THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES
Vol. 9, Nº 3, 2021, pp. 465–473

UDC: 811.111'37 https://doi.org/10.22190/JTESAP2103465H

Original scientific paper

SYSTEMATIC RESEARCH SYNTHESIS ON RATING IN ESP SPEAKING ASSESSMENTS

Yuko Hijikata, Jiyu Min

University of Tsukuba, Faculty of Humanities and Social Sciences, Japan,
The Ohio State University, USA
E-mail: hijikata.yuko.fe@u.tsukuba.ac.jp, min.145@buckeyemail.osu.edu

Abstract. *Jacoby and McNamara (1998) insist that domain specialists (e.g., doctors) and language specialists (e.g., teachers) differ in their approaches to evaluating profession-specific communication tasks. However, due to the scarcity of studies examining speaking assessments for occupational purposes, the differences between the two rater groups have not yet been clearly revealed. In response to this gap in the literature, this systematic research synthesis study examines how the workplace speaking skill has been evaluated, focusing on rater groups and rating scales. The major findings are as follows. First, research on ESP speaking assessments tends to include more than one rater group such as domain specialists and language specialists. Second, while domain specialists and language specialists generally demonstrate high intergroup correlations, the rater group notably differs in terms of field-specific criteria. Third, compared with linguistic scales, field-specific criteria have not been developed. Based on these results, directions for future research are discussed.*

Key words: *English for specific purposes, English for occupational purposes, speaking, assessment, rating, systematic research synthesis*

1. INTRODUCTION

Language assessment for professional purposes (LAPP) is defined as "any assessment process, carried out by and for invested parties, which is used to determine a person's ability to understand and/or use the language of a professionally-oriented domain to a specified or necessary level" (Knoch and Macqueen 2020, 3). Is LAPP, or language for specific purposes (LSP) assessment, different from other general language assessments? Douglas (2013) claimed that they are similar in that test developers must consider the purposes of tests, characteristics of test-takers, the target language-use situations, and the reliability, validity, and impact of each potential test. However, he also stated that these two types of test differ in the two aspects: (a) the authenticity of a task and (b) the interaction between language knowledge and domain-specific content knowledge.

Although there has been an increasing need to develop performance-based workplace English assessments, it is difficult to develop speaking assessments of English for occupational

Submitted January 10th, 2021, accepted for publication March 1st, 2021
Corresponding author: Yuko Hijikata. University of Tsukuba, Faculty of Humanities and Social Sciences, Japan
E-mail: hijikata.yuko.fe@u.tsukuba.ac.jp

© 2021 by University of Niš, Serbia | Creative Commons License: CC BY-NC-ND

purposes (EOP). The difficulty largely stems from the fact that developers of ESP assessments are not necessarily 'experts' in the learners' fields and thus do not have sufficient knowledge for understanding a learner's communication in their workplace. For example, language professionals are not familiar with technical terms used at a certain workplace, necessary language function, or real conversations because they are not experts in the relevant domain (e.g., medicine or engineering). Nevertheless, language teachers or language program developers must develop an authentic context for EOP learners and appropriate speaking assessments. In order to ensure that ESP practitioners can collect language samples and understand learners' requirements at their workplaces, a needs-analysis is generally conducted (e.g., Cowling 2007). However, discrepancies can be noted between highly customized learning materials, which use authentic content and tasks related to the jobs at hand, and standardized tests targeting a general work-related language, such as Comprehensive Adult Students Assessment System or Adult Basic Learning Examination (Ekkens and Winke 2009).

Another significant challenge is evaluating learners' speaking performances without content knowledge in the specific job-related domain, such as hospitality industries, health medical fields, businesses, or construction companies. Jacoby and McNamara (1999) proposed the notion of "indigenous criteria," which emphasizes elements closely related to a specific domain of the language (Elder and McNamara 2016), particularly for the Occupational English Test (OET). Furthermore, Douglas (2000) admitted that the most crucial and complicated concern in developing LSP speaking tests is developing evaluation criteria or rating scales. Douglas (2001) insisted that assessment criteria must be derived by incorporating a target language to use as the test content. This issue raised the following question: Who should be trained and included as a rater in ESP speaking tests? Domain specialists are experts in the subject area but not in the language. By contrast, language professionals are familiar with evaluating speaking performances and developing tasks used in speaking tests. However, they are not specialists in the specific fields of the test-takers.

