UDC 371.3::811.111
https://doi.org/10.18485/bells.2017.9.3

Jelena Grubor, PhD[*]
State University of Novi Pazar,
Department of Philology
Novi Pazar, Serbia

ANALYTICAL MEASUREMENT OF L2 SPEAKING PROFICIENCY IN A FINAL EXAM IN ENGLISH

Abstract
The main aim of the study was to determine whether the assumed components of L2 speaking proficiency constitute the construct itself. The subsidiary aim was to check inter-rater reliability, in order to establish whether the raters used similar criteria while testing speaking proficiency. Thirty philological class students were tested individually by two independent raters on five trait categories (pronunciation, grammar, lexis, fluency, content). The results indicate that the ratings are reliable and that the construct of L2 speaking proficiency comprises all of the assumed components. This provides support for the use of a five-trait-category model in testing speaking proficiency in an L2 in practice. However, due to the size of the sample and its nature, the model needs further empirical testing.

Key words: inter-rater reliability, perceived speaking proficiency, trait categories

[*] E-mail: bram.english@yahoo.co.uk

1 Introduction

Speaking, as a productive skill, is exceedingly difficult to test (Bacham and Palmer 1981). Consequently, assessment and/or testing of speaking proficiency is widely regarded as subjective both in SLA literature and practice, because raters normally use a holistic approach by which a student's level of proficiency is assessed *en général* without any clear criteria in mind. The analytical approach, on the other hand, is less rater-biased, at least due to the fact that raters have some focus of assessment. Since the subject matter of this study is analytical measurement of speaking proficiency, we shall briefly present the theoretical components of L2 speaking proficiency, which is generally considered a 'multifaceted' construct (e.g. De Jong et al. 2012; Housen and Kuiken 2009; Norris and Ortega 2009).

The componential structure of L2 speaking proficiency is normally believed to include complexity, accuracy and fluency, and the individual components themselves are deemed to be multidimensional as well (Norris and Ortega 2009). Broadly speaking, *accuracy* refers to the degree to which the produced language deviates from or conforms to certain norms (Hammerly 1991; Pallotti 2009; Skehan 1998), *fluency* to the speed, ease and smoothness with which learners use their linguistic knowledge in an L2 (Lennon 1990), *complexity* to the diversity, i.e. elaboration and variety of the language produced (Ellis 2003). In the context of perceived fluency testing, complexity may pertain to the intrinsic property of other testing dimensions, in the sense that raters use the level of complexity as a criterion for accuracy, fluency, lexis (Grubor 2013).

Two more dimensions have been reported in research relative to testing L2 speaking proficiency: *lexis* and *adequacy*. The former refers to the 'lexical diversity' (Kormos and Denés 2004) or the 'lexical richness' (De Jong et al. 2012) of the produced language. This variable has not been sufficiently investigated in SLA literature (Skehan 2009), despite the fact that knowledge of vocabulary has proved to be a good predictor of speaking proficiency (Beglar and Hunt 1999; De Jong et al. 2012; Zareva et al. 2005). The latter refers to "appropriateness to communicative goals and situations" or "degree to which a learner's performance is more or less successful in achieving the task's goals efficiently" (Pallotti 2009). Besides the communicative adequacy defined in the previous sentence, there is also the functional (informational) adequacy of speaking, which pertains to the success of conveying messages through speaking (De Jong 2012).

This variable has also been insufficiently investigated in SLA research (De Jong et al. 2007; De Jong et al. 2012; Housen, Kuiken and Vedder 2012; Pallotti 2009)[1], although Pallotti (2009), for example, maintains that it should be viewed not only as a separate dimension of proficiency, but also as a way of interpreting CAF features (i.e. complexity, accuracy, fluency).

There is a considerable body of research dealing with testing *utterance proficiency* within the framework of psycholinguistics. These studies aim at determining the psycholinguistic mechanisms and processes underlying L2 acquisition and the manifestation of the acquired knowledge (Grubor 2013). The scope of our study, on the other hand, is within the area of testing *perceived proficiency* since testing speaking proficiency at schools almost invariably takes the form of teachers' perceptions of students' proficiency levels[2]. Kormos and Dénes (2004) point out that studies on perceived fluency are not very numerous, and this is true of all the other components of speaking proficiency.

