SHS Web of Conferences **224**, 04006 (2025) https://doi.org/10.1051/shsconf/202522404006
*ICAP-H 2025*

# A comparative analysis of 2-option and 3-option response formats for computerized work aptitude test

*Rafli* Sodiq Bagaskara*[1] and *Anjar* Kartaputra[1]

[1]Anvesana Research Institute, Indonesia

**Abstract.** This study aims to identify the optimal response format for a work aptitude test by evaluating its psychometric properties. This study employed a quantitative, quasi-experimental method utilizing a post-test only control group design. Following data cleaning, a total of 45 participants, divided into two groups, were included in the final analysis. The instruments utilized were the Focusync Test, administered with both 2-option and 3-option response sets, and the Pauli Test. The results revealed that the 3-option format demonstrated higher reliability, as indicated by Cronbach's alpha, in measuring aspects such as total responses, errors, and variability, compared to the 2-option format. Nevertheless, the 3-option response format showed more adequate convergent validity than the 2-option format. Therefore, the 3-option response format is recommended for future computerized work aptitude tests.

## 1 Introduction

The operationalization of work aptitude tests has undergone a transformation from traditional paper-and-pencil tests (PBT) to computer-based testing (CBT). The shift in the medium of aptitude tests, commonly referred to as "Tes Koran," has been driven by various considerations, including efforts to standardize test administration, improve scoring efficiency, and reduce the risk of human error. Conventional aptitude tests such as the Kraepelin and Pauli, officially recognized by the Indonesian Psychological Association (Himpunan Psikologi Indonesia; HIMPSI) under Decree No. 024/SK/PP-HIMPSI/VIII/18, have long been subject to conceptual and procedural inconsistencies among psychologist in Indonesia [1]. The operationalization of these tests requires candidates to perform rapid and repetitive mathematical operations [2], which renders both administration and scoring vulnerable to error [3]. As a result, assessment outcomes are at risk of lacking validity and reliability. This is particularly concerning given that aptitude tests are high-stakes assessments that demand careful consideration of validity, reliability, and fairness [4]. The importance of aptitude measurement lies in its ability to capture critical psychological dynamics of prospective employees, ranging from adaptability to work speed [5]. In practical terms, work aptitude is a psychological attribute that supports individuals in achieving various accomplishments across activities, situations, and professional domains [6]. The lack

∗. Corresponding author: sodiq@anvesana.id

of validity and reliability in these paper-based tests has prompted the development of computer-based tests.

The development of computer-based aptitude tests has generally followed two approaches: adaptation and test construction. The adaptation approach involves transferring the existing test from a paper-and-pencil format to a digital platform without altering its theoretical constructs, as exemplified by the online versions of the Kraepelin test developed by Indriani et al. [7] and Nada et al. [8]. In this approach, responses are collected in an open-ended format, requiring participants to manually enter their answers in designated fields. In contrast, the test construction approach seeks to redesign aptitude tests in a manner more compatible with computer-based media. Examples include the Faxtor's Endurance and Speed Test (F-EAST) [9], Pauline, Qutest [10], and the Galang-Satrio Adaptability Test (GSAT). These tests generally employ closed-response formats, either dichotomous or multiple-choice. Notably, the variation in item formats across these instruments indicates a lack of consensus among researchers regarding the most effective item type for measuring work aptitude.

In fact, one of the most salient features of CBT-based aptitude tests is their reliance on dichotomous response formats. For instance, the F-EAST requires participants to solve mathematical operations and then select one of two responses: "1" to indicate an odd result and "0" to indicate an even result [11]. Similarly, Pauline employs a dichotomous format but provides more explicit answer options: "odd" and "even". In contrast, Qutest utilizes a trichotomous response format with figural items [12]. The differences across these instruments suggest the absence of agreement among test developers regarding the optimal number of distractors for aptitude test measurement.

Theoretical analyses have consistently suggested that three-option items tend to yield superior psychometric properties compared to two-option formats. Statistically, trichotomous items reduce the probability of correct guessing to 33.3%, whereas dichotomous items allow up to a 50% chance of correct guessing. This makes three-option formats more effective in mitigating guessing effects that threaten item validity. The reduced susceptibility to guessing also lowers error rates and enhances item discrimination. These findings are consistent with Rodriguez [13], who demonstrated that reducing response formats from five options to two led to declines in item difficulty, discrimination, and reliability coefficients, whereas reducing from five to three options decreased item difficulty but preserved discrimination and reliability. Similarly, Gierl et al. [14] concluded that three-option formats with two distractors were the most effective overall in balancing item difficulty, discrimination, and reliability.

