Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 2, 2020, pp 305-312
https://doi.org/10.18510/hssr.2020.8234

APPLYING ITEM RESPONSES THEORY FOR MEASURING STUDENT'S ABILITY IN ACADEMIC SPEAKING

Memet Sudaryanto[1*], Kundharu Saddhono[2], Lina[3]
[1,2]Sebelas Maret University, Indonesia, [3]Institut Ilmu al-Quran An Nur Yogyakarta, Indonesia.
Email: [1*]memetsudaryanto@staff.uns.ac.id, [2]kundharu.uns@gmail.com, [3]madahafiyya@gmail.com

Article History: Received on 07th January 2020, Revised on 27th February 2020, Published on 28th March 2020

Abstract

Purpose of the study: Every human being is required to be skilled at communicating, skilled at expressing thoughts, ideas, and feelings. This research aims to explain student's ability in academic speaking based on the framework of item responses theory.

Methodology: This research uses a mix method research approach; qualitative and quantitative approaches are used together to answer the formulated problems. A qualitative approach is used for digging up information about the needs to develop the speaking test. The try out subjects was 25 university student taking Bahasa Indonesia subject; while 125 students were used as the measurement subject. The data were students' responses to the speaking test which was scored by two people (rater). The reliability of the instrument was estimated by the Generalizability theory. Rasch model analysis was used to estimate the item parameter; while the Maximum Likelihood was used to estimate the students' ability.

Main Findings: The value of $\sigma^2(o)$ s influenced by the similarity to the average mean score observed in academic speaking. The value of $\sigma^2(p)$ and $\sigma^2(o)$ suggests that the distribution of variability in person and item is the same and high. A sufficiently large value of $\sigma^2(pi)$ implies the fact that the value involves all residual sources for a variance. Balance alternatives, weigh consequently and decide rationally.

Applications of this study: Based on the design proposed for the Indonesian Language Proficiency assessment, other generalizability design (G-Design) used is a cross design because each student (p) becomes the object of observation of two observers (r) who both assess four aspects of observation/indicator. It is used to determine the reliability (i.e., reproducibility) of measurements under specific conditions in academic speaking.

Novelty/Originality of this study: Some rater qualifications that must be met include the process of gathering and using the appropriateness of background information before assessing. Measuring the ability of students in academic speaking by applying G theory and conducting an IRT analysis approach needs.

Keywords: *Item Responses Theory, Language Test, Academic Speaking, Generalizability, Rasch.*

INTRODUCTION

Speaking skill is a form of productive skills which susceptible to bias and error measurement. Most rater on speaking tests not only measure the speaker's speaking ability, but also focus on other factors such as how the students' dress, attractiveness, or personal factors that hard to explain. These factors cause the bias of the measurement result. High subjectivity speech style will influence the assessment of one's speaking abilities (Nash, Crimmins, & Oprescu, 2016).

Academic speaking is a language skill that is mandatory mastered by advance language speakers. Speakers' classifications are divided based on not only their ability to use complex speech styles but also the need to communicate in a work and study (Andrade, 2006). Academic speaking is the highest level of communication because it requires an ability to speak that is not biased and acceptable speech topic for people who are listening to them for the first time. Speaking skill is a productive skill consists of several components, namely: (a) the use of spoken language which functions as a medium of speech through the vocabulary of language structure, pronunciation, and intonation, and various languages; (b) mastery of the content of the conversation that depends on what is the topic of the conversation; (c) mastery of techniques and speaking performance that are adapted to the situation and type of conversation, such as conversing, giving speeches, telling stories, and so on (Mattarima & Rahim Hamdan, 2011). Mastery of technique and appearance is essential in the types of formal speech, such as speeches, lectures, and discussions.

Complex speaking skills can be accepted as a form of productive skills that master a variety of speech styles, topics, and mastery of a massive and comprehensive audience. This ability will be seen from the speech delivered until the presentation received by all audiences. Speaking skills can be demonstrated by mastering the topics presented, being responsive to responses, developing ideas, ice-breaking references, and improvising. This causes the results of the assessment is not comprehensive because basically the quality of communication and the suitability of the language context is something that should not be ignored (Rahmawati, Suwandi, Saddhono, & Setiawan, 2019).