In a nutshell, the current study has been initiated by previous large-scale research, which investigated the componential structure of perceived speaking proficiency in a paired-testing format (Grubor 2013). Based on these results, we have made some modifications in the present study (eg. divided the concept of accuracy into grammar and pronunciation), on the one hand, and on the other, we replicated the methodology of the original study to a large extent.

2 Methodology

The main aim of the study was to test the construct of perceived L2 speaking proficiency and determine its components. The selection of independent variables hypothesised to incorporate the stated construct was based on research findings. Generally put, the research has shown that listeners' perception of fluency may be influenced by pronunciation, grammar and vocabulary (e.g. Kormos and Dénes 2004; Rossiter 2009). In addition,

[1] These studies have dealt with the adequacy variable and are, therefore, an exception to the previously stated statement.

[2] The terms perceived and utterance proficiency were developed from perceived and utterance fluency (Segalowitz 2010, cited in De Jong et al. 2013) to emphasise the subjective nature of testing productive skills on the part of teachers in real classrooms. The same term has been used in Grubor (2013).

adequacy was also reported as part of speaking performance (De Jong et al. 2012).

Consequently, we hypothesised that the following trait categories constitute the construct itself: pronunciation (Pron.), grammar (Gram.), fluency (Flu.), lexis (Lex.), and content (Cont.). As we concluded in the initial study (Grubor 2013) and in line with some previous research (eg Bonk and Ockey 2003), we divided the accuracy variable into grammar and pronunciation. Accuracy and fluency represent the variables typically measured in research so far. The former pertained to error-free language (Lennon 1990) relative to grammar and pronunciation, the latter to 'smoothness and ease of oral linguistic delivery' (De Jong et al. 2013) or 'speed and spontaneity of speech' (Grubor 2013). Lexis, in our study, pertained to the range of vocabulary the participants used, i.e. 'lexical diversity' (Kormos and Denés 2004) or 'lexical richness' (De Jong et al. 2012). Finally, no language is produced independently of the communicative context in which any conversation takes place (Grubor, 2013), therefore, the content of speakers' communicative messages becomes particularly significant. The adequacy variable pertained to conveying ideas relevant to the conversational topic (eg whether they responded to the questions they were asked, or conversely, whether they used, for example, circular arguments or lacked any argumentation at all).

2.1 Sample and testing format

The sample recruited for the study included thirty secondary school students, who sat the English school-leaving exam at the Philological Grammar School in Kragujevac, course: Modern Languages. The participants were at the *upper-intermediate to advanced* level, all female (f=30), aged 18 and 19. They had had English for four years, five classes per week, which is approximately 180 classes per year. The sample involved students from two consecutive school years for two reasons. Firstly, the maximum number of students in a philological class is 25, and frequently there are fewer students than this. Secondly, students who have obtained an A in the written test, and who have also had an A in English in the first, second, third and fourth year are exempt from the oral part.

As regards the English school-leaving exam, it consists of the written and oral part. The oral part that we are interested in includes two broad parts. The first part is to do with unfamiliar texts, which cover topical issues students are acquainted with (eg. genetic engineering, eating habits, beauty, moral issues etc). Students are required to answer questions related to text analysis (eg. derivatives, synonyms, antonyms, grammatical structures etc), and they also have an argumentative conversational topic (eg. *Cloning is acceptable*, *Vegetarianism is a philosophy rather than a practical exercise*, *Beauty is only skin-deep*, *Capital punishment should not exist*)[3]. The second part is to do with the reading assignments that students have had (eg. *To Kill a Mockingbird, Animal Farm, Sense and Sensibility*, etc.) and it included argumentative literary topics (eg. *Atticus as a role model father*, *Not the corrupt doctrine but individuals in power*, *Willoughby and Marianne as a (mis)match*, etc.)[4].

The participants' speaking proficiency level was assessed by two independent raters on a pre-defined six-point scale (range *0*: not at all able – *5*: most able), which included the five mentioned trait categories (Pron., Gram., Lex., Flu., Cont.). Both raters were female, had approximately ten years of teaching experience and no previous testing-specific training with regard to testing speaking proficiency. One of them was the participants' subject teacher, the other was a school colleague. The role of the raters was not equally balanced on purpose. First of all, teachers normally do what the first rater did (ask students questions and sub-questions, listen and assess), whereas the second rater played the role of a *supervisor* (i.e. was not involved in task-setting, but instead was able to focus on the content of the produced language). The raters used the Common European Framework of Reference for Languages as a reference point for their assessment as regards the participants' level of speaking proficiency.