Despite these advantages, the three-option format is more commonly employed in knowledge-based tests than in aptitude assessments, largely due to the limited research on the latter. Given the psychometric potential of the three-option response format, this study aims to compare the quality of psychometric properties between dichotomous and trichotomous formats in work aptitude tests. The findings are expected to provide a fundamental basis for the future development of computer-based aptitude assessments.

# 2 Methods

## 2.1 Participants

A total of 54 undergraduate students from a university in Bandung participated in this study. Following data cleaning for completeness and accuracy, nine participants were excluded. The remaining 45 cases ($M_{\text{age}} = 20.4$, $SD_{\text{age}} = 1.12$; 67.4% female) were retained for subsequent data analyses. The total sample size was determined by practical considerations. Given that each group only contained 22–23 participants, the results could be threatened by errors arising from limited statistical power. Therefore, these findings should be considered a preliminary study, highlighting the need for larger samples in future research. Data collection was conducted simultaneously in person under the supervision of a licensed psychologist and three trained assistants who served as observers. Each participant completed the test using a laptop connected to the Scalify platform, which was equipped with a proctoring system. This procedure was intended to ensure that participants engaged seriously with the tasks and refrained from dishonest behavior. All participants read and signed an informed consent form prior to the commencement of the experiment.

## 2.2 Research Instrument

### 2.2.1 Focusync

Focusync is an assessment tool developed to measure work aptitude. It evaluates aptitude through a series of repetitive tasks involving simple arithmetic operations, requiring participants to respond as quickly as possible. The test consists of 20 sheets (items), each with a maximum of 100 arithmetic problems. The assessment is designed to elicit psychological attributes associated with work capacity, including productivity (measured by the number of responses), accuracy (measured by the number of errors), emotional stability (measured by variability in responses), achievement motivation (measured by the height of the performance curve), and vitality (measured by the peak point). In this study, Focusync was modified into two response formats: trichotomous (three-option) and dichotomous (two-option). The total test duration ranged from 20 to 30 minutes.

**Table 1.** Definitions of Each Aspects Measured by Focusync

| Aspects | Definition |
|---|---|
| Total Responses | The overall number of items completed by the participant, regardless of accuracy. |
| Errors | The number of incorrect responses produced during the test. |
| Variability | The degree of fluctuation or inconsistency in performance across test items. |
| Curve Height | The highest point reached in the performance curve, reflecting peak productivity. |
| Peak Position | The point in the test at which maximum performance occurs. |
| Corrections | The number of errors that were identified and corrected by the participant. |

https://doi.org/10.1051/shsconf/202522404006

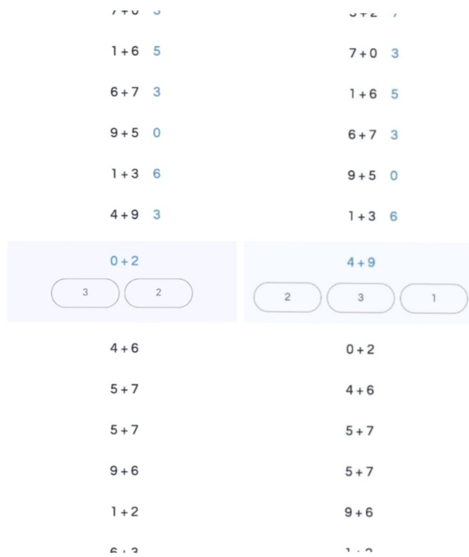


**Fig. 1.** Focusync Dichotomous (Left) and Trichotomous (Right) Response Formats

#### 2.2.2 Pauli Test

The Pauli Test, also widely known as the "Tes Koran," is a psychological assessment primarily intended to measure work aptitude, although some psychologists have also used it to assess personality traits and work attitudes. The test requires participants to repeatedly perform simple arithmetic operations. The resulting data are used to capture various psychological attributes, including productivity (total number of responses), accuracy (errors), variability (response fluctuations), curve height, peak position, and number of corrected operations. These dimensions are intended to provide an overview of individual work-related psychological attributes, including: Endurance, defined as the ability to work consistently over extended periods under pressure; Work Tempo (Speed), defined as the rate at which monotonous tasks are completed; Accuracy, defined as the ability to remain focused and minimize errors in repetitive tasks; Emotional Stability, reflected in whether performance patterns remain steady or fluctuate drastically; and Achievement Motivation (Willpower), defined as the extent of an individual's drive to complete demanding and monotonous tasks.