In contrast to the content and development of speech material, an advance speaker should master articulation, intonation, pauses, choice of words (diction), and voice volume. A variety of vowel games are used to show forms of speech variations that show the quality of the speaker. One of the keys to quality speech is the good articulation and can be

heard with every word. The selected words certainly have different meanings even though the pronunciation is very close. Therefore, articulation is essential through many other linguistic factors that are able to support message delivery.

Every human being is required to be skilled at communicating, skilled at expressing thoughts, ideas, and feelings (Pyrik, 2015). Speaking skill plays an essential role in human life. Everyday life is faced with various activities require a speaking skill. In the purpose of conversation to reach its objectives, the speaker must have the ability and skills to convey information to others. It implies that the speaker must understand very well how to speak harmoniously and effectively so that other people (listeners) can receive the information conveyed by the speaker effectively. That is why it is so important to have language skills, especially speaking skills (Laufer & Aviad-Levitzky, 2017).

Error in receiving messages in the communication process is indicated by several factors, such as, (1) material that is not relevant to everyday life, (2) delivery too fast, (3) unclear articulation with double-meaning words, and (4) technical vocals such as sleeping on the tongue. Speaking as one aspect of productive language skills, the ability to change the form of thoughts or feelings into meaningful sounds of language (Paul& Smith, 1993). Speaking is the ability to say articulation sounds or words to express state and convey thoughts, ideas, and feelings (Townsend, Kim, & Mesquita, 2014).

Based on the explanation above, it is clear that measuring speaking ability is a complicated process. It needs a good instrument that can minimize measurement bias. This research developed a reliable instrument to measure students' speaking ability and applied it in the measurement process.

LITERATURE REVIEW

Language is the only communication medium that humans use to give and receive responses. As a media, language is divided into two namely written language and spoken language. As a form of skill, language is divided into productive and receptive skills. Brooks states that listening and reading are receptive. Listening and reading skills are in the category of receptive skills while writing and speaking skills are productive skills (Brooks-Lewis, 2009). This classification underlies the function of language as a medium of communication.

In addition to the communication function, language has a central function as the cultural identity of a society. Indonesia recognizes Indonesian as the national language, the language of unity, and the language of communication. These three functions underlie the importance of standardized and applicable Indonesian nationally and internationally. Moreover, currently, Indonesian is being pioneered as an international language in Asia.

The main points of language identification are divided into four skills namely, listening skills, speaking skills, reading skills, and writing skills. Listening skills are closely related to speaking skills, as are reading skills and integrated writing skills. Agreeing with Cunnings worth's statement asserts that the basic knowledge of language includes grammar, lexemes, and four language skills such as listening, speaking, reading and writing, all of which must go hand in hand (Cunningsworth & Horner, 1985).

Horwitz adds that the core of using language must be authentic because language is scientific that is practical and needs to be sharpened (Horwitz, 2001). With the more frequent and steady use of good grammar, the language used does not need to laboriously explore the use of standard Indonesian spelling. The intensity of the use of language capable of steady helps the language used to learn independently while using good and correct spelling.

Academic activity in the language is shown by discussion, problem-solving, and the ability to play roles (Oradee, 2013). This shows that each activity can refer to different interactions that require different technical complexities (Zyoud, 2016). Speaking skills need to be supported by routine daily listening activities by listening to good language rules and not biased meaning in each speech. Someone's speaking ability will be predictable from his listening and reading skills both in quantity (routine) and the quality of the simulation.

Improving listening skills also means helping improve the quality of one's speech. A good listening concept is listening comprehension because it requires skills to store, use, and master a number of facts that they refer to (Winarni, Slamet, & Saddhono, 2018). Saddhono states that language as an agency symbol for humans to communicate and interact in various social groups so as to produce various interpretations of messages socio-cultural meaning (Saddhono, 2015). The academic acumen of lecturers, students, scientists, and researchers in interpreting this meaning has the potential to produce academic products that can spur the growth of science and technology. As a means of academic scientific communication in the form of scientific texts, Beaugrande state that discourse must meet seven standards of textuality, namely cohesion, coherence, intentionality, acceptability, informativity, situational, and intertextuality (de Beaugrande, 2006).

