Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (10s2): 1021-1029

Inter-Rater Reliability for Assessing Digital Leadership Situational Judgement Test Linguistic Validation using Cohen Kappa

Nurhafizah Abdul Musid[1], Mohd Effendi Ewan Mohd Matore[2]*, Aida Hanim A. Hamid[3]

[1]Abdul Musid, N., Faculty of Education, The National University of Malaysia, Malaysia, *p113289@siswa.ukm.edu.my*

[2]*Mohd Matore, M. E. E., Research Centre of Education Leadership and Policy, Faculty of Education, The National University of Malaysia, Malaysia, *effendi@ukm.edu.my*

[3]A. Hamid, A. H., Research Centre of Education Leadership and Policy, Faculty of Education, The National University of Malaysia, Malaysia, *aidahanim@ukm.edu.my*

Received: 18- June -2023
Revised: 13- July -2023
Accepted: 20- August -2023

Abstract

The validity assessment using experts in nominal data in linguistic context to assess digital leadership is the Situational Judgement Test (SJT) item was rarely highlighted, especially the digital leadership aspect of teachers using the SJT. Linguistic validity is always neglectable compared to any other types of validation such as content and construct validity. Linguistic validation can be so impactful because it ensures that the majority of the respondents in a given population can understand the instrument and maintain appropriate reading and comprehension levels. Hence, the aim of this study is to examine the linguistic validity of the SJT for teachers' digital leadership using Cohen Kappa analysis. Cohen Kappa beneficial to analyze the agreement between two linguistic experts. This study used a purposive sampling, and it was conducted online. The experts were contacted through e-mail and online messaging applications. These two experts are language experts called *Munsyi Dewan* (*Bahasa*) and *Munsyi Muda Bahasa* who have more than eight years of experience in Malay Language. The two experts have to score the survey which contains 45 items for the nine constructs of digital leadership. Each construct has five items and the scale used is a dichotomous scale or binary data. The result shows a 97.78 percent agreement between the two experts. This means almost perfect agreement between the two experts. However, both experts made several comments to improve the items in the SJT. The comments on the linguistic aspects include particles, the use of italics, prepositions, passive voice, prefixes, vocabulary, and the use of English terms. This study can produce items on SJT for teachers' digital leadership with good linguistic validation and can be easily understood by the respondents. A follow-up study is planned to examine SJT items' psychometric properties on teachers' digital leadership using the Rasch Measurement Model.

Keywords: Situational Judgement Test, digital leadership, Cohen Kappa, teacher, linguistic validity.

INTRODUCTION

There are a number of instruments used to gather information, such as standardized and non-standardized tests as well as rating scales and checklists (Benson et al., 2019), rubrics, and questionnaires (Papadakis et al., 2020). However, for a thorough assessment of a person's actions, the Situational Judgement Test (SJT) is presented as an alternative (Abdul Musid et al., 2022). The SJT is not only a reliable instrument for measuring non-academic traits but also affordable (Patterson, Knight, et al., 2016; Patterson, Zibarras, et al., 2016). As such, the SJT provides a trustworthy instrument that can be used to evaluate emotional competencies, including professionalism and leadership (Goss et al., 2017; Patterson, Zibarras, et al., 2016). Therefore, the SJT is an appropriate instrument to assess teachers' digital leadership competencies.

However, the validity and reliability of any instrument developed and used in a study must be tested. First, the extent to which something indicates what it should measure is known as validity (Impellizzeri & Marcora, 2009). Validity is important as it is the most fundamental consideration in test development and test evaluation (American Educational Research Association (AERA) et al., 2014). Therefore, the validity of the SJT, which was developed to measure teachers' digital leadership competencies, must be examined. There are several types of validity that are commonly discussed, namely construct validity, face validity, criterion validity, and content validity. However, this study will focus on one more type of validity that is of less concern to researchers, namely linguistic validity. Most of the studies that discussed about linguistic validation focused on the translation aspect such as studies by Mandysova and Herr (2019), Farooq and Malik (2021), and Faran and Malik (2021).

