Jurnal Penelitian dan Pengukuran Psikologi
Volume 15, Nomor 01, April 2026 http://doi.org/10.21009/JPPP.151.10

PSYCHOMETRIC RE-EVALUATION OF IST IN INDONESIAN DEFENSE SELECTION

Sukmo Gunardi[1]
[1]*Prodi Psikologi Industri, Fakultas Teknik, Universitas Nurtanio Bandung, Bandung, Indonesia*

Email: *sukmo.gunardi@gmail.com*

p-ISSN: 2337-4845
e-ISSN: 2620-7486

Received	*Revised*	*Accepted*	*Published*
05 Januari 2026	*20 Februari 2026*	*29 April 2026*	*30 April 2026*

Abstract

This study re-evaluated the psychometric characteristics of the Intelligenz-Struktur-Test (IST) in an Indonesian defense-selection context using de-identified archival testing data. The study examined four observed domains, namely verbal (WA), analytical (AN), numerical (ZR), and figural (FA), based on a cleaned analytic sample with complete data across all domains. The final sample consisted of 2,222 participants. Responses were scored dichotomously, and the analyses focused on descriptive statistics, internal consistency, standard error of measurement, gender comparisons, and age correlations. Mean scores were 10.50 for WA, 10.38 for AN, 7.88 for ZR, and 8.77 for FA. Reliability estimates varied across domains, with AN showing the strongest internal consistency and FA the weakest. Gender differences were statistically significant for WA, AN, and ZR, with females scoring higher than males, although effect sizes were small; no significant gender difference was found for FA. Age was positively associated with all four domains. The findings indicate that the IST yields interpretable cognitive domain scores in this defense-related sample, but they do not justify broad claims of fairness or definitive structural validity. The study supports continued local psychometric evaluation, transparent data screening, subgroup analysis, and stronger item-level and latent-structure testing in future research

Keywords: ist, psychometrics, defense selection, gender differences, reliability

Abstrak

Penelitian ini menelaah kembali karakteristik psikometrik Intelligenz-Struktur-Test (IST) dalam konteks seleksi personel pertahanan di Indonesia menggunakan data arsip asesmen yang telah dide-identifikasi. Penelitian mencakup empat domain teramati, yaitu verbal (WA), analitik (AN), numerik (ZR), dan figural (FA), berdasarkan sampel analitik yang telah dibersihkan dan memiliki data lengkap pada seluruh domain. Sampel akhir terdiri atas 2.222 partisipan. Respons diskor secara dikotomis, dan analisis difokuskan pada statistik deskriptif, konsistensi internal, standard error of measurement, perbandingan berdasarkan gender, dan korelasi dengan usia. Rerata skor adalah 10,50 untuk WA, 10,38 untuk AN, 7,88 untuk ZR, dan 8,77 untuk FA. Estimasi reliabilitas bervariasi antardomain, dengan AN menunjukkan konsistensi internal paling kuat dan FA paling lemah. Perbedaan gender signifikan secara statistik ditemukan pada WA, AN, dan ZR, dengan perempuan memperoleh skor lebih tinggi daripada laki-laki, meskipun ukuran efeknya kecil; tidak ditemukan perbedaan gender yang signifikan pada FA. Usia berkorelasi positif dengan keempat domain. Temuan menunjukkan bahwa IST menghasilkan skor domain kognitif yang dapat diinterpretasikan, tetapi belum cukup untuk mendukung klaim luas mengenai fairness atau validitas struktural yang definitif. Penelitian ini mendukung perlunya evaluasi psikometrik lokal yang berkelanjutan, pembersihan data yang transparan, analisis subkelompok, serta pengujian tingkat item dan struktur laten yang lebih kuat di masa mendatang.

Kata kunci: ist, psikometrik, seleksi pertahanan, perbedaan gender, reliabilitas

1. Introduction

Cognitive ability testing remains central in educational, organizational, and military selection because it provides standardized information relevant to learning capacity, reasoning performance, and adaptation to complex task environments. In defense-related settings, these concerns become especially important because decisions made from test scores may influence recruitment, training allocation, and institutional perceptions of fairness. For that reason, cognitive tests used in high-stakes settings should be supported not only by tradition or operational familiarity, but also by evidence regarding score reliability, subgroup comparability, and internal structure.

The Intelligenz-Struktur-Test (IST), originally developed within the Amthauer tradition, is widely recognized as a multidimensional intelligence test intended to assess several domains of cognitive performance. In broad theoretical terms, the IST is relevant to longstanding debates in intelligence research concerning the relation between general cognitive ability and more specific domains. Spearman's concept of a general factor suggests that common variance across cognitive tasks reflects a broad underlying capability, whereas multidimensional and hierarchical traditions, including fluid–crystallized and CHC-oriented perspectives, emphasize that cognitive performance is also differentiated across verbal, numerical, analytical, and figural domains. From this standpoint, the IST is psychometrically important because it may capture both shared and domain-specific variance, and recent open-access adaptation work has continued to document its relevance in verbal, numerical, and figural assessment contexts (Jokste et al., 2024).