### 2.3 Data Analysis Techniques

This study employed a quantitative experimental approach. Data analysis began with a cleaning process, in which cases with missing values were excluded. Following this procedure, 45 participants who had completed the assessments in full were retained for analysis. Participants were randomly assigned into two groups: the dichotomous group (completing the two-option Focusync) and the trichotomous group (completing the three-option Focusync). The dataset was then analyzed for psychometric properties, specifically internal consistency reliability and validity. Internal consistency reliability was examined using Cronbach's alpha based on the indices of productivity, errors, and variability. Validity was assessed through both convergent validation procedures. Convergent validity was examined by correlating Focusync scores with Pauli scores. All correlation analyses were conducted using Spearman's rho.

### 2.4 Experimental Procedure

The study employed an experimental design using a Posttest-Only Control Group framework, whereby reliability and validity were analyzed after participants completed the assessments. The procedure involved the following steps: 1) Preliminary Data Collection: Participants completed a demographic questionnaire including age, gender, and domicile; 2) Preparation and Administration: Participants signed an informed consent form and were given instructions for the tests; 3) Random Assignment: Upon entering the platform, participants were randomly assigned by an automatically generated token to either the two-option test (control group) or the three-option test (experimental group); 4) Test Administration: Participants first completed the Focusync test, followed by a 15-minute break, and then the Pauli Test; 5) Data Collection and Debriefing: All Pauli test sheets were collected, and participants were provided with a brief explanation of the actual purpose of the study.

## 3 Result & Discussion

A total of 45 participants were included in the analysis, with ages ranging from 18 to 22 years ($M_{\text{age}} = 20.4$, $SD_{\text{age}} = 1.12$). The sample consisted of 30 females (66.7%) and 15 males (33.3%). Most participants were from West Java, predominantly residing in Bandung City (n = 18; 40.0%), followed by Bandung Regency (n = 9; 20.0%), with the remainder distributed across other cities and regencies such as Bekasi, Bogor, and Tasikmalaya. Participants were randomly assigned to two groups: the trichotomous group (n = 23) and the dichotomous group (n = 22).

### 3.1 Reliability Analysis

Reliability analysis is a statistical procedure used to assess the consistency of a test. In this study, reliability was examined using an internal consistency approach with Cronbach's alpha coefficients. All analyses were conducted using Jamovi (Version 2.6.22.0).

**Table 2.** Comparison of Cronbach's Alpha Reliability Coefficients for Trichotomous and Dichotomous Response Formats

| Aspects | N Item | Alpha Trichotomous (N=23) | Alpha Dichotomous (N=23) |
|---|---|---|---|
| Total Responses | 20 | 0.991 | 0.984 |
| Errors | 20 | 0.718 | 0.582 |
| Variability | 16 | 0.676 | 0.680 |
| Curve Height | 1 | - | - |
| Peak Position | 1 | - | - |

Overall, the three-option format demonstrated superior internal consistency compared to the two-option format across the dimensions of total responses, errors, and variability. The most pronounced difference was observed in the error dimension, where the three-option group reached $\alpha = 0.718$ (above the .70 threshold), while the two-option group remained at $\alpha = 0.582$. According to Hair et al. [13], reliability coefficients above .70 are considered acceptable, suggesting that the three-option format is more reliable and could provide consistency score measurement. This finding is consistent with theoretical arguments that three-option formats, by providing two distractors, reduce the probability of correct guessing [11]. On the other hand, the variability coefficients for both formats fell below the reliability cutoff, with only a minor difference of .04. Nevertheless, values in the .60–.70 range may be considered marginally acceptable for exploratory studies.

### 3.2 Convergent Validity

Convergent validity was examined to determine whether the indicators used in this study accurately measured the same underlying construct. Because each group contained fewer than 30 participants, achieving a normal distribution was considered unlikely. Therefore, non-parametric correlations (Spearman's rho) were employed.

