند (Bond & Fox, 2008). برازش پرسش بدین‌معناست که پرسش با دیگر سؤال‌های آزمون هم‌ردیف است و به تعریف یک مقیاس برای سنجش سازه‌ی مورد نظر که در اینجا توانایی

---

۱. Kuder-Richardson
۲. Split half Experiment
۳. Cronbach’s Alpha
۴. Test-Retest experiment
5. Amos
6. WINSTEPS
7. Infit
8. Outfit

نوشتن است، کمک می‌کند. جدول (۲) نشان می‌دهد که همه‌ی پرسش‌ها، آمارهای درون‌برازش و برون‌برازش مناسبی دارند؛ بنابراین، آزمون به میزان قابل قبولی تک‌بُعدی است. شرط دیگر به‌کارگیری مدل آماری راش این است که هر سؤال آزمون از دیگر سؤالات مستقل باشد. یعنی پاسخ‌گویی به یک پرسش تحت تأثیر پاسخ‌گویی به پرسش‌های دیگر نباشد. این نکته نیز در مورد سؤالات بخش نوشتن آزمون زبان دانشگاه فردوسی صدق می‌کند.

برای محاسبه‌ی برازش مؤلفه‌ها، تعامل هر مؤلفه در هر تکلیف با هر نمره‌دهنده یک مؤلفه‌ی مجزا در نظر گرفته شد. در این آزمون دو تکلیف وجود دارد و هر تکلیف سه مؤلفه دارد که هر مؤلفه توسط دو نمره‌دهنده نمره داده شدند. بنابراین آزمون ۱۲ مؤلفه دارد. جدول (۳) نشان می‌دهد که فقط مؤلفه‌ی اول/ نمره‌دهنده‌ی اول در تکلیف یک، از برازش مناسب برخوردار نیست. به عبارت دیگر نمره‌دهنده‌ی اول در تکلیف (۱) مؤلفه‌ی اول را ناهماهنگ و سختگیرانه‌تر نسبت به دیگر مؤلفه‌ها و نمره‌دهنده‌ی دوم نمره‌گذاری کرده است. بقیه‌ی مؤلفه/ نمره‌دهندگان از برازش مناسب برخوردارند که این خود گواهی است بر تک‌بعدی بودن آزمون. پایایی راش آزمون ۰,۹۶ به دست آمد که بسیار خوب است و نشان‌دهنده‌ی هماهنگی بین دو نمره‌دهنده در نمره‌گذاری و تکرارپذیری نمرات آزمون دهندگان در اجرای دوباره‌ی آزمون است.

جدول ۳. مقادیر درون برازش و برون برازش برای هر مؤلفه/ نمره‌دهنده

شماره تکلیف	شماره نمره‌دهنده	عنوان مؤلفه	درون برازش	برون برازش	ضریب تمیز
اول	اول	شیوه‌ی بیان	۱,۴۹	۱,۴۱	۰,۸۶
		کیفیت زبان	۰,۹۲	۰,۸۷	۰,۸۹
		بسط موضوع	۱,۰۰	۰,۹۶	۰,۸۹
	دوم	شیوه‌ی بیان	۱,۱۷	۱,۱۱	۰,۸۷
		کیفیت زبان	۰,۸۰	۰,۷۶	۰,۹۰
		بسط موضوع	۰,۹۰	۰,۸۳	۰,۹۰
دوم	اول	شیوه‌ی بیان	۱,۱۰	۱,۱۰	۰,۸۵
		کیفیت زبان	۰,۸۳	۰,۸۰	۰,۸۵
		بسط موضوع	۱,۲۳	۱,۲۳	۰,۷۹
	دوم	شیوه‌ی بیان	۰,۸۲	۰,۸۳	۰,۸۶
		کیفیت زبان	۰,۷۹	۰,۸۰	۰,۸۶
		بسط موضوع	۰,۹۸	۱,۰۰	۰,۸۲

۵. ۱. مقیاس نمره‌دهی

در آزمون نوشتن از یک مقیاس شش درجه‌ای برای نمره‌دهی استفاده شد. بررسی عملکرد مقیاس و یا استفاده‌ی درست نمره‌دهندگان از مقیاس، بخش مهمی از تحلیل آزمون‌های عمکردمحور است. جدول (۳) ویژگی‌های