The effects of raters on speaking assessments have been examined in many studies (e.g., Duijm et al. 2018), although most did not have a specific focus on ESP. Duijm et al. (2018) compared linguistically trained and non-trained raters in terms of accuracy and fluency in speaking. They found that the former group focuses on accuracy while the latter group paid more attention to fluency. In'nami and Koizumi (2016) conducted a meta-analysis focused on generalizability studies in speaking and writing. Their analysis compared the impacts of tasks and task-related interaction and raters and rater-related interaction to see which was more responsible for score variances and found that task and task-related interaction had greater influence. They also compared types of tasks, contexts, and scoring methods to the degrees of person-by-task interactions and found that person-by-task interactions had a greater influence on assessments when they were based on both academic and general contexts compared to interactions based only on academic contexts.

Considering the above, we predict that, in the occupational context, the effects of raters would be large depending on whether the rater had much background knowledge in each professional domain and whether the rater had experience with evaluating language performance. To examine this possibility, we need to systematically review the effects of raters and evaluation criteria. To the best of our knowledge, no synthesis has been performed regarding ESP speaking assessment rating. Therefore, in this systematic research synthesis study, which examined the existing literature in ESP speaking assessment, we addressed the following research questions (RQs):

RQ1: Which group—domain specialists, language specialists, or both—served as raters in EOP speaking assessment research?

RQ2: What assessment criteria have been used in research on EOP speaking assessments?

RQ3: Do domain and language specialists differently rate speaking performance?

2. METHOD

This study adopted the systematic research synthesis to synthesize the existing literature in ESP speaking assessments. The Educational Resources Information Center (ERIC), PsycINFO, and ScienceDirect were used to acquire data because they have been frequently used in previous research syntheses and meta-analyses. We first examined papers published before August 2019 in 10 peer-reviewed journals in applied linguistics: *Applied Linguistics, English for Specific Purposes, Language Assessment Quarterly, Language Learning, Language Teaching Research, Language Testing, Modern Language Journal, Studies in Second Language Acquisition, System, and TESOL Quarterly*. Among these journals, seven were included in this synthesis based on the selection of comprehensive international peer-reviewed journals in applied linguistics by Hashemi and Babaii (2013): *Applied Linguistics, English for Specific Purposes, Language Learning, Language Teaching Research, Language testing, Modern Language Journal,* and *TESOL Quarterly*. Considering the focus of this synthesis (i.e., speaking assessment) and In'nami and Koizumi's (2009) meta-analysis, three journals (*Language Assessment Quarterly, Studies in Second Language Acquisition,* and *System*) were additionally selected. The following 15 keywords were used to retrieve empirical studies: (1) ESP assessment, (2) speaking assessment, (3) oral assessment, (4) ESP testing, (5) speaking test, ESP, (6) speaking test, professional English, (7) oral test, ESP, (8) spoken fluency, (9) speaking fluency, (10) interpreter test, (11) English aviation test, (12) "business English" test, (13) health professional English test, (14) legal English test, and (15) professional English test. This list of keywords was developed on the basis of the key terms and synonyms used in the initially reviewed books and articles, authors' experiences, and major professional fields of interest as described by Douglas (2013).

We performed a secondary screening of papers that were selected from the initial screening of abstracts using a spreadsheet. We coded each paper as "Include" or "Not Include" in the synthesis based on the following criteria: (a) adult learners, (b) English as the target language, (c) domain-specific EOP, (d) non-self-evaluation, (e) focus on speaking assessment, and (f) empirical or corpus study. Regarding (a) the participants, we included undergraduate, graduate, and international teaching assistants engaging in particular professional domains, as well as adult learners, who were working professionals from specific fields. Concerning (b) the target language, we excluded papers that addressed languages other than English. For (c) domain-specific EOP, papers that included EAP or immigration issues were not selected. In contrast, we included working professionals and university students (i.e., undergraduate, graduate, and international teaching assistants), who studied English for specific professional purposes, if the study was specific to a certain professional domain. In other words, we did not include papers which were EOP in general. For (d), we excluded studies conducted based only on self-assessment data because judgments regarding an adequate level of domain-specific English proficiency do not necessarily match the self assessments (Knoch and Macqueen 2020, 9) and we aimed to analyze scoring rubrics. Regarding (e) and (f), we included both empirical and corpus studies focusing on ESP

speaking assessment as far as working professionals' speaking skills were actually measured. In other words, research that investigated how English is used in the workplace was excluded. Regarding (f), we did not include papers without empirical data even if the topic was related to ESP assessment.

The agreement percentage between the two researchers was 95.76%, whereas disagreements were resolved through a discussion. Among 6,795 studies that were initially retrieved, 34 studies were finally selected using the aforementioned process.