Linguistic validation is performed by linguistic experts who have academic qualifications in languages at the level of a baccalaureate degree or higher. Normally, these experts have been in service for at least five years. If they are accredited by a professional association for languages, this confirms their expertise. The purpose of linguistic validation is to ensure that the items in the SJT have the correct use of terms or words, as well as the correct sentence structure and grammar. As Karthikeyan et al. (2015) and Mazurek et al. (2018) explain, linguistic validation is important to ensure that the majority of the respondents in a given population can understand the instrument and maintain appropriate reading and comprehension levels. In simpler language, the items in the instrument are clear and easy to understand. Another reason for using linguistic validation was to strengthen the reliability of the questionnaire (Althumiri et al., 2021).

Linguistic validation can be demonstrated empirically. The method used to analyze linguistic validation depends on the data type. The linguistic validation data collected in the present study are in binomial form, i.e., yes or no. Considering the fact that two raters are involved in this linguistic validation, Cohen Kappa analysis is suitable for obtaining empirical data. This fact is consistent with the view Armstrong et al. (2023) that Cohen Kappa may be used to quantitatively indicate the reliability two similar item measurements. While a zero kappa value shows no relationship, positive and negative values show agreement and disagreement, respectively (McHugh, 2012). In addition, inter-rater reliability can also be assessed using the Cohen Kappa analysis (Ayub et al., 2023).

LITERATURE REVIEW

Definition of Cohen Kappa

Jacob Cohen introduced Cohen-Kappa index analysis as a method for measuring the reliability of qualitative data (Main et al., 2021). Mahamod and Mohd Ishak (2003) stated that to determine a high reliability value for any item used to describe a theme, inter-rater agreement is required. These items whose reliability was measured can be used in the development of questionnaires (Main et al., 2021), and in this study, it refers to the SJT for teachers' digital leadership. The Cohen Kappa coefficient is the most common for measuring inter-rater agreement (Cohen, 1960), making it a metric for this particular measurement (Di Eugenio & Glass, 2004).

This method is mainly used in speech recognition (Mohd Zaman et al., 2014), biology (Fattahi et al., 2015), clinical (Wongpakaran et al., 2013), and other fields (Md Juremi et al., 2017). Essentially, as a measure of inter-rater agreement on item categories (Cohen, 1960), Cohen's Kappa coefficient is basically a method for measuring the degree to which two raters on qualitative items (categories) are in accord (Md Juremi et al., 2017). It is defined as:

$$K = \frac{p_A - p_E}{1 - p_E} \quad \text{..... (1)}$$

where p_A represents the relative observed inter-rater agreement and p_E represents the chance agreement between raters (Cohen, 1960). The value of K varies from 1 to -1 depending on the degree of agreement between raters (Md Juremi et al., 2017). $K = 1$ implies perfect agreement, and $K = 0$ implies that there is no agreement between raters beyond what would be expected by chance (Cohen, 1960). A perfect agreement of K that is 1 is rarely achieved, but a value close to 1 means that the degree of agreement is excellent (Md Juremi et al., 2017). On the other hand, if the value is -1, it means that the raters do not agree with the categories (Md Juremi et al., 2017).

According to Mahamod and Mohd Ishak (2003), no specific value is appropriate to indicate the level of agreement between experts, although according to Fleiss & Cohen (1973), a complete agreement and a consistent non-agreement is indicated by a kappa value of 1.0 and -1.0, respectively, and a kappa of 0 indicates a random level of agreement/disagreement between the two raters. Wood (2007) on the other hand, argues that a kappa value of at least 0.60 or 0.70 indicates strong agreement.

Linguistic Validation and SJT Development

Limited studies have been done on SJT to measure leadership as shown in Table 1. Only three research that use SJT to measure leadership for the past 14 years. From the previous study, it appears that several studies have been conducted to measure digital leadership, and these measurements were not made in terms of the SJT. However, there are SJT studies for leadership in general, namely studies by Grant (2009), Peus et al. (2013), and Grossman and Sharf (2018). Regarding the linguistic validation aspect, the three studies show that this has not been implemented. Therefore, there is room for strengthening linguistic validation in the instruments.