Moore designs language skills to maximize one's learning to use language in stages (1) describe in general terms; (2) determine specific objectives (objectives) with specific strategies both directly and indirectly, and (3) summarize and integrate the content of the discourse with the needs in the field into daily activities (Moore, 2005). Academic understanding is the interrelation between cohesion and coherence between elements as elements related to syntactic elements while cohesion and coherence are more related to semantic meaning.

A speaker must choose a variety of languages that are appropriate to the social and cultural language (environment). If the linguistic structure is wrong and does not fit into the variety and ecology of language (Sudaryanto, Mardapi, & Hadi, 2019), it will lead to obstruction of communication, misinterpretation, and misplacement of the desired meaning. Likewise, the choice of words used must be per the ecology of the language, the topic of conversation, and the level of the recipient of the conversation. Thus, speaking skill is a complex skill.

Miller suggests that in general, the form of tests that can be used in measuring speech are subjective tests that contain instructions for carrying out speaking activities (Miller, 2013). Some tests that can be used include, (a) A test of speaking skills based on pictures, carried out by giving questions in connection with a series of pictures or telling a series of pictures, (b) An interview test, used to measure adequate language skills. (c) Storytelling, done by expressing something (his experience or a particular topic), (d) Discussion, by asking to discuss a particular topic, and (e) Structured speech, which includes retelling, reading quotations, changing sentences and constructing sentences (Brooks-Lewis, 2009).

METHODOLOGY

This research applied a mixed-method research approach, namely qualitative and quantitative, which were used together to answer the problem formulated. A qualitative study was used in the process of digging up information about the needs for speaking tests for advance speakers containing topics that are relevant to their daily needs, through document and content analysis. Document analysis was applied to view the current test/non-test instrument for measuring speaking skills. Furthermore, content analysis was carried out to investigate the current speaking skills assessment document. The results of the document and content analysis were used as the basis of developing the instrument of this research which is in the form of test and observation rubric. The developed instrument then was used to measure students' academic speaking ability. 25 students were taking Bahasa Indonesia class which were asked as the subjects of the instrument try out, while 125 students were involved in the measurement process. Two raters were asked to score the students' speaking ability x. The scoring of the instrument try out results then was analysed to estimate the reliability; while the score of the measurement than was analysed to estimate students' speaking ability. The reliability analysis of the instruments applied Generalizability theory and the estimation of students' speaking ability applied Maximum Likelihood Estimation which is in the framework of item response theory.

RESULT/FINDINGS

A speaker must choose a variety of languages that are appropriate to the social and cultural language (environment). If the linguistic structure is wrong and does not fit into the variety and ecology of language, it will lead to obstruction of communication, misinterpretation, and misplacement of the desired meaning (Winarni et al., 2018). Likewise, the choice of words used must be following the ecology of the language, the topic of conversation, and the level of the recipient of the conversation. Thus, speaking skill is a complex skill.

Table 1: ANOVA for p × i × o Design Using Synthetic

with np =	25	ni = 4		no = 2		σ^2 (α)
P	24	12	15615,83	63,83	2,66	24,5385
I	3	75	15600,51	48,51	16,17	-0,4737
R	2	100	15552,86	0,86	0,43	2,7291
Pi	72	3	15720,00	55,66	0,77	-8,9883
Pr	48	4	11754,38	-3862,32	-80,46	-27,0507
Ir	6	25	15606,56	5,19	0,87	-1,0749
Pir	144	1	15858,00	3994,26	27,74	27,7380
Mean (μ)			15552,00			
Total	299			306.00		

Based on the results of the scores observed in table 1 above, the observed score ranges from 0 to 9. The value of $\sigma(p)$ s the highest compared to the estimated value of variance component means score items and occasions and is influenced by considerable variability in the average score of the person observed (Brennan, 2010).