3. RESULTS

3.1. Raters in EOP Speaking Assessment Research

Raters were categorized as (i) domain specialists such as doctors and pilots, (ii) language professionals, (iii) both domain and language specialists, and (d) no raters were included or rater information was not clearly described. Table 1 presents the raters in each professional domain.

As can be seen here, approximately 18 percent of studies included both domain specialists and language professionals, nearly 30 percent of studies only involved domain specialists, and over 30 percent only included language professionals as raters. This means that nearly half of the previous studies included domain specialists in the rating stage. It is also worth noting that some studies had several domain specialist groups such as pilots and air controllers (Kim 2018; Kim and Billington 2016), as well as medical and nursing clinical educators (Manias and McNamara 2016; Pill and McNamara 2016).

When comparing within professional domains, no differences in tendency were evident. Both Health and Aviation, the two major domains, showed similar tendencies in that (a) over half of the studies included domain professionals as raters, (b) less than a quarter of the studies only involved language professionals, and (c) multiple studies included more than one professional group (doctors and nurses, pilots and controllers).

Table 1 Raters in previous studies

Raters	Professional fields				
	Health	Aviation	Education	Others*	Total
Domain specialists	6 (35.29)	2 (33.33)	0 (0.00)	2 (28.57)	10 (29.41)
Language specialists	4 (23.53)	1 (16.67)	2 (50.00)	4 (57.14)	11 (32.35)
Both domain and language specialists	3 (17.65)	1 (16.67)	1 (25.00)	1* (14.29)	6 (17.65)
No raters / Not clearly mentioned	4 (23.53)	2 (33.33)	1 (25.00)	0 (0.00)	7 (20.59)
Total	17 (100.00)	6 (100.00)	4 (100.00)	7 (100.00)	34 (100.00)

Note. Figures in parentheses indicate percentages. "Others" included Customer Service, Interpretation, Business and Management, and Science and Technology. Raters in Friginal (2013) were professionals in service quality who had worked as university TESL instructors before; although there were not two different rating groups, this study was counted in "Both domain & ESL specialists – Others" because the raters were qualified as both domain and ESL specialists.

In sum, ESP speaking research tended to include domain specialists as raters, and some studies even included different profession groups in one professional field.

3.2. Assessment Criteria in EOP Speaking Assessment Research

Not all papers explicated the definition of each dimension, and thus, we evaluated 23 papers that measured speaking skills by using rating scales.

Table 2 shows that fluency (k = 18), intelligibility (k = 21), appropriateness of language (k = 13), and grammar and expression (k = 17) were included in the majority of studies regardless of the professional category. The rating scales in more than 50% of the studies included the appropriateness of language (k = 13). That is, these dimensions can be considered general speaking skills. However, we must note that numerous studies, excluding Friginal (2013), have not presented detailed scales. Therefore, comparing these studies was not simple. For example, the majority of these studies (18 out of 23) included fluency in their rating scales. However, definitions of "fluency" were different among studies. For example, Knoch (2014) emphasized "rate of delivery," whereas Han's (2016) disfluency (un/filled pauses and long silence) may overlap with "intelligibility." Therefore, clearly defining each dimension is a necessary task for future research.

Table 2 Assessment criteria used in previous studies

Dimension		*k* (/23)
Overall communicative effectiveness		10
Linguistic	Fluency	18
	Intelligibility (pronunciation, intonation, stress, rhythm, accent)	21
	Appropriateness of language (e.g., use of suitable professional language)	13
	Grammar and expression (range and accuracy of language)	17
	Comprehensibility	4
	Comprehension	6
Professional	Clinician engagement (professional manner, patient awareness)	5
	Management of interaction (information gathering, information giving)	11
	Professional tasks, information completeness, handling questions	6
Others	L2 perception / Accuracy of information / Knowledgeableness about content / Voice quality / Interlocutor influence	5

Note: One study (Lumley 1998) used different criteria for doctors and language teachers; more specifically, doctors evaluated only based on "overall communicative effectiveness" while language teachers evaluated based on linguistic categories. Other studies adopted the same criteria for language and domain specialists.

The results reveal another trend; that is, more field-specific criteria have been developed and used in ESP speaking assessments in recent studies (e.g., Friginal 2013; Manias and McNamara 2016; Wette and Hawken 2016). That is, recent studies have considered factors beyond the general criteria of the overall communicative skills and effectiveness of test-takers. They also included criteria that reflected more specific context-related characteristics. For example, health professionals interact and communicate with not only colleagues with the same profession but also patients who do not have professional health-related knowledge. Thus, two new criteria, namely clinician engagement and management interaction, were developed to reflect the real language requirements and situations of test-takers (e.g., Manias and McNamara 2016; O'Hagan et al. 2016). By contrast, aviation professionals (e.g., pilots and air-traffic controllers) prominently communicate with people engaged in the same field. Therefore, speaking assessment criteria, such as the one developed by the international civil aviation organization (ICAO), primarily focuses on target language usage among professionals (e.g., Knoch 2014). Other professional domains did not derive field-specific criteria. Despite

the fact that non-verbal communication is important in workplace communication, non-verbal factors were not included in rating scales. Considering their importance, criteria to assess non-verbal skills should be developed.