Table 1: - Research on SJT for leadership

Author & Year	Country	Title of article	Linguistic validation
(Grant, 2009)	United States of America	The Validation of a Situational Judgment Test to Measure Leadership Behavior	Null
(Peus et al., 2013)	Germany	Situation-based Measurement of the Full Range of Leadership Model- Development and Validation of a Situational Judgment Test	Null
(Grossman & Sharf, 2018)	United States of America	Situational Judgment Tests and Transformational Leadership: An Examination of the Decisions, Leadership, and Experience in Undergraduate Leadership Development	Null

From the aspect of contributing to the psychometric aspect, this study contributes to the improvement of item quality through linguistic validation. This study also contributes to the knowledge aspect in education when it generates new knowledge through appropriate items and constructs to measure teachers' digital leadership. The contribution of this study to the methodological aspect is the use of the SJT as an instrument to measure digital leadership compared to the commonly used instrument, namely the questionnaire with a Likert scale. Furthermore, this study uses a more complex and widely accepted analysis, namely the Cohen Kappa procedure to measure agreement between two parties. By conducting linguistic validation in this study, it is able to provide empirical evidence of the validity of language for a study in a local context. This is because linguistic inaccuracies in an SJT instrument can lead to misinterpretation in a measurement.

At this time, Malaysia began paying more and more attention to the growth of SJT. Studies on SJT development that use Cohen Kappa analysis are relatively rare, nevertheless. Essentially, the Cohen Kappa analysis can be applied at any point in the SJT development process, such as when individuals agree to a construct or item. It is also possible to assess expert agreement on the validity aspect using the Cohen Kappa statistic. The Cohen Kappa analysis only uses two raters; hence it might not be extensively applied. The Cohen Kappa analysis should not be used if there are more than two raters in the analysis.

METHODOLOGY

This section will explain on panel experts, instrument and data analysis of linguistic validity using Cohen Kappa.

Panel Experts

This study used a purposive sampling, and it was conducted online. The experts were contacted through e-mail and online messaging applications. Expert sampling is one of the variants of purposive sampling in which participants can be selected based on their specific skills or knowledge related to the topic of interest (Etikan et al., 2016). In this study, two experts were selected to determine linguistic validity using Cohen Kappa. Both had at least a master's degree in Bahasa Melayu and they have more than eight years of experience in Bahasa Melayu. They also held the title of *Munsyi Dewan (Bahasa)* and *Munsyi Muda Bahasa*. The first expert is an academic lecturer at a teacher training institute and the second expert is a teacher at a primary school. The validity of the study is compromised if the selected experts do not have sufficient expertise in the subjects studied (Zulkifli et al., 2022). Their feedback is crucial in determining the items in the SJT in terms of grammatical laws and the use of appropriate terms.

Instrument

This linguistic validation form consists of two parts, namely Part A: Basic information of the expert and Part B: Linguistic validation of the SJT item for digital leadership in teachers. In Part A, experts need to provide their full name, years of work experience, place of work, and academic qualifications. In Part B, there are a total of 45 items for the experts to score. These 45 items refer to nine constructs, each of which includes five items. The nine constructs include (1) student engagement, learning, and outcomes, (2) learning environment and spaces, (3) professional growth and learning, (4) communication, (5) public relations, (6) branding, (7) opportunities, (8) empowered professionals, and (9) learning catalyst.

Each item consists of a situation and four possible responses. For this linguistic validation, the expert must evaluate the situation and the four possible responses in sequence. Each item must be rated yes or no in terms of agreement with the linguistic aspects. For each item, there is also a comment box where

experts can make comments on the items of the SJT for teacher digital leadership, if needed. At the end of the form are overall comments and suggestions for improvement. Experts must sign this form agreeing to maintain the confidentiality of this SJT assignment. This is to avoid plagiarism issues if these SJT items are shared with the public.