Based on the results of the analysis of the ability and analysis of needs in communicating in daily life, four things must be mastered by speakers in the ability to speak (a) Presenting knowledge and experiences, (b) Participating in discussions, (c) Presenting/interpreting literary texts, and (d) Giving talks on various issues. Learning activities that can be developed based on the above indicators are (a) Identifying statements in problem-solving in a formal conversation, (b) Identifying the reasons stated or not stated in an opinion, (c) Seeing the similarities and differences between two or more opinions, (d) Finding, underlining, and ignoring (if appropriate) seems irrelevant in an opinion, (e) Describing the

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 2, 2020, pp 305-312
https://doi.org/10.18510/hssr.2020.8234

logical or structure of opinion, and (f) Concluding an opinion. Several indicators of speaking skills are constructed in items that are tested for reliability based on rater ratings.

Table 2: ANOVA for p × i × o Design Using Synthetic Data Set No. 3*with np = 25, ni=4, no=2

Effect (α)	df(α)	f(α)	T(α)	SS(α)	MS(α)	σ^2 (α)
p	24	8	9920,13	50,00	2,0833	-0,0678
i	3	50	9870,98	0,85	0,2850	0,0097
o	1	100	9875,57	5,44	5,4450	0,0681
pi	72	2	9958,50	37,52	0,5211	-0,2639
po	24	4	10001,25	75,68	3,1533	0,5261
io	3	25	9878,28	1,86	0,6183	-0,0431
pio	72	1	10117,00	75,52	1,0489	1,0489
Mean (μ)		80	9870,13			
Total	199			246,87		

The value of $\sigma(o)$ s influenced by the similarity to the average mean score observed. The value of $\sigma(p)$ and $\sigma(o)$ suggests that the distribution of variability in person and item is the same and high. A sufficiently large value of $\sigma(pi)$ implies the fact that the value involves all residual sources for variance (Kuzar, 2003). The value of $\sigma(pi)$ is greater than $\sigma(po)$ and $\sigma(io)$ which suggest that item-person interaction is important and needs to be considered when designing a procedure.

$$rxx' = \frac{\sigma^2 true}{\sigma^2 obs}$$

While the components of pure score variance and observed score variance can be described as follows

$$\frac{\sigma^2 true}{\sigma^2 obs} = \frac{\sigma^2 p}{\sigma^2 p + \frac{\sigma^2 i}{ni} + \frac{\sigma^2 r}{nr} + \frac{\sigma^2 pi}{ni} + \frac{\sigma^2 pr}{nr} + \frac{\sigma^2 ir}{ninr} + \frac{\sigma^2 pir}{ninr}}$$

By entering the value of the variance field score above in the equation below will produce the reliability coefficient as follows

$$rxx' = \frac{24,5385}{24,5385 + \frac{0}{4} + \frac{2,7291}{3} + \frac{0}{4} + \frac{0}{3} + \frac{0}{12} + \frac{27,7380}{12}}$$

$$rxx' = \frac{24,5385}{24,5385 + 0,9097 + 2,3115}$$

$$rxx = \frac{24,5385}{27,7597}$$

rxx'= 0,883961

Some rater qualifications that must be met include, (a) the process of gathering and using the appropriateness of background information before assessing, (b) identifying the consequences of conclusions that can be made and weighing the consequences before conclusions are made, (c) identifying alternative actions and values its value, and (d) balance alternatives, weigh consequently, and decide rationally. Based on the design proposed for the Indonesian Language Proficiency assessment, the generalizability design (G-Design) used is a cross design, because each student (p) becomes the object of observation of two observers (r) who both assess four aspects of observation/indicator (i). Thus, the generalizability design used is p x i x r.

Construct reliability is estimated by the loading factor of each indicator that makes up the instrument (λ) and the unique error index of each indicator (δ), which is formulated as follows

$$CR = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + (\sum \epsilon_i)}$$

(1) (Geldhof, Preacher, & Zyphur, 2014).

$$\omega = \frac{(\sum \lambda_j)^2}{(\sum \lambda_j)^2 + \sum \sigma^2_{\varepsilon_j}}$$

(2) (Kamata, A., Turhan, A., & Darandari, 2003).