مقیاس استفاده شده در این آزمون برای نمره‌دهی را نشان می‌دهد. اولین ستون از راست نمره‌ی هر درجه از مقیاس را نشان می‌دهد که بین ۰ تا ۵ است. ستون دوم تعداد دفعاتی است که هر نمره داده شده است. مثلاً نمره‌ی صفر ۱۷۸ بار و نمره‌ی پنج ۴۵ 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 افرادِ تواناتر، قاعدتاً نمره‌ی بیشتری گرفته‌اند. ستون چهارم دشواری آستانه‌ها را نشان می‌دهد. دشواری آستانه‌ها باید به ترتیب نمره اضافه شود. به‌هم‌ریختگی ترتیب بزرگی آستانه‌ها نشانه‌ی استفاده نادرست از مقیاس است (Linacre, 2009). آستانه‌ها نقاطی بر روی طیف توانایی هستند که احتمال دو نمره کنار هم در آن نقاط ۵۰ درصد است. نمره‌ی اول، یعنی (۰) آستانه‌ای ندارد، چون قبل از آن نمره‌ی دیگری نیست. آستانه بین ۰ و ۱ دشواری آن ۴٫۶۷- است. یعنی اگر فردی توانایی او در نوشتن برابر این مقدار باشد به احتمال ۵۰ درصد نمره‌دهنده به او نمره‌ی ۰ و یا ۱ می‌دهد. به عبارت دیگر برای این‌که یک آزمون‌دهنده نمره‌ی ۰ یا ۱ بگیرد (در هر تکلیف و در هر مؤلفه) باید توانایی‌اش ۴٫۶۷- باشد. اگر فردی توانایی‌اش ۲٫۴۵- منفی باشد، ۵۰ درصد احتمال دارد ۱ و یا ۲ بگیرد.

اگر از مقیاس نمره‌گذاری درست استفاده شده باشد، انتظار می‌رود که دشواری آستانه‌ها به ترتیب با اندازه‌ی نمره زیاد شود که در اینجا همین‌گونه است. نمرات بین ۰ تا ۵ هریک معرف میزان متفاوتی از توانایی نوشتن هستند و جای متفاوتی بر روی مقیاس دارند. نمره‌دهندگان باید بتوانند تفاوت‌های بین این مقادیر را تشخیص دهند و از مقیاس برای بیان تفاوت‌های بین آزمون دهندگان در میزان سازه‌ی مورد اندازه‌گیری درست استفاده کنند. همان‌طور که جدول (۴) نشان می‌دهد ترتیب آستانه‌ها مطابق ترتیب نمرات است و به‌هم‌ریختگی ندارد که این خود گواهی بر استفاده درست از مقیاس و نمره‌گذاری صحیح آزمون است.

جدول ۴. آماره های مقیاس نمره‌گذاری

مقیاس نمره‌گذاری	تعداد دفعات	میانگین توانایی برای کسب نمره	آستانه
۰	۱۷۸	-۴٫۹۰	بدون آستانه
۱	۳۴۰	-۳٫۱۶	-۴٫۶۷
۲	۳۹۴	-۱٫۲۴	-۲٫۴۵
۳	۲۲۰	۰٫۶۷	۰٫۱۴
۴	۷۱	۲٫۶۸	۲٫۹۴
۵	۴۵	۳٫۹۳	۴٫۰۴