2.2 Procedures

For the purpose of analysing the gathered data, we used the statistical programme PASW Statistics[5]. We performed some basic statistical procedures, such as descriptive statistics (means, standard deviations, etc.). The main aim of this small-scale, introductory study was to determine whether the assumed components of L2 speaking proficiency constitute the construct itself. With that in mind, we employed the exploratory factor

[3] This part is referred to as "Conversation" in our study.

[4] This part is referred to as "Literature" in our study.

[5] Upgraded version of the SPSS (18.0) statistical programme.

analysis to test the structural form of the perceived speaking proficiency construct (Principal Component Analysis). With a view to bringing such a measurement format into real classrooms, we wanted to determine whether the raters used similar criteria while testing speaking proficiency. In other words, we looked into the reliability of ratings assigned to each assumed trait category and tested the inter-rater reliability using two general approaches (Pearson's Product Moment correlation and Cronbach's alpha).

3 Results

As regards *conversation*, we may notice that the raters assigned quite high values to each trait category, based on the mean ratings (cf. Table 1).

Table 1. Descriptive statistics: Conversation

Construct	**Rater 1**				**Rater 2**			
	M	Min	Max	SD	M	Min	Max	SD
Pron.	4.62	3	5	.696	4.72	3	5	.663
Gram.	4.46	2	5	.853	4.28	2	5	.936
Flu.	4.74	3.5	5	.481	4.56	2	5	.833
Lex.	4.54	3	5	.676	4.72	3	5	.678
Cont.	4.88	4	5	.332	4.68	3	5	.690

Similar results may be reported for *literary topics* (cf. Table 2). The results of the descriptive statistics analysis, therefore, indicate that the majority of the participants are, on the whole, high-achievers.

Table 2. Descriptive statistics: Literature

Construct	Rater 1				Rater 2			
	M	Min	Max	SD	M	Min	Max	SD
Pron.	4.68	3	5	.610	4.68	2	5	.734
Gram.	4.42	2.5	5	.838	4.42	2	5	.799
Flu.	4.70	3	5	.612	4.68	3	5	.557
Lex.	4.62	3	5	.545	4.76	3	5	.502
Cont.	4.92	4	5	.277	4.76	3	5	.502

With the aim of determining the reliability of the ratings, or statistically speaking, the inter-rater reliability, we performed two general approaches. The first of them was Pearson's Product Moment correlation, which showed consistent ratings. As we can see from the table presenting the results of *conversation*, the correlations (reliability estimate *r*) were strong, positive and of the first level of significance (cf. Table 3). Drawing on researchers in the field of applied linguistics (Brown 2004; Larson-Hall 2010), we subsequently adjusted the inter-rater correlations using the Spearman-Brown prophesy formula, because Pearson's coefficient cannot account for the number of raters in the study. This is important, statistically speaking, because it may influence the strength of correlations, as was the case with only two raters in our study (cf. reliability estimate *r* adjusted). In other words, when the correlations were adjusted according to the number of raters, they were even stronger, indicating very reliable ratings.

Table 3. Reliability estimates (ratings correlation): Conversation

Construct	Reliability estimate (r)		Reliability estimate (r adjusted)
	p	r	rho
Pronunciation	p=.000	r=.888	.941
Grammar	p=.000	r=.915	.956
Fluency	p=.000	r=.846	.917
Lexis	p=.000	r=.889	.941
Content	p=.000	r=.735	.847

With respect to *literary topics*, the correlations, for the most part, were strong, positive and of the first level of significance (cf. Table 4). However, the fluency trait category had the significance of the second-level and moderate-to-strong correlation. Although this correlation was improved when adjusted (i.e. was shown to be strong), still this correlation was the weakest in comparison to other values.