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 2, 2020, pp 305-312
https://doi.org/10.18510/hssr.2020.8234

$$\Omega_w = \frac{\sum_{i=1}^{p} \frac{l_i^2}{(1-l_i^2)}}{1+\sum_{i=1}^{p} \frac{l_i^2}{(1-l_i^2)}}$$

(3) (Penev & Raykov, 2006).

The three formulas above estimate the reliability of the instrument, namely the estimation with the construct of reliability, the formulation with omega reliability (ω) uses a factor load (λ) sedan meanwhile the maximum reliability (Ω) uses a factor load symbolized by ℓ. This construct reliability can be estimated after the instrument developer proves the construct validity by confirmatory factor analysis until obtaining a suitable model (fit model) (Retnawati, 2016).

Examinee Ability

Specific strategies for conversations or speaking activities such as debates or talk shows also need to be discussed in depth so that the speaker will not have difficulty in determining the technique of practicing speaking skills. Practice material contained in textbooks to help learners respond to conversations that were unpredictable beforehand. This needs to be studied, especially for academic languages that are not well mastered by speakers. The speaker or learner will know things outside of his knowledge in a real context. Similar mistakes made are often rushed to speak. Another mistake that often happens is to talk at the same time (or even some time before) other people finish talking.

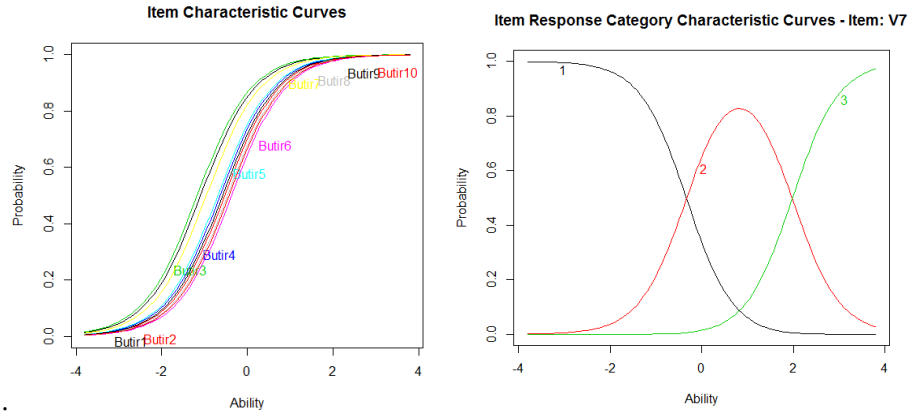

Figure 1: Item Characteristic Curves of the Speaking Test

There is a need to practice constantly in order to get good results in time management in speaking. In time management communication skills can be interpreted as "when the most appropriate time to stop talking and provide opportunities to the other person". The assessment of speaking skills is carried out differently at each level. Nurgiyantoro argues that the chosen form of speaking ability should allow students to not only speak their language skills, but also express their ideas, thoughts, or feelings (Nurgiyantoro, 2009). Thus the test is functional in addition to also revealing the ability of students to speak in the language concerned is approaching its normal use. Some tasks that can be used include discussion based on pictures, interviews, storytelling, discussion, and giving speeches (Sudaryanto, Saddhono, et al., 2019). In each form of speaking skills, each assessment model can be developed. For example, there are several ways of evaluating speech assignments Jakobvits and Gardon put forward speech and story evaluation techniques. This assessment model can also be used in the assessment of discussions.

Item Responses Theory Analysis

The score for estimating skills on each skill needs to be standardized in standard rules. Speaking is a productive language skill that is first performed by a child. This skill needs to be practiced by listening to and watching a variety of contexts from the softest to the loudest and clearest. Practicing sensitivity in listening skills starts with audiovisual media from the easiest to the most difficult.

Table 3: KMO and Bartlett's Speaking Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.808
Bartlett's Test of Sphericity	Approx. Chi-Square	2210.111
	df	306
	Sig.	.000

All student responses to speaking items were analyzed using SPSS to obtain factors and measure Kaiser-Meyer-Olkin (KMO), which found 0.808 results. Based on an analysis of the adequacy of the sample shows the chi-square value in the Bartlet test of 2210.9 with a degree of freedom 306 and a p-value of less than 0.01. In addition to the adequacy of the sample in the assumption test with the item response theory need to be proven by three tests, namely unidimensional, local independence, and parameter invariance (Hambleton & Swaminathan, 1985).