3.3. Differences Between Domain and Language Professionals

In this section, we examine how rating performances differ between the evaluation groups, domain, and language specialists. We qualitatively analyze the results from the five studies, which had both domain and language professional groups. The raters in Friginal (2013) are excluded here because they belong to both groups, and comparison is impossible.

Three quantitative studies (Elder 1993; Lumley 1998; van Naerssen and Riggenbach 1987) estimated correlation coefficients between domain and language rater groups, and they all showed that the two rater groups had moderate to high correlations for many dimensions, while some dimensions led to lower coefficients. In Elder (1993), where eight math/science teachers were compared with seven language specialists, the two rater groups demonstrated high correlations (r = .85 - .96) for linguistic dimensions (Intelligibility, Fluency, Accuracy, Comprehension, Interaction, and Overall Performance). However, a lower correlation coefficient (r = .73) revealed that domain specialists rated differently from language professionals in the dimension of subject-specific language. Another finding was that domain specialists disagreed within the group on linguistic rating categories (i.e., accuracy, fluency) but not on subject-specific language; meanwhile, the opposite was the case for language professionals. Lumley (1998) examined all correlation coefficients between nine doctors and ten ESL teachers. The language professionals generally had tolerable agreements with the doctors because the average coefficient was around .70. However, individual differences were found even in the same rater group, and some doctor pairs had a coefficient of .44. Van Naerssen and Riggenbach (1987) also showed high correlations between domain and language specialists who were native speakers of English for professional tasks (r = .96) and for general proficiency questions (r = .99). However, domain specialists and nonnative language specialists had only a moderate correlation (r = .54) when grading for general proficiency questions. Thus, they concluded that grading was primarily influenced by whether English was the raters' first language, rather than the raters' professions.

Two other studies (Knoch 2014; Wette and Hawkin 2016) had qualitative features. Wette and Hawkin (2016) examined correlation coefficients in a quasi-experimental study, but the raters were one domain specialist and one language professional. In the pre-test, the correlation coefficients (r) between the two raters were .77 for language criteria and .80 for medical criteria. However, in the post test, the medical criteria did not show significant correlation (r = .51) because the medical educator marked lower scores. Knoch (2014) investigated whether two language professionals and ten pilots agreed on the appropriate proficiency level for operational flying through focused group interviews. While the pilots and language experts generally had similar evaluations, some discrepancies were also found. For example, while the language specialists judged that a certain speaker's language ability was not sufficient, the majority of the pilots regarded that he should be operational. In contrast, another speakers' language skills were regarded as insufficient by most of the pilots although the language specialists were positive. These may have been caused by differences in the degrees to which the rater groups weigh on the technical knowledge of the speaker.

The similarities and differences between domain and language professionals are summarized as follows. First, while the ratings of the two groups were largely consistent, domain specialists tended to record lower scores. Second, among the rating criteria, the domain specialists emphasized technical knowledge. Third, raters' first languages and individual differences in harshness also affected their rating performances.

4. DISCUSSION

We conducted a systematic research synthesis of 34 studies on speaking assessment in the field of ESP, and revealed the following results. Firstly, significant finding is that research on ESP speaking assessments tends to have more than one rater group, that is domain specialists and language specialists. Secondly, although domain specialists and language specialists generally demonstrate high intergroup correlations, the rater group notably differs in terms of field-specific criteria, and the domain specialists emphasized technical knowledge. Last but not least, field-specific criteria have not been developed compared with linguistic scales.

Based on the aforementioned results, we suggest several directions for future research on ESP speaking assessment rating. First, future researchers must promote approaches for developing appropriate rater training programs and to facilitate cooperation between the two rater groups. When language specialists serve as raters, it is crucial to develop rating scales for each professional domain. Second, non-verbal communication should be included in the rating scale.

This study has some limitations. The first one is the use of only three search engines, namely ERIC, PsycINFO, and ScienceDirect. Although we retrieved 6,795 studies, other databases may offer other papers that we would have done well to include. The second limitation is that we limited our synthesis to ESP and did not include other languages, such as Japanese (e.g., Brown, 1995). To capture the complete picture of LSP, future research synthesis would do well include other languages. Last, due to the small number of studies which compared domain and language specialist rater groups, we could not conduct a meta-analysis. Accumulating correlation coefficients and estimating a synthesized coefficient would enhance the results.