۵. ۲. نقشه‌ی آزمون‌دهنده-پرسش

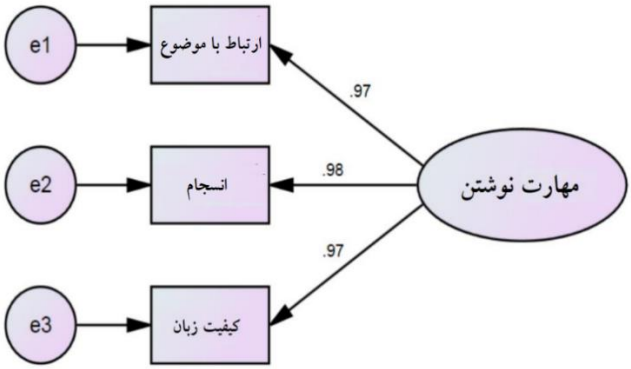
شکل (۲) نقشه‌ی رایت[۱] یا آزمون‌دهنده ـ پرسش است. در این نقشه سطح توانایی زبانی آزمون‌دهندگان و سطح دشواری سازه‌های موجود در معیار نمره‌دهی که توسط هریک از نمره‌دهندگان به آزمون‌دهندگان اختصاص داده شده است، به طور همزمان نمایش داده می‌شود و درنتیجه به‌طور مستقیم قابل مقایسه هستند. در سمت راست نقشه، سطح سختی سازه‌ها و در سمت چپ، سطح توانایی آزمون‌دهندگان قرار دارند. مؤلفه‌های پایین نقشه نشان‌دهنده‌ی سازه‌های ساده‌تر و آزمون‌دهندگان ضعیف‌تر و مؤلفه‌های بالای نقشه نشان‌دهنده‌ی سازه‌های دشوارتر و آزمون‌دهندگان قوی‌تر است. (w1) سازه‌ی ارتباط با موضوع برای نمره‌دهنده‌ی اول است. همچنین (w2) سازه‌ی انسجام معنایی و(w3) کیفیت زبان برای این ارزیاب است. w1.1 نشاندهنده‌ی نمره‌ی یک برای سازه‌ی ارتباط با موضوع است که توسط نمره‌دهنده‌ی اول اختصاص یافته است. به همین ترتیب w1.2 نشان‌دهنده‌ی نمره‌ی دو برای این سازه است. از سوی دیگر w4، w5، w6 به ترتیب نمره‌هایی هستند که نمره‌دهنده‌ی شماره‌ی دو به هریک از سازه‌های ارتباط با موضوع، انسجام معنایی و کیفیت زبان داده است. بنابراین، w4.3، یعنی نمره‌ی ۳ برای سازه‌ی ارتباط با موضوع که توسط نمره‌دهنده‌ی شماره‌ی 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 مورد بازبینی قرار بگیرد. نقشه نشان می‌دهد که مؤلفه‌ها و درجه‌های نمره‌گذاری (۰ تا ۵) همه‌ی گستره توانایی آزمون‌دهندگان را در برمی‌گیرند و با این مقیاس می‌توان تمام داوطلبان با هر میزان توانایی نوشتن را اندازه گرفت.

۱. Wrightmap

شکل ۲. نقشه‌ی آزمون‌دهنده- پرسش

۵. ۳. تحلیل عاملی تأییدی

مدل اندازه‌گیری بخش نوشتن آزمون در حالت تخمین استاندارد در شکل زیر نشان داده شده است. در این مدل مسئله‌ی مهم در درجه‌ی اول مقادیر بار عاملی بین عوامل و سؤالات مربوطه است که تمامی مقادیر بالاتر از ۳۰ درصد است. با توجه به بارهای عاملی می‌توان میزان اهمیت و تأثیر هریک از ابعاد برای متغیر مورد نظر را مشخص کرد. بعد انسجام مهمترین عامل در بخش نوشتن (۰,۹۸۴) است.



شکل ۳. تحلیل عاملی تأییدی بخش نوشتن

در جدول زیر مقادیر بارهای عاملی، برآورد پارامترها و معنی‌داری آن‌ها توسط آزمون *تی*[1] گزارش می‌شود.

جدول ۵. معنی‌داری برآورد پارامترهای بخش نوشتن

متغیر اول	بار عاملی	ضریب رگرسیون	خطای برآورد	مقدار آماره‌ی تی
ارتباط با موضوع	۰,۹۶۹	۰,۹۳۷	۰,۰۲۵	۳۹,۵۹۲
انسجام	۰,۹۸۳	۰,۹۸	۰,۰۲۲	۴۵,۵۴۷
کیفیت زبان	۰,۹۷۲	۱		

نتایج جدول فوق نشان می دهد که هر سه بعد بخش نوشتن تأثیر معنی‌داری در سطح خطای ۵ درصد دارند. اعتبار این مدل توسط شاخص‌های نیکویی برازش در جدول زیر گزارش شده است.