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 2, 2020, pp 305-312
https://doi.org/10.18510/hssr.2020.8234

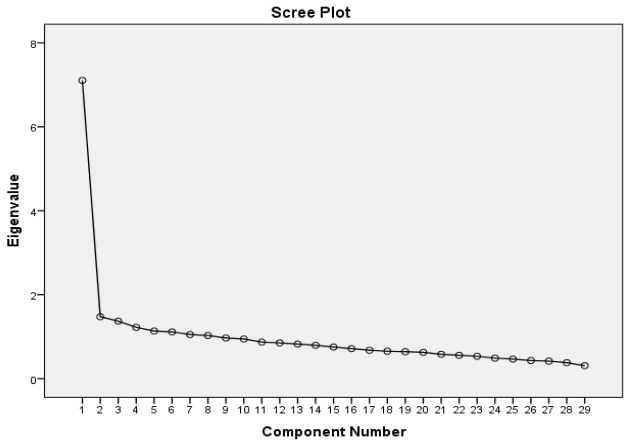

Figure 2: Scree Plot of Eigenvalues of Speaking Test

Based on the screen plot above shows the items used to measure speaking skills consist of one dimension tested (Fraley, Waller, & Brennan, 2000). While the parameter invariance is seen from the functioning of even and odd items through different levels and difficulty index.

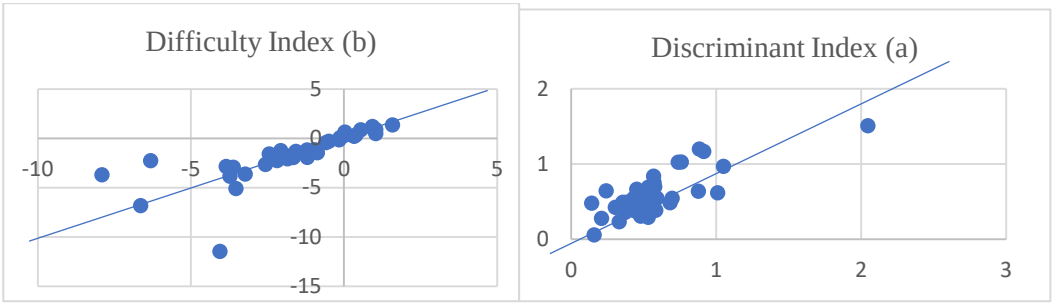

Figure 3: Parameter (*b*, *a*) Invariance Analysis

In Figure 3, each point can be observed relatively close to the line on slope 1. This shows that there is no variation in the parameters of the estimated group of test-takers. Similarly, the different power can be observed in the two diagrams above (Retnawati, 2008).

DISCUSSION/ANALYSIS

Rater who measures self-control and delivery of ideas evaluates speaking skills. Assessment in speaking skills is to control self-control to be better. The use of low and dominant notes can make the speaker sound confident. Also, the tone is also able to release the tension experienced by the speaker. One stretch that can be done is to relax the muscles in the shoulders, neck, jaw, and throat all affect the tone of voice so they want to relax the muscles when talking. To relax those muscles here is an exercise that can be tried: start by taking a deep breath. When exhaling makes sounds that are half-yawned / half-sighing and soften the tension felt in the jaw, throat, neck, and shoulders.

Based on the generalisability design construction with the 2 facet model, namely the facets of the assessment/indicator aspects (i), and the observer/observer facets (o). As for students, in this case, it acts as an object of observation, not as a facet, because it is not a source of error in observation (Matt & Sklar, 2015). Furthermore, by entering the value of the variance field score above in the equation will produce a reliability coefficient of rxx'= 0.88. Based on the result of reliability estimation, the reliability coefficient in the category is high.

Important note from the panelists, Indonesian as a practical skill has aspects of assessment, which is different from other theoretical subjects. Listening and reading skills as receptive (receiving responses) have different characteristics, especially with other types of productive skills such as speaking and writing. Proficiency in this skill is measured by the ability of a person to receive information from his hearing instrument. Four language skills such as listening, speaking, reading and writing have different characteristics (Markhamah, Ngalim, Muinudinillah Basri, & Sabardila, 2017).