Despite the aforementioned limitations, this study revealed rating characteristics and challenges in ESP speaking assessment research. Since little research synthesis has been conducted on the ESP speaking field, this study contributes significantly to the discussion of rater-related issues.

ACKNOWLEDGMENTS: *An earlier version of this paper was presented at the 2018 American Association for Applied Linguistics Conference in Chicago, USA. This research was supported by the University of Tsukuba Basic Research Support Program Type S.*

REFERENCES

Brown, Anne. 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12: 1–16. doi:10.1177/026553229501200101

Cowling, Jeremy. 2007. Needs analysis: Planning a syllabus for a series of intensive workplace courses at a leading Japanese company. *English for Specific Purposes* 26:426–442. doi:10.1016/j.esp.2006.10.003

Douglas, Dan. 2000. *Assessing Language for Specific Purposes.* Cambridge: Cambridge University Press.

Douglas, Dan. 2001. Language for specific purposes assessment criteria: Where do they come from? *Language Testing* 18:171–185. doi:10.1177/026553220101800204

Douglas, Dan. 2013. ESP and assessment. In *The Handbook of English for Specific Purposes*, edited by B. Paltridge and S. Starfield, 367–383. New York: John Wiley & Sons.

Duijm, Klaartje, Rob Schoonen, and Jan Hulstijn. 2018. Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing* 35:501–527. doi:10.1177/0265532217712553

Ekkens, Kristin, and Paula Winke. 2009. Evaluating workplace English language programs. *Language Assessment Quarterly* 6:265–287. doi:10.1080/15434300903063038

Elder, Catherine, and Tim McNamara. 2016. The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing* 33:153–174. doi:10.1177/0265532215607398

Elder, Catherine. 1993. How do subject specialists construe classroom language proficiency? *Language Testing* 10:235–254. doi:10.1177/026553229301000303

Friginal, Eric. 2013. Evaluation of oral performance in outsourced call centres: An exploratory case study. *English for Specific Purposes* 32:25–35. doi:10.1016/j.esp.2012.06.002

Han, Chao. 2016. Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly* 13:186–201. doi:10.1080/15434303.2016.1211132

Hashemi, Mohammad, and Esmat Babaii. 2013. Mixed methods research: Toward new research designs in applied linguistics. *Modern Language Journal* 97:828–852. doi:10.1111/j.1540-4781.2013.12049.x

In'nami, Yo, and Rie Koizumi. 2009. A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26:219–244. doi:10.1177/0265532208101006

In'nami, Yo, and Rie Koizumi. 2016. Task and rater effects in L2 speaking and writing: A synthesis of generalizability. *Language Testing* 33: 341–366. doi:10.1177/0265532215587390

Jacoby, Sally, and Tim McNamara. 1999. Locating competence. *English for Specific Purposes* 18:213–241. doi:10.1016/S0889-4906(97)00053-7

Kim, Hyejeong. 2018. What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing* 35:403–426. doi:10.1177/0265532218758127

Kim, Hyejeong, and Rosey Billington. 2016. Pronunciation and comprehension in English as a lingua franca communication: Effect of L1 influence in international aviation communication. *Applied Linguistics* 39:135–158. doi:10.1093/applin/amv075

Knoch, Ute. 2014. Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes* 33:77–86. doi:10.1016/j.esp.2013.08.002

Knoch, Ute, and Susy Macqueen. 2020. *Assessing English for Professional Purposes*. New York: Routledge.

Lumley, Tom. 1998. Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes* 17:347–367. doi:10.1016/S0889-4906(97)00016-1

Manias, Elizabeth, and Tim McNamara. 2016. Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing* 33:235–249. doi:10.1177/0265532215608411

O'Hagan, Sally, John Pill, and Ying Zhang. 2016. Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective. *Language Testing* 33:195–216. doi:10.1177/0265532215607920

Pill, John, and Tim McNamara. 2016. How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing* 33:217–234. doi:10.1177/0265532215607402

van Naerssen, Margaret, and Heidi Riggenbach. 1987. Evaluating oral skills in an EST program: Native English-speaking scientists respond to Chinese scientists' English. *English for Specific Purposes* 6:219–232. doi:10.1016/0889-4906(87)90005-6

Wette, Rosemary, and Susan Hawken. 2016. Measuring gains in an EMP course and the perspectives of language and medical educators as assessors. *English for Specific Purposes* 42:38–49. doi:10.1016/j.esp.2015.11.002