Table 4. Reliability estimates (ratings correlation): Literature

Construct	Reliability estimate (r)		Reliability estimate (r adjusted)
	p	r	rho
Pronunciation	p=.000	r=.808	.894
Grammar	p=.000	r=.799	.888
Fluency	p=.011	r=.501	.668
Lexis	p=.000	r=.794	.885
Content	p=.000	r=.755	.860

The second approach to testing inter-rater reliability is computing Cronbach's alphas (α) for each assessed construct. Howell (2002) maintains that the best way to test inter-rater reliability for cases of raters assessing people is to examine the intraclass correlation. Larson-Hall (2010) explains that looking into the intraclass correlation will take into account the correlation between the raters, but it also enables the researcher to determine whether the actual assigned scores differ. Thus, we employed the scale reliability (method: Two-way random), which again indicated that the ratings were concordant, judging from the values of Cronbach's coefficients (cf. Table 5).

Table 5. Reliability estimates (Cronbach's alphas): Conversation & Literature

Construct	**Reliability estimate (α)**		**Reliability estimate (α adjusted)**	
	Con.	**Lit.**	**Con.**	**Lit.**
Pronunciation	.940	.885	.941	. 894
Grammar	.953	.888	.955	.888
Fluency	.846	.666	.916	.668
Lexis	.941	.883	.941	.885
Content	.729	.779	.847	.860

Con: Conversation, **Lit**: Literature

After establishing that the ratings were concordant and the raters were in agreement, we went on to test the hypothesis that the five-trait categories constitute the construct of speaking proficiency by conducting the exploratory factor analysis (EFA). We included separate variables for each rater's composite score to check whether the assessed categories would cluster together, i.e. load on the same factor. The extraction method was the Principal Component Analysis. Only one factor was extracted (cf. Table 6), explaining 82.5% of the total variance. This means that Pronunciation, Grammar, Fluency, Lexis and Content constitute one theoretical construct, i.e. perceived speaking proficiency, and that together they explain 82.5% of the said construct.

Table 6. Exploratory factor analysis: Factor loadings > .60

Component Matrix[a]	
Mean	Component
	1
Pronunciation T1	.948
Pronunciation T2	.902
Grammar T1	.937
Grammar T2	.895
Fluency T1	.791
Fluency T2	.929
Lexis T1	.933
Lexis T2	.963
Content T1	.817
Content T2	.951
Extraction Method: Principal Component Analysis. a. 1 components extracted.	

4 Discussion

In this section, we will briefly comment on the results, bearing in mind the stated aims. The first aim was to check whether the raters, despite their lack of training in testing speaking proficiency, were consistent in their subjective ratings of the assumed trait categories. This aim was important because of the validity and reliability of the results in the first place, and also because of potential implications for theory and practice. As Derwing et al. (2004) stated regarding fluency, an examination of the reliability of raters' judgements is essential to determine the construct validity of perceived fluency, which in our case applies to other dimensions of speaking proficiency as well.

The results show that the raters were in agreement, or more precisely, that their ratings were concordant, which is supported by the very high correlations and values of Cronbach's alpha. In other words, despite the lack of previous official training in testing speaking proficiency, these raters/teachers similarly assessed the given categories. Other studies reported similar results, showing no differences with regard to the raters' training (Caban 2003; Derwing and Munro 1997; Munro and Derwing 1999). This may imply that, in general, raters rely more on their experience and intuition than on a set rating scale (Teng 2007), or that the raters measure speaking proficiency with similar criteria in mind.

The second aim was to determine the components of the perceived speaking proficiency construct. Speaking is a very complex construct, and consequently it is difficult to devise an appropriate measurement of oral proficiency, given a wide range of aspects that should or could be assessed/tested (Grubor 2013).

Our results suggest that there is a unique but multifaceted construct of speaking proficiency, which comprises pronunciation, grammar, fluency, lexis and content. In the initial large-scale study, which investigated perceived speaking proficiency in a paired-testing format, it was concluded that there is a unique construct of language use, which further divides into *purely linguistic* and *sociolinguistic* features (Grubor 2013). In the current study (one-to-one testing format), the unique construct includes linguistic features (Pron., Gram., Lex., Flu.) and the adequacy of delivered communicative messages (Cont.).