**Table 3.** Comparison of Cronbach's Alpha Reliability Coefficients for Trichotomous and Dichotomous Response Formats

| Focusync | | Pauli | | | | | |
|---|---|---|---|---|---|---|---|
| Methods | Aspects | Total Responses | Errors | Variability | Curve Height | Peak Position | Corrections |
| Dichotomous (N=22) | Total Responses | 0.227 | -0.031 | 0.025 | 0.122 | 0.031 | -0.341 |
| | Errors | 0.128 | 0.043 | -0.226 | 0.057 | *-0.453** | 0.421 |
| | Variability | -0.332 | -0.179 | 0.190 | 0.128 | -0.053 | 0.185 |
| | Curve Height | 0.013 | 0.196 | 0.229 | 0.197 | -0.048 | 0.213 |
| | Peak Position | -0.047 | 0.497* | 0.361 | 0.053 | -0.021 | -0.137 |
| Trichotomous (N = 23) | Total Responses | **0.637**** | 0.041 | -0.402 | 0.092 | 0.382 | 0.144 |
| | Errors | 0.044 | 0.279 | 0.084 | -0.057 | -0.05 | 0.259 |
| | Variability | -0.299 | 0.136 | 0.302 | 0.01 | -0.104 | 0.137 |
| | Curve Height | 0.461* | 0.163 | -0.195 | 0.181 | 0.282 | 0.123 |
| | Peak Position | -0.148 | -0.071 | 0.071 | 0.262 | 0.094 | -0.114 |

*Note.* Correlations with a large effect size ($\geq |.50|$) are in boldface, correlations with a medium effect size ($\geq |.30|$) are in italics.
* < 0,05
** < 0,01
*** < 0,001

The results indicated stronger convergent validity for the three-option format. Specifically, the three-option format demonstrated superior correlations between analogous aspects (e.g., total responses, variability) with the Pauli test, whereas the two-option format showed no significant correlations across these aspects. Moreover, the two-option format displayed violations of convergent validity, as reflected in a significant negative correlation between errors and peak position, and a positive correlation between peak position and errors in the Pauli test. These results suggest that the two-option format failed to capture the intended constructs, while the three-option format more accurately reflected the relevant psychological attributes.

The findings support the psychometric superiority of the three-option response format. This aligns with prior evidence indicating that three-option formats provide the optimal balance between efficiency and psychometric quality. Rodriguez's meta-analysis identified the three-option format as the most effective in preserving validity and reliability, while reducing guessing. Similarly, Tarrant and Ware [14], and Esmaeeli et al. [15] demonstrated that three-option items reduce cognitive load, thereby optimizing test performance without compromising psychometric quality. However, previous studies were conducted within the context of knowledge-based tests and did not explore this format's potential for work aptitude tests. This study shows that even in the context of work aptitude assessment, the trichotomous response format has more robust psychometric properties than the dichotomous response format. Thus, in the context of work aptitude assessment, the three-option format appears to be more effective.

Despite these contributions, the study has several limitations. Convergent validity was not established across all dimensions, such as errors, variability, height, and peak position. Notably, the three-option format demonstrated higher correlations between height and peak position than within the same dimensions, although these did not reach statistical significance. Furthermore, this study relied solely on the Pauli Test as a reference instrument, without examining associations with the Kraepelin Test. The relatively homogeneous sample, drawn from a single educational institution, also limits generalizability. Finally, the small sample sizes in both groups (n=22 and n=23) restricted the statistical power of the analyses. Future research with larger and more diverse samples is needed to further validate the robustness of the three-option format.

## 4 Conclusion

This study demonstrates that the three-option response format is superior to the two-option format in terms of psychometric properties. Reliability analyses indicated higher internal consistency for the three-option format, particularly in the dimensions of total responses and errors. Convergent validity results further confirmed the advantage of the three-option format, as it yielded stronger and more construct-consistent correlations with the Pauli Test. These findings suggest that three-option items provide a more accurate and psychometrically sound measurement of work aptitude. Future studies should extend this line of research with larger sample sizes and broader predictive variables.