Data Analysis

Linguistic validation is analyzed in this study using Cohen Kappa, since two experts are involved. This study uses a Cohen Kappa template found on a website. Cohen Kappa is one of the most popular ways to assess how well two different raters or analysis approaches are in accord (Craig, 1981). In this regard, the Kappa test is contingent on the random adjusted agreement coefficient, *K* (Vergni et al., 2021), as follows:

$$K = \frac{p_o - p_e}{1 - p_e} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

such that p_o represents the overall frequency of agreement observed, whereas p_e denotes the randomly expected agreement percentage. Practically, *K* denotes the extent to which the observed match frequency exceeds the match frequency p_e, expected from a random classification. Thus, a qualitative assessment of the strength of agreement (see Table 2) is occasionally used to assess the relative strength of agreement per the Kappa statistics.

Table 2: - Kappa statistics, *K* and the strength of agreement (Landis & Koch, 1977)

Kappa statistic	Strength of agreement
0.81–1.00	Almost perfect
0.61–0.80	Substantial
0.41–0.60	Moderate
0.21–0.40	Fair
0.00–0.20	Slight
< 0.00	Poor

RESULTS AND DISCUSSION

Table 3 presents the outcome for the agreement between two experts. Each item requires a yes or no decision from the experts.

Table 3: - Agreement between the two experts

Item	Expert 1	Expert 2
1	Yes	Yes
2	No	Yes
3	Yes	Yes
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	Yes	Yes
8	Yes	Yes
9	Yes	Yes
10	Yes	Yes
11	Yes	Yes
12	Yes	Yes
13	Yes	Yes
14	Yes	Yes
15	Yes	Yes
16	Yes	Yes
17	Yes	Yes
18	Yes	Yes
19	Yes	Yes
20	Yes	Yes
21	Yes	Yes
22	Yes	Yes
23	Yes	Yes
24	Yes	Yes

25	Yes	Yes
26	Yes	Yes
27	Yes	Yes
28	Yes	Yes
29	Yes	Yes
30	Yes	Yes
31	Yes	Yes
32	Yes	Yes
33	Yes	Yes
34	Yes	Yes
35	Yes	Yes
36	Yes	Yes
37	Yes	Yes
38	Yes	Yes
39	Yes	Yes
40	Yes	Yes
41	Yes	Yes
42	Yes	Yes
43	Yes	Yes
44	Yes	Yes
45	Yes	Yes

From the analysis on the website, the number of observed agreements and agreements expected by chance is 44 (97.78 percent of observations), respectively. This shows a 97.78 percent agreement between the two experts. Table 1 shows almost perfect agreement between the two experts. Of the 45 items, 44 items were rated 'yes' by both experts. Meanwhile, one item was rated 'no' by the first expert and 'yes' by the second expert. Even though most of the items were rated 'yes' by both experts, there are still comments from these experts and the items need to be revised. However, the corrections are considered minor corrections in terms of grammatical point of view.

All items in this digital SJT for teacher leadership are in Malay. One of the comments received refers to particles in grammar. There are several particles in Malay such as -lah, -kah, -tah and -pun. According to Chye and Subramaniam, (2012), in the interrogative verse structure of Malay, when the verse begins with an interrogative word, the use of -kah is required, and conversely, when the interrogative word is used at the end of the verse, the particle -kah is aborted. This confusion arises because the usage of everyday language and formal writing is different (Chye & Subramaniam, 2012). For example, the form of the interrogative verse for item 5 "Bagaimana cara anda memberi tunjuk ajar kepada murid?" is ungrammatical. So, the grammatical form of this interrogative sentence should be "Bagaimanakah cara anda memberi tunjuk ajar kepada murid?". In written language, interrogative sentences that go through the process of prioritizing interrogative words must receive the particle -kah in the element that is brought forward (Karim et al., 2008).