جدول ۶. شاخص‌های نیکویی برازش تحلیل عاملی تأییدی بخش نوشتن

شاخص	کای دو بر درجه‌ی آزادی	میانگین مجذور خطاهای مدل	شاخص نیکویی برازش
مقدار	۱,۷۳۶	۰,۰۶۱	۰,۹۹۴
حالت مطلوب	$1 \leq \& \leq 3$	$0 \leq \& \leq 0.08$	$0.9 \leq \& \leq 1$

[1] T

با توجه به خروجی نرم‌افزار که در جدول فوق ارائه شده، شاخص‌های نیکویی برازش مدل در حد مطلوب قرار دارند. با توجه به شاخص‌ها و خروجی‌های نرم‌افزار آموس، می‌توان گفت که داده‌ها با مدل منطبق هستند و شاخص‌های ارائه شده نشان‌دهنده‌ی این موضوع هستند که درمجموع مدل ارائه شده مدل مناسبی است و داده‌های تجربی به‌خوبی با آن منطبق هستند.

۶. نتیجه‌گیری

تاکنون در زبان فارسی چندین معیار برای نمره‌دهی مهارت نوشتاری غیرفارسی‌زبانان ارائه شده است (Jalili, 2011; Golpour, 2014; Jalili, 2017; Golpour, 2018)، اما اعتبار هیچکدام از آن‌ها توسط روش‌های پژوهشی کمّی، مشخص نشده است. در این پژوهش تلاش شد تا اعتبار سازه‌ای آزمون جامع زبان فارسی دانشگاه فردوسی، مورد بررسی قرار گیرد. معیار نمره‌دهی مهارت نوشتاری در این آزمون بر مبانی اصول نظری آزمون طراحی گردیده است. در این معیار مهارت نوشتاری از سه سازه‌ی کیفیت زبان، انسجام و ارتباط با موضوع تشکیل شده است (ETS, 2014) . بدیهی است که این معیار برگرفته از نگاهی ارتباطی به مفهوم زبان است؛ یعنی نه تنها دانش زبانی افراد در سطوح فروجمله، جمله و گفتمان را می‌سنجد، بلکه توانایی انجام فعالیت زبانی را نیز مورد ارزیابی قرار می‌دهد. بر اساس این معیار نمره‌دهندگان علاوه بر کیفیت زبان و انسجام متن به بررسی این نکته می‌پردازند که متن نوشته شده تا چه میزان با هدف فعالیت زبانی هم‌خوانی دارد.

نتایج تحلیل عاملی نشان داد که سه سازه‌ی مشخص شده از میزان اعتبار بالایی برخوردار هستند؛ یعنی می‌توان مهارت نوشتاری را به سازه‌های کیفیت زبان، انسجام و ارتباط با موضوع تقسیم کرده و آن‌ها را به صورت جداگانه نمره‌دهی کرد. در این میان سازه‌ی انسجام با ۰,۹۸ بیشترین میزان و دو سازه‌ی دیگر هر کدام با ۰,۹۷ دومین میزان بار عاملی را داشتند. بر اساس این آمار می‌توان گفت که تقسیم مهارت نوشتاری به سازه‌های کیفیت زبان، انسجام و ارتباط متن با موضوع تقسیم‌بندی معتبری است و هر کدام از این مؤلفه‌ها، سازه‌ی جداگانه‌ای را اندازه‌گیری می‌کند.

در معیار نمره‌دهی آزمون بسندگی زبان فارسی دانشگاه فردوسی، از یک مقیاس شش درجه‌ای برای نمره‌دهی سازه‌ها استفاده شده است. مدل آماری راش نشان داد که هریک از نمره‌دهندگان توانسته‌اند به شکل نسبتاً صحیحی از این معیار برای نمره‌گذاری استفاده کنند، زیرا ترتیب آستانه‌ها مطابق ترتیب نمرات است و به‌هم‌ریختگی ندارد. از سویی دیگر نقشه‌ی آزمون‌دهنده-پرسش، گویای این امر بود که این مقیاس نمره‌دهی توانایی تمییز آزمون‌دهندگان ضعیف، متوسط و قوی را از یکدیگر دارد و مؤلفه‌ها و درجه‌های نمره‌گذاری (۱

تا ۵) همه‌ی گستره‌ی توانایی آزمون‌دهندگان را دربرمی‌گیرند و با این مقیاس می‌توان تمام داوطلبان با هر میزان توانایی نوشتن را اندازه گرفت.