## 5 Implication

The findings provide practical recommendations for researchers and test developers: the three-option response format should be prioritized in the design of work aptitude tests, as it offers superior psychometric properties compared to dichotomous formats.

**Acknowledgements:** The authors would like to thank Muhammad Khansa Fikri and Yoga Aditya for their assistance as enumerators and research assistants in this study. The authors would also like to thank Scalify for providing the assessment platform for this study.

**Funding:** This study was supported by Anvesana Research Institute.

**Data Availability Statement:** The datasets of the current study are available from the corresponding author on reasonable request.

**Authors' Contributions Statement:** RSB did data curation, data analysis, and Writing – original draft. AK did Conceptualization, Project administration, Funding acquisition, and Writing – original draft. All authors read and approved the final version of the manuscript.

**Competing Interest:** The authors declare that they have no conflict of interest.

## References

1. F. I. Baharuddin and R. Hidayat, Konsensus Dan Kontradiksi Tes Kraepelin Pada Kalangan Psikolog, Universitas Gadjah Mada, 2019
2. E. J. Engstrom, Tempering Madness: Emil Kraepelin's Research on Affective Disorders, Osiris **31**, 163 (2016)
3. S. Sukadji, Kecepatan Dan Ketelitian Kerja Yang, J. Psikol. **20**, (1993)
4. D. R. Eignor, The Standards for Educational and Psychological Testing, in APA Handbook of Testing and Assessment in Psychology, Vol. **1**: Test Theory and Testing and Assessment in Industrial and Organizational Psychology (American Psychological Association, Washington, DC, US, 2013), pp. 245–250 https://doi.org/10.1037/14047-013
5. A. Adhha and D. F. Darusmin, Asesmen Potential Review pada Karyawan Rumah Sakit Islam YY Pekanbaru, Psyche 165 J. **14**, 247 (2021) https://doi.org/10.35134/jpsy165.v14i2.107
6. E. Gallardo-Gallardo, N. Dries, and T. F. González-Cruz, What is the meaning of 'talent' in the world of work?, Hum. Resour. Manag. Rev. **23**, 290 (2013) https://doi.org/10.1016/j.hrmr.2013.05.002
7. N. Indriani, D. Sakethi, and A. Syarif, Pengembangan Simulasi "Stress Test" Menggunakan Tes Kraepelin Pada Tes Psikologi, J. Pepadun **1**, 63 (2020) https://doi.org/10.23960/pepadun.v1i1.11
8. Q. Nada, M. Arhami, and Z. K. Simbolon, Pengukuran Aptitude dengan Uji Kraepelin Menggunakan Metode Linear Congruential Method (LCM), J. Teknol. **22**, 1 (2022) https://doi.org/10.30811/teknologi.v22i1.2418
9. W. Yudiana, A. B. Adjie, N. A. Assa, and A. K. Putra, Psychometric Properties of a Newly Online Work Aptitude Test: Faxtor's Endurance and Speed Test, J. Educ. Health Community Psychol. **1**, 33 (2025) https://doi.org/10.12928/jehcp.v1i1.25292
10. I. M. Febriawan, A. A. Fauzan, C. Adelina, H. N. Afifah, R. Pari, and A. Januarsjaf, Qutest construction and psychometric evaluation as test of attention and willpower for employee selection screening, JP3I J. Pengukuran Psikol. Dan Pendidik. Indones. **11**, 55 (2022)
11. M. C. Rodriguez, Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research, Educ. Meas. Issues Pract. **24**, 3 (2005) https://doi.org/10.1111/j.1745-3992.2005.00006.x
12. M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang, Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review, Rev. Educ. Res. **87**, 1082 (2017) https://doi.org/10.3102/0034654317726529

13. J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, Eighth edition (Cengage, Andover, Hampshire, 2019)

14. M. Tarrant and J. Ware, A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments, Nurse Educ. Today **30**, 539 (2010) https://doi.org/10.1016/j.nedt.2009.11.002

15. B. Esmaeeli, E. Esmaeili Shandiz, S. Norooziasl, H. Shojaei, A. Pasandideh, R. Khoshkholgh, B. Fazli, and F. Barkhordari Ahmadi, The Optimal Number of Choices in Multiple-Choice Tests: A Systematic Review, Med. Educ. Bull. **2**, 253 (2021) https://doi.org/10.22034/meb.2021.311998.1031