Finally, although all the correlations were positive and strong, we need to point out that the correlations on the subscale of fluency in literary topics were lower in comparison to all the other values. This raises the question whether one of the raters applied the "speed-and-spontaneity-of-speech" criterion quite rigidly or overlooked the fact that the test-takers might have been deciding on their opinion on the spot, or else may have been unable to remember some instances to support their views. In a word, the inherent nature of 'free topics' included in the conversational part and literary topics is quite different. In the event of discussing certain protagonists and/or events in their literary pieces, the participants might not have had 'ready-made' answers or formed opinions, but instead needed some extra time to shape their thoughts, which might have been misinterpreted as lack of fluency. This finding implies that further research should be conducted in this direction to determine whether there should be different criteria involved as regards giving opinions on literary topics.

To conclude, the results suggest that the scale can be used in the classroom, as was the case with the sample of this study. However, we need to add some words of caution at this point with respect to the limitations of the study. The first one is concerned with the nature of the sample: all the participants were female and they were largely high-achievers. The second one is concerned with the sample size. As we have already pointed out, the number of students in philological classes is quite small, thus it was impossible to include a larger sample. This obstacle can be overcome by replicating the study with tertiary students, or by conducting a study over a few years, thereby obtaining a larger sample. Finally, due to the said limitations, we must emphasise that these results refer only to the sample of this study and not to L2 learners in general.

5 Conclusion

Assessing and/or testing productive skills, such as speaking and writing, reflects a certain amount of subjectivity, even if there are clear descriptors (such as in CEFR). The current study has thus placed emphasis on *perceived proficiency*, because that seems to be the reality of everyday classrooms. Accordingly, we have opted for a more objective measurement system, an analytical, not a holistic one.

The results suggest that there are five dimensions of speaking proficiency: pronunciation, grammar, fluency, lexis and content, or else that these five dimensions play an important role in testing speaking proficiency. The statistical analyses performed have indicated that teachers can use such an instrument in the classroom. The main advantage of the five-trait-category model is that it can readily be employed in schools. In addition, when the speaking skill is 'divided' into certain dimensions such as these, teachers may easily give their students instant feedback on the areas to work on in the future in order to improve their speaking skills. In addition, teachers may devise certain activities which will help their learners to enhance the specific aspects of the speaking skill that they have problems with. In a word, although testing speaking proficiency is indeed subjective, using such a clear-cut model is much more objective than a holistic approach where the teacher's overall impression of the learner's level of speaking proficiency plays the one and only role in giving a grade to a student and enables vague or no feedback at all.

References

Bacham, L. and A. Palmer (1981). The construct validation of the FSI oral interview. *Language Learning*, 31(1), 67–86.

Beglar, D. and A. Hunt (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16(2), 131–162.

Bonk, W. J. and G. J. Ockey (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89–110.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies*, 22(2), 91–139.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, 21(2), 1–44.

De Jong, N. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(5), 5–34.

De Jong, N. H. et al. (2007). The effects of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele et al. (eds.), *Complexity, Accuracy and Fluency in Second Language Use, Learning and Teaching*, Brussels: Contact forum, 53–63.

De Jong, N. H. et al. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34.

De Jong, N. H. et al. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(5), 893–916.

Derwing, T. M. and M. J. Munro (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16.

Derwing, T. M. et al. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655–680.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford: Oxford University Press.

Grubor, J. (2013). The construct of perceived L2 speaking proficiency in a paired testing format. *Poznań Studies in Contemporary Linguistics*, 49(2), 185–203.

Hammerly, H. (1991). *Fluency and Accuracy: Toward Balance in Language Teaching and Learning*. Clevedon, UK: Multilingual Matters.

Housen, A. and F. Kuiken (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.

Housen, A, F. Kuiken and I. Vedder (2012). *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins Publishing Company.

Howell, D. C. (2002). *Statistical Methods for Psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Kormos, J. and M. Dénes (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164.

Larson-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS*. Second language acquisition research series. New York: Routledge.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417.

Munro, M. J. and T. M. Derwing (1999). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, 49(suppl. 1), 285–310.

Norris, J. M. and L. Ortega (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, 19(2), 169–192.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.

Rossiter, M. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, 65(3), 395–412.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. New York: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.

Teng, H-C. (2007). A study of task type for L2 speaking assessment. *ISLS Readings in Language Studies*, 433–446.

Zareva, A, P. Schwanenflugel and Y. Nikolova (2005). Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition*, 27(4), 567–595.