Some assessment techniques used by teachers cannot measure the ability of test-takers. The instructor explains the use of judgment techniques to measure the competency of test-takers from the results of the assessment for different competencies/dimensions. In the principle of measurement, competencies/dimensions that are unidimensional cannot be used as a reference for assessing other dimensions (Retnawati, 2017). Interdimensional composite scores used by the instructor's reference to provide an assessment are a balanced mean, that is, each skill has the same score value for the average participant's learning outcomes.

An easy way to judge someone's conversation is from the questions asked. Scientifically speaking that is interesting will provoke many questions and comments from listeners, even though comments that are contra/contradictory. Conversely, if there are no questions at all from the listener, it means that our scientific speaking is failing; our conversation is not interesting or boring (the listener might want our session to finish quickly to move on to the next speaker).

CONCLUSION

Each measurement produces information about the measurement results. Desired measurement information is not based on the individual being measured, but based on the information on the focus of measurement. The measurement information is based on the relationship between the test and the individual. Based on the results of the measurement of the ability to speak to speakers, the level of difficulty of the items produced is in the range that is categorized as good. At this level of ability, the test takers can show conversations that identify detailed information in the space and self-description of a particular person. The competency is illustrated by dialogue, activities, space division, and room specifications (store description). The test taker can distinguish unusual concepts, deceptive prepositions, and incorrect designations. This level of competence also measures the ability to show parts of simple to complex conversation following the ability level of each speaker.

LIMITATION AND STUDY FORWARD

This research has an adequate sample size, but it is better to involve a greater sample size to gain more reliable estimation. Item fit, item information function, and comparison reliability estimation of the test based on IRT (Item Response Theory) and CTT (Classical Test Theory) can be done as further analysis, as well as the DIF (Differential Item Function) analysis on the test.

AUTHORS CONTRIBUTION

Memet Sudaryanto (First Author) developing the instrument, gathering the data, analyzing the data, arranging research reports, and comparing the data. **Kundharu Saddhono** (Second Author) is validating and constructing theories, designing the blueprint, gathering data. **Lina** (Third Author) analyzing the data, proofing the draft, gathering other relevant theories.

REFERENCES

1. Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors. *Journal of Research in International Education*. https://doi.org/10.1177/1475240906065589
2. Brennan, R. L. (2010). Generalizability theory. In *International Encyclopedia of Education*. https://doi.org/10.1016/B978-0-08-044894-7.00246-3
3. Brooks-Lewis, K. A. (2009). Adult learners' perceptions of the incorporation of their L1 in foreign language teaching and learning. *Applied Linguistics*. https://doi.org/10.1093/applin/amn051
4. Cunningsworth, A., & Horner, D. (1985). The role of simulations in the development of communication strategies. *System*. https://doi.org/10.1016/0346-251X(85)90035-1
5. de Beaugrande, R. (2006). Critical discourse analysis: History, ideology, methodology. *Studies in Language and Capitalism*. https://doi.org/diskursanalyse; qualitative methode; ideologie
6. Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.78.2.350
7. Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). *Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework*. *19*(1), 72–91. https://doi.org/10.1037/a0032138
8. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory : principles and applications. In *Evaluation in education and human services;* https://doi.org/10.1017/CBO9781107415324.004
9. Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190501000071
10. Kamata, A., Turhan, A., & Darandari, E. (2003). Estimating reliability for multidimensional composite scale scores. *Annual Meeting of American Educational Research Association, Chicago, April 2003*.
11. Kuzar, R. (2003). Constructions: A construction grammar approach to argument structure. *Journal of Pragmatics*. https://doi.org/10.1016/S0378-2166(97)81937-6
12. Laufer, B., & Aviad-Levitzky, T. (2017). What Type of Vocabulary Knowledge Predicts Reading Comprehension: Word Meaning Recall or Word Meaning Recognition? *Modern Language Journal*. https://doi.org/10.1111/modl.12431
13. Markhamah, Ngalim, A., Muinudinillah Basri, M., & Sabardila, A. (2017). Comparison of Personal Pronoun between Arabic and Its Indonesian Translation of Koran. *International Journal of Applied Linguistics and English Literature*. https://doi.org/10.7575/aiac.ijalel.v.6n.5p.238
14. Matt, G. E., & Sklar, M. (2015). Generalizability Theory. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. https://doi.org/10.1016/B978-0-08-097086-8.44027-4
15. Mattarima, K., & Rahim Hamdan, A. (2011). The Teaching Constraints of English as a Foreign Language in Indonesia: The Context of School Based Curriculum. *Sosiohumanika*.
16. Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language and*