In addition, the experts also comment on the use of italics. This is because some of the proper nouns used in the items of SJT use English words. Examples include WhatsApp, Telegram, Facebook, and Google Drive. Therefore, these proper nouns are written in italics. The function of italic letters in writing is to show the use of foreign language words or phrases in writing (Osman & Yusoff, 2019). However, the linguist consulted for this study explained that the English proper nouns in these items of SJT do not need to be italicised.

Continuing with the use of prepositions in a sentence. Prepositions are words that come before the name phrase (Karim et al., 2008). In Malay, there are several prepositions such as 'di', 'bagai', 'sejak', 'untuk', 'daripada' and 'kepada'. Among the warnings of language experts is the use of 'di dalam' and 'dalam'. For example, the correct phrase for item 1 is 'di dalam kelas' instead of 'dalam kelas' because the class is a room. As Karim et al. (2008) explains, the preposition 'di' is used specifically before nouns or noun phrases that denote a place. This preposition is written separately from the noun or noun phrase that follows it. The preposition 'di' cannot be used before nouns or noun phrases that describe time, period, or age.

The next aspect of language commented on by the experts is the passive voice. The passive has the precedence of the subject over the heading (Karim et al., 2008). Karim et al. (2008) further explains that in languages such as Malay and English, the relationship between active and passive is generally to change the place of the name phrase, which is the subject and object of the verse in question. An example of a passive given by a language expert is item 2, namely '...diberi guru...' must be changed to '...diberi oleh guru...'. The passive is also formed

from the transitive active and contains verbs that focus on the object of origin as the title or element being explained (Karim et al., 2008).

In addition, several words with English terms are used in this SJT items. Linguists recommend writing these words in Malay first and putting the English terms in parentheses. Nevertheless, the English terms must be inserted because they are more familiar to the respondents. For example, English terms such as ‘pen drive’ and ‘external hard drive’ are better recognized by respondents than similar terms in Malay. Although both Malay and English terms are provided, the translation must be accurate. Ensuring valid translation quality can help reduce sampling errors, increase the number of questionnaire responses, and improve the generalizability of the results (Kalfoss, 2019).

Furthermore, the aspect of language on which the language experts’ comment, namely the use of prefixes, also needs to be explored in the SJT items. Prefixes are affixes that are added or appended to the beginning of the base to form a new word (Katamba, 2005). Prefixes generally have an easily understood meaning applied to the word origin that they function in (Katamba, 2005). Examples of prefixes commonly used in Malay are 'ber-' and 'mem-'. Language experts have commented on item 7, so the word prefix 'ber-' is deleted. Language experts believe that the root word, namely 'kumpulan', is more suitable to be used without a prefix in the verse in question.

Another suggestion for improvement made by the language experts concerns vocabulary. There are some words that need to be replaced, for example, 'platform' in item 9 needs to be changed to 'tapak' and 'menggunakan' in item 10 should be replaced with 'melalui'. By definition, vocabularies are words that we must be familiar with in order to ensure effective communication, particularly through receptive vocabulary (listening) and expressive vocabulary (speaking) (Neuman & Dwyer, 2009). Therefore, the choice of words used must be appropriate and able to meet the intent of the verse being promoted. This can also prevent respondents from understanding the verse to be formed differently.

CONCLUSION

The results show almost perfect agreement between the two experts for nine constructs of digital leadership among teachers. Both experts made several comments to improve the items include particles, the use of italics, prepositions, passive voice, prefixes, vocabulary, and the use of English terms. Corrections to the SJT items were made according to the experts' suggestions to ensure that all items conform to the grammatical laws of Bahasa Melayu. The unexpected finding in this study is the use of italics for English proper nouns in the SJT items. Although all items are phrased in Malay, the experts indicated that English proper nouns such as Facebook, Telegram, and Google Drive do not need to be italicized. Besides, this study is restricted to two language experts, and such a practical constraint makes the study unable to provide an extensive review from numerous experts. It would be interesting to compare experiences from many experts from different kinds of linguistic aspects. In addition, this linguistic validation study is unique compared to others because linguistic validation is conducted on instruments in the form of SJTs that include situations and action responses. In contrast, other linguistic validations use survey instruments via questionnaires in which a Likert scale is employed. The results of this study should be of interest to stakeholders as this study develops appropriate SJT items that can be presented to teachers to assess their digital leadership. The SJT items with good linguistic validation can be easily understood by the respondents. Further research might use a larger number of language experts with other methods of data analysis. This study is important because it discusses the grammatical laws of Bahasa Melayu, which is written in English, so those who are interested in studying a foreign language can use this paper as a reference. Future trials should assess the psychometric properties of the items of the SJT on teachers’ digital leadership using the Rasch Measurement Model as a way forward. Thus, knowledge is developed by using a new, modern theory compared to Classical Test Theory.