از سوی دیگر میزان پایایی نمره‌دهنده در این آزمون ۰,۹۶ برآورد شد که رقم بسیار قابل قبولی به‌شمار می‌رود. این نتیجه نشان می‌دهد که نمره‌دهندگان از معیار نمره‌دهی به شکل همسانی استفاده کرده‌اند و در تعریف سازه‌های مشخص شده و درجه‌بندی نمره‌دهی آن‌ها اتفاق نظر دارند. البته باید این نکته را در نظر داشت که نتیجه‌ی به‌دست آمده مربوط به نمره‌گذاریِ دو نمره‌دهنده است. اگر تعداد نمره‌دهندگان بیشتر باشد، مسلماً رقم به‌دست آمده تغییر خواهد کرد.

پژوهش حاضر تلاشی برای بررسی میزان اعتبار معیار نمره‌دهی به‌کاررفته در بخش نوشتاری آزمون جامع زبان فارسی در دانشگاه فردوسی است. بر اساس نتایج این پژوهش، میزان اعتبار این معیار نمره‌دهی تا حدودی روشن شده است. با این حال رسیدن به سنجش دقیق از مهارت نوشتاری زبان فارسی برای غیرفارسی‌زبانان مستلزم پژوهش‌های دیگری نیز هست. در وهله‌ی نخست باید مشخصات متنی متون نوشتاری فارسی مشخص باشد. به عبارتی ما باید بدانیم که متون نوشتاری غیرفارسی‌زبانان در سطوح زبانی مقدماتی، متوسط و قوی چه ویژگی‌های واژگانی، دستوری و گفتمانی‌ای دارند. جای پژوهش‌های مانند کامینگ و همکاران (Cumming et al., 2000) در زبان فارسی خالی است. از سوی دیگر استفاده از تکنولوژی نوین رایانه‌ای می‌تواند کمک شایانی به نمره‌دهی مهارت نوشتن بکند. ارزیاب‌های الکترونیکی تا کنون در زبان فارسی مورد استفاده قرار نگرفته‌اند. امید است تا در آینده شاهد طراحی ابزارهای دقیق سنجش زبان فارسی باشیم.

فهرست منابع:

جلیلی، سیّد اکبر. (۱۳۹۰). *آزمون مهارتی فارسی (آمفا) بر پایه‌ی چهار مهارت اصلی زبانی* (پایان‌نامه‌ی کارشناسی ارشد).دانشگاه علاّمه طباطبائی، تهران، ایران.

جلیلی، سیّد اکبر. (۱۳۹۶). ارزیابی تولیدات نوشتاری فارسی‌آموزان سطح پیشرفته: طراحی یک چارچوب جزئی‌نگر. *پژوهش‌نامه‌ی آموزش زبان فارسی به غیرفارسی‌زبانان*. ۶ (۱) (پیاپی ۱۳)، ۶۴-۳۱.

گل‌پور، لیلا. (۱۳۹۴). *طرّاحی و اعتباربخشی آزمون بسندگی زبان فارسی بر پایه‌ی چهار مهارت زبانی* (پایان‌نامه‌ی دکتری). دانشگاه پیام‌نور مرکز، تهران، ایران.

گل‌پور، لیلا. (۱۳۹۷). طراحی آزمون مهارت نوشتاری ویژه‌ی غیرفارسی‌زبانان: راهکارها و تحلیل خطاها. *پژوهش‌نامه‌ی آموزش زبان فارسی به غیرفارسی‌زبانان*. ۷ (۲) (پیاپی ۱۶)، ۶۸-۴۵.

متولیان نائینی، رضوان و استوار ابرقویی، عباس. (۱۳۹۳). نقش تداخل در پیدایش خطاهای نحوی در نگارش فارسی‌آموزان عرب‌زبان. *پژوهش‌های زبانشناختی در زبان‌های خارجی*. ۳ (۲). ۳۸-۳۵۱.

متولیان نائینی، رضوان و ملکیان، رسول.(۱۳۹۳). تحلیل خطاهای نحوی فارسی‌آموزان اردوزبان. در *پژوهش‌نامه‌ی آموزش زبان فارسی به غیرفارسی‌زبانان*. ۳ (۱) (پیاپی ۶)، ۵۶-۲۹.