Communication Disorders. https://doi.org/10.1111/1460-6984.12061
17. Moore, K. D. ((2005). Effective instructional strategies: From theory to practice. In *Sage Publications.*
18. Nash, G., Crimmins, G., & Oprescu, F. (2016). If first-year students are afraid of public speaking assessments what can teachers do to alleviate such anxiety? *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2015.1032212
19. Nurgiyantoro, B. (2009). *Penilaian dalam pengajaran bahasa dan sastra.* Yogyakarta: BPFE.
20. Oradee, T. (2013). Developing Speaking Skills Using Three Communicative Activities (Discussion, Problem-Solving, andRole-Playing). *International Journal of Social Science and Humanity*. https://doi.org/10.7763/IJSSH.2012.V2.164
21. Paul, R., & Smith, R. L. (1993). Narrative skills in 4-year-olds with normal, impaired, and late- developing language. *Journal of Speech and Hearing Research*. https://doi.org/10.1044/jshr.3603.592
22. Penev, S., & Raykov, T. (2006). On the relationship between maximal reliability and maximal validity of linear composites. *Multivariate Behavioral Research*. https://doi.org/10.1207/s15327906mbr4102_1
23. Pyrik, J. (2015). Communicating Risk. In *Intelligence Communication in the Digital Era: Transforming Security, Defence and Business*. https://doi.org/10.1057/9781137523792_4
24. Rahmawati, L. E., Suwandi, S., Saddhono, K., & Setiawan, B. (2019). Need analysis on the development of writing competency test for foreign university students. *Humanities and Social Sciences Reviews*, *7*(3). https://doi.org/10.18510/hssr.2019.7368
25. Retnawati, H. (2008). *Validitas dan reliabilitas konstruk skor tes kemampuan calon mahasiswa*. (1), 126–135.
26. Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*. https://doi.org/10.21831/reid.v2i2.11029
27. Retnawati, H. (2017). Learning trajectory of item response theory course using multiple softwares. *Olympiads in Informatics*. https://doi.org/10.15388/ioi.2017.10
28. Saddhono, K. (2015). Integrating culture in Indonesian language learning for foreign speakers at Indonesian universities. *Journal of Language and Literature*. https://doi.org/10.7813/jll.2015/6-2/58
29. Sudaryanto, M., Mardapi, D., & Hadi, S. (2019). How foreign speakers implement their strategies to listen indonesian language? *Journal of Advanced Research in Dynamical and Control Systems*.
30. Sudaryanto, M., Saddhono, K., Wahyono, H., Widiatmi, T., Ino, L., Hidayat, H., … Pramono, P. (2019). Indonesian as a Foreign Language: Standard Setting and Materials Development Issues. *1st Workshop on Environmental Science, Society, and Technology, WESTECH*, 178–184. https://doi.org/10.4108/eai.8-12-2018.2284065
31. Townsend, S. S. M., Kim, H. S., & Mesquita, B. (2014). Are You Feeling What I'm Feeling? Emotional Similarity Buffers Stress. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550613511499
32. Winarni, R., Slamet, S. Y., & Saddhono, K. (2018). Development of Indonesian literature textbook with character education through information and communication technology (ICT) learning based. *International Journal of Engineering and Technology(UAE)*. https://doi.org/10.5465/amr.2009.35713291
33. Zyoud, M. M. (2016). Theoretical Perspective on How To Develop Speaking Skill. *An International Multidisciplinary Journal*.