CONFLICT OF INTEREST

The authors declare that there were no financial or commercial relationships that might cause a possible conflict of interest throughout the execution of this research.

AUTHOR CONTRIBUTIONS

Conceptualization, Nurhafizah Abdul Musid; design, Mohd Effendi Ewan Mohd Matore and Aida Hanim A. Hamid; drafting manuscript, Nurhafizah Abdul Musid and Mohd Effendi Ewan Mohd Matore; writing, Nurhafizah Abdul Musid and Mohd Effendi Ewan Mohd Matore; critical version of manuscript, Mohd Effendi

Ewan Mohd Matore and Aida Hanim A. Hamid; supervision, Mohd Effendi Ewan Mohd Matore and Aida Hanim A. Hamid. The published version of the manuscript has been thoroughly read and agreed upon by all authors.

FUNDING

The Faculty of Education, Universiti Kebangsaan Malaysia (UKM) provided funding for this research under GP-2021-K021854 (Publication Reward Grant) and GG-2022-020 (Research Fund of FPEND).

ACKNOWLEDGEMENTS

A token of appreciation goes to Mohd Effendi Ewan Mohd Matore and Aida Hanim A. Hamid, both my supervisors from the Faculty of Education, UKM, and Ministry of Higher Education (MOHE) for the great opportunity to carry out this research as well as for their guidance throughout the entire research process. Special thanks also to Kumpulan Penyelidikan Universiti (KPU), Penilaian Pendidikan (Educational Evaluation) UKM.

REFERENCE

1. Abdul Musid, N., Mohd Matore, M. E. E., & A. Hamid, A. H. (2022). The Prospective of Situational Judgement Test in Assessing Individual Performance. *International Journal of Academic Research in Progressive Education and Development*, *11*(3), 1810–1819. https://doi.org/DOI:10.6007/IJARPED/v11-i3/15401
2. Althumiri, N. A., Basyouni, M. H., AlMousa, N., AlJuwaysim, M. F., Almubark, R. A., BinDhim, N. F., Alkhamaali, Z., & Alqahtani, S. A. (2021). Obesity in Saudi Arabia in 2020: Prevalence, Distribution, and Its Current Association with Various Health Conditions. *Healthcare*, *9*(3), 1–8.
3. American Educational Research Association (AERA), American Psychological Association (APA), & National Council for Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.
4. Armstrong, C. M., He, Y., Chen, C. Y., Counihan, K., Lee, J., Reed, S., & Capobianco, J. (2023). Use of a Commercial Tissue Dissociation System to Detect Salmonella-Contaminated Poultry Products. *Analytical and Bioanalytical Chemistry*, 1–6.
5. Ayub, I., Rehman, M., Nawaz, M., Jabbar, M., Butt, H., & Jamil, F. (2023). Inter-Rater Reliability to the Assessment of Ramus Relationship of Mandibular Impacted Third Molar Among Denitsts: An Orthopantomographic Study. *Pakistan Journal of Medical & Health Sciences*, *17*(1), 394–396.
6. Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test Use and Assessment Practices of School Psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, *72*, 29–48.
7. Chye, Y. C., & Subramaniam, V. (2012). Analisis Kesilapan dalam Pembelajaran Bahasa Melayu oleh Pelajar Asing. *GEMA Online Journal of Language Studies*, *12*(2), 667–692.
8. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104
9. Craig, R. T. (1981). Generalization of Scott's Index of Intercoder Agreement. *Public Opinion Quarterly*, *45*, 260–264.
10. Di Eugenio, B., & Glass, M. (2004). The Kappa Statistic: A Second Look. *Computational Linguistics*, *30*(1), 95–101.
11. Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11
12. Faran, M., & Malik, F. (2021). Translation and Linguistic Validation of Perma Profiler in Music Engagers Context. *Palarch's Journal of Archaeology of Egypt/Egyptology*, *17*(3), 4102–4120. https://www.archives.palarch.nl/index.php/jae/article/view/6095%0Ahttps://www.archives.palarch.nl/index.php/jae/article/download/6095/5985
13. Farooq, S., & Malik, F. (2021). Translation, Cultural Adaptation and Linguistic Validation of Motives for Risk-Taking Scale. *ASEAN Journal of Psychiatry*, *22*(9), 1–12.
14. Fattahi, S., Othman, Z., & Zulaiha Ali Othman, Z. (2015). New Approach for Imbalanced Biological Dataset Classification. *Journal of Theoretical and Applied Information Technology*, *72*(1), 40–57.
15. Fleiss, J. L., & Cohen, J. (1973). The Equivalence of Weighted Kappa and The Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, *33*(3), 613–619.