References:

Attali, Y. (2007).*Construct validity of e-rater in scoring TOEFL essays.* (TOEFL Research Rep. 07-21). Princeton, NJ: Educational Testing Service.

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.

Bachman L. & Palmer A. (1996). *Language testing in practice*. New York: Oxford University Press.

Bachman L. & Palmer A. (2010). *Language assessment in practice*. New York: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah,NJ: Lawrence Erlbaum.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Canale, M. (1983). From Communicative Competence to Communicative Language Pedagogy. In Richards, C. and Schmidt, R. W. (eds) *Language and Communication*. London: Longman. 2-27

Canale, M. & Swain, M. (1980). Theoretical basis of Communicative Approaches to second language teaching and testing.*Applied Linguistics* 1, 1, 1-47

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative Competence: a pedagogical motivated model with content specifications. *Issues in Applied Linguistics* 2, 5-35.

Chapelle, C., Grabe, W., & Berns, M. (1997). Communicative language proficiency: definition and implications for TOEFL 2000. (TOEFL Monograph No. 10) Princeton, NJ: Educational Testing

Cumming, A., Kantor, R., Powers, D., Santos, T., &Taylor, C. (2000). *TOEFL 2000 writing framework: a working paper* (TOEFL Monograph No. 18). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. (TOEFL Monograph No. 22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67-96.

ETS: Educational Testing System. (2012). *The official Guide to the TOEFL test*. (4th ed). New York: Mc Graw Hill.

Erdosy, M. U. (2004). *exploring variability in judging writing ability in a second language: a study of four experienced raters of ESL compositions*. (TOEFL Research Rep. No.70). Princeton, NJ: Educational Testing Service.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment and advanced resource Book.* New York: Routledge.

Golpour, L. (2018). Developing of a writing skill test for non-Persian learners: Approaches and analysis of errors. *Journal of teaching Persian to Speaker of Other Languages.(7) 2, 45-68.* [in Persian].

Golpour, L. (2014). *Designing and validating Persian proficiency test based on four language skills*. (PhD. Dissertation). Pyamnour markaz University, Iran. [in Persian].

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

Hambleton, R. K., Swaminathan, H. & Rogers, J.(1991*).Fundamentals of Item Response Theory*. Newbury Park: Sage Publication.

Jalili, A. (2017). Assessing advanced Persian language learners written production: developing a detailed rubric. *Journal of teaching Persian to Speaker of Other Languages.(6) 1, 158.* [in Persian].

Jalili, A. (2011). Persian Language proficiency test based on four main langage skills. (MA. Dissertation). Allameh Tabataba'I University, Iran. [in Persian].

Kelly, T. L. (1927). *Interpretation of educational Measurements*. New York: World Book Company.

Lado, R. (1961). *Language Testing*. London: Longman.

Linacre, J. M. (2009). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.

Lissitz, R. W. (ed.), (2009). *The concept of validity: revisions new directions and applications*. Charlotte, NC: Information Age Publishing, INC.

Messick, S. (1987). *Validity* (Report no. RR-87-40). Princeton: ETS.

Motavallian Nayini, R., & Abarghouyi, A. (2013). The study of Persian syntactic errors by Arabic – speaking learners, *Journal of teaching Persian to Speaker of Other Languages*.(2) 2.[in Persian].

Motavallian Nayini, R., & Malekian R. (2013). Syntactic error analysis of urdu-speaking learners of Persian. *Journal of teaching Persian to Speaker of Other Languages.(3) 1, 31-64.* [in Persian].

Mousavi, S. A. (2012). Item Response Theory. In *An Encyclopedic Dictionary of Language Testing*. 5th ed. Tehran. Rahnama.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukish, k. (2000). *Comparing the validity of automated and human essay scoring*. (GRE No.98-08 aR). Princeton, NJ: Educational Testing Service.

Richards J. C. & Rodgers, T., S. (2014). *Approaches and methods in language teaching.* Third ed. Cambridge: Cambridge University press.

Weigle, S, C. (1999). Investigating Rater prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*. 6(2).145-178.

Zhang, M., Breyer, F. J.,& Lorenz, F. (2013). *Investigating the suitability of implementing the e-rater scoring engine in a large scale English language testing program*. (TOEFL Research Rep. 13-36). Princeton, NJ: Educational Testing Service.