https://doi.org/10.1177/001316447303300309
16. Goss, B. D., Ryan, A. T., Waring, J., Judd, T., Chiavaroli, N. G., O'Brien, R. G., Trumble, S. C., & McColl, G. J. (2017). Beyond Selection: The Use of Situational Judgement Tests in the Teaching and Assessment of Professionalism. *Academic Medicine*, *92*(6), 780–784.
17. Grant, K. L. (2009). *The Validation of a Situational Judgment Test to Measure Leadership Behavior*. (Master's thesis, Western Kentucky University). Retrieved from https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1064& context=theses
18. Grossman, G., & Sharf, R. (2018). Situational Judgment Tests and Transformational Leadership: An Examination of the Decisions, Leadership, and Experience in Undergraduate Leadership Development. *Journal of Leadership Education*, *17*(1), 114–131. https://doi.org/10.12806/v17/i1/r4
19. Impellizzeri, F. M., & Marcora, S. M. (2009). Test Validation in Sport Physiology: Lessons Learned from Clinimetrics. *International Journal of Sports Physiology and Performance*, *4*(2), 269–277.
20. Kalfoss, M. (2019). Translation and Adaption of Questionnaires: A Nursing Challenge. *SAGE Open Nursing*, *5*(0319), 1–13. https://doi.org/10.1177/2377960818816810
21. Karim, N. S., M. Onn, F., Musa, H., & Mahmood, A. H. (2008). *Tatabahasa Dewan* (Edisi ke-3). Kuala Lumpur: Dewan Bahasa dan Pustaka.
22. Karthikeyan, G., Manoor, U., & Supe, S. S. (2015). Translation and Validation of the Questionnaire on Current Status of Physiotherapy Practice in the Cancer Rehabilitation. *Journal of Cancer Research and Therapeutics*, *11*(1), 29–36.
23. Katamba, F. (2005). *English Words, Structure, History, Usage* (Edisi ke-2). New York: Routledge.
24. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.
25. Mahamod, Z., & Mohd Ishak, N. (2003). Analisis Cohen Kappa dalam Penyelidikan Bahasa: Satu Pengalaman. *Prosiding Seminar Penyelidikan Guru Peringkat Kebangsaan 2003*, 1–7. http://www.ipbl.edu.my/portal/penyelidikan/seminarpapers/2003/zamriUKMkk1.pdf
26. Main, A., Yahya, R., & Mahat, H. (2021). Inter-Rater Reliability Assessment for Motivating Factors in Blood Donation Using Cohen's Kappa Analysis. *Politeknik & Kolej Komuniti Journal of Social Sciences and Humanities*, *6*(1), 13–22.
27. Mandysova, P., & Herr, K. (2019). The Translation and Linguistic Validation of the Revised Iowa Pain Thermometer into Czech for a Clinical Study Involving Czech Stroke Patients. *Kontakt*, *21*(1), 55–64. https://doi.org/10.32725/kont.2019.015
28. Mazurek, J., Sutkowska, E., Szcześniak, D., Urbańska, K. M., & Rymaszewska, J. (2018). FIMA, the Questionnaire for Health-Related Resource Use in the Elderly Population: Validity, Reliability, and Usage of the Polish Version in Clinical Practice. *Clinical Interventions in Aging*, *13*, 787–795.
29. McHugh, M. L. (2012). Interrater Reliability: The Kappa Statistic. *Biochemica Medica*, *22*(3), 276–282. https://hrcak.srce.hr/89395
30. Md Juremi, N. R., Zulkifley, M. A., Hussain, A., & Wan Zaki, W. M. D. (2017). Inter-Rater Reliability of Actual Tagged Emotion Categories Validation Using Cohen's Kappa Coefficient. *Journal of Theoretical and Applied Information Technology*, *95*(2), 259–264.
31. Mohd Zaman, M. H., Mustafa, M. M., & Hussain, A. (2014). Inter-rater Reliability of Accessing the Intelligibility of Band-limited Transformed Speech Using Nonlinear Frequency Compression. *2014 IEEE 2014 Proceedings of International Conference on Computer, Communications, and Control Technology*, 126–129. https://doi.org/10.1109/I4CT.2014.6914160
32. Neuman, S. B., & Dwyer, J. (2009). Missing in Action: Vocabulary Instruction in Pre-K. *The Reading Teacher*, *62*(5), 384–392. https://doi.org/10.1598/rt.62.5.2
33. Osman, Z., & Yusoff, N. (2019). Retorik Penulisan Ilmiah: Penilaian Berdasarkan Prinsip Kerjasama Grice. *International Journal of Language Education and Applied Linguistics*, *09*(1), 69–83. https://doi.org/10.15282/ijleal.v9.1196
34. Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020). Developing and Exploring an Evaluation Tool for Educational Apps (E.T.E.A) Targeting Kindergarten Children. *Sustainability*, *12*(4201), 1–10.
35. Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How Effective are Selection Methods in Medical Education? A Systematic Review. *Medical Education*, *50*(1), 36–60.
36. Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational Judgement Tests in Medical Education and Training: Research, Theory and Practice: AMEE Guide No.100. *Medical Teacher*, *3*(1), 3–17.
37. Peus, C., Braun, S., & Frey, D. (2013). Situation-based Measurement of the Full Range of Leadership Model-Development and Validation of a Situational Judgment Test. *The Leadership Quarterly*, *24*(5), 777–795.

https://doi.org/10.1016/j.leaqua.2013.07.006
38. Vergni, L., Todisco, F., & Di Lena, B. (2021). Evaluation of the Similarity between Drought Indices by Correlation Analysid and Cohen's Kappa Test in a Mediterranean Area. *Natural Hazards*, *108*(2), 2187–2209.
39. Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A Comparison of Cohen's Kappa and Gwet's AC1 when Calculating Inter-Rater Reliability Coefficients: A Study Conducted with Personality Disorder Samples. *BMC Medical Research Methodology*, *13*(61), 1–7. https://doi.org/10.1186/1471-2288-13-61
40. Wood, J. M. (2007). Understanding and Computing Cohen's Kappa: A Tutorial. *WebPsychEmpiricist*. Retrieved from http://works.bepress.com/james_wood/22/
41. Zulkifli, H., Rashid, S. M. M., Mohamed, S., Toran, H., Raus, N. M., Pisol, M. I. M., & Suratman, M. N. (2022). Designing the Content of Religious Education Learning in Creating Sustainability among Children with Learning Disabilities: A Fuzzy Delphi Analysis. *Frontiers in Psychology*, *13*(1036806), 1–13. https://doi.org/doi: 10.3389/fpsyg.2022.1036806