Despite the longstanding use of intelligence testing internationally, empirical evaluation of the IST in Indonesia remains limited, especially in restricted-access operational settings such as defense-related selection. This gap matters for at least three reasons. First, psychometric properties established in one cultural or institutional setting cannot be assumed to generalize automatically to another. Second, score interpretation requires transparency about the sample and data processing procedures from which psychometric evidence is derived. Third, fairness claims in high-stakes testing should be grounded in evidence rather than assumed from routine use. In the Indonesian context, open-access work on IST-related item quality and DIF has underscored the importance of local psychometric evaluation rather than direct assumption from external settings (Tarigan & Fadillah, 2022). The initial version of this manuscript attempted to address this need by integrating classical reliability analysis, confirmatory factor analysis, and network analysis. However, the earlier paper was too brief, underdeveloped theoretically, and insufficiently transparent regarding sample construction, analytic strategy, and interpretation of findings. Re-examination of the archival raw dataset also showed that several empirical statements in the earlier draft were not adequately aligned with the cleaned participant-level data used for defensible reporting. The present revised manuscript therefore adopts a more cautious and data-grounded approach. It also recognizes that network-based interpretation must be linked to estimation accuracy and stability considerations rather than substantive claims alone (Epskamp et al., 2018).

The study re-evaluates the psychometric characteristics of the IST in an Indonesian defense-selection context using de-identified archival operational data. The focus is on four observed domains—verbal (WA), analytical (AN), numerical (ZR), and figural (FA)—and on several foundational questions: the consistency of domain scores, descriptive differences across cohorts, mean differences across gender, and associations between age and the observed domains.

Accordingly, the revised study has five aims: to report the final analytic sample and screening process transparently; to estimate domain-level internal consistency for WA, AN, ZR, and FA; to estimate the standard error of measurement for each domain; to examine gender differences using both significance testing and effect sizes; and to estimate associations between age and each observed domain.

In this revised framing, the manuscript no longer treats fairness as established merely because group means may or may not differ significantly. Nor does it infer theoretical centrality from reliability coefficients alone. Instead, it advances a more modest but more defensible claim: in this defense-related archival sample, the IST yields interpretable domain scores, but stronger conclusions about fairness, invariance, and structural superiority require continued local validation and more rigorous follow-up analyses, including formal invariance procedures when such analyses are available (Rodríguez-Cancino & Concha-Salgado, 2023).

2. Research Method

Participants

Participants were drawn from archival operational testing records collected in defense-related selection contexts during the period in which the author served in the relevant institution. The available database consisted of multiple test-administration cohorts distributed across four IST domains: verbal (WA), analytical (AN), numerical (ZR), and figural (FA). Because the present study aimed to examine the psychometric pattern of the full four-domain battery, the analytic sample was restricted to participants with data available across all four domains.

To construct the final analytic dataset, records were matched across WA, AN, ZR, and FA using a composite identifier consisting of participant number, gender, and age. This strategy was adopted because participant number alone was not sufficiently unique across administration modules. Cases were excluded if they contained invalid gender coding, missing or implausible age information, more than 25% missing item responses within any subtest, or ambiguous duplicate identifiers that could not be aligned confidently across sheets. After screening, the final analytic sample consisted of 2,222 participants.

Of the final sample, 1,543 were male and 679 were female. Participants ranged in age from 15 to 56 years. The final sample was drawn from three complete operational cohorts: I 2025, II 2025, and II 2025.

Research Design

The instrument under study was the Intelligenz-Struktur-Test (IST), administered in four broad domains in the archived dataset: verbal (WA), analytical (AN), numerical (ZR), and figural (FA). Each domain contained 20 items in the operational records used for this study. Responses were recorded in multiple-choice format and scored dichotomously for the present analyses, with keyed responses coded as 1 and non-keyed responses coded as 0. Domain scores were represented by the total raw score for each subtest.

For descriptive purposes, WA was treated as a verbal domain, AN as an analytical domain, ZR as a numerical domain, and FA as a figural domain. These labels are used in the present paper as practical descriptive categories for the archived operational dataset. More specific theoretical mapping of these domains to broader intelligence models should be made cautiously and ideally supported by explicit structural evidence from the same cleaned analytic sample (Amthauer et al., 2001; Carroll, 1993).

Procedure and Ethical Considerations

This study used secondary analysis of de-identified archival testing data originating from operational defense-selection contexts. Before analysis, the dataset used for research was stripped of direct personal identifiers, and analyses were conducted only on anonymized records.

For ethical clarity, the data are best described as archival institutional records collected in the course of operational testing during the period in which the author held professional responsibility, rather than as private personal records. At the time of revision, institutional authorization regarding the research use of the archived de-identified dataset was being sought from the relevant former institution. In the final submitted version, this section should be aligned explicitly with the resulting authorization, ethics approval, waiver, or exemption status, as applicable.

Analiytic Strategy

The analytic strategy proceeded in four stages. First, descriptive statistics were calculated for age, gender composition, and total scores on WA, AN, ZR, and FA. Second, internal consistency was estimated for each domain using Cronbach's alpha and McDonald's omega. Alpha was retained because it remains a widely recognized conventional reliability coefficient, whereas omega was included as a complementary estimate that is more appropriate under less restrictive assumptions (Malkewitz et al., 2023).

Third, independent-samples t-tests were conducted to compare male and female participants on each domain score, and effect sizes were expressed using Cohen's d. In the present study, these t-tests were used only to evaluate observed-score mean differences, not to establish the absence of differential item functioning or item bias. Accordingly, the results should not be interpreted as evidence that no DIF exists; stronger conclusions would require dedicated DIF procedures such as Mantel–Haenszel, IRT-based DIF, or formal multi-group invariance analysis (Meredith, 1993; Meade et al., 2008; Rutkowski & Svetina, 2017).

Fourth, Pearson correlations were computed to examine associations between age and each domain score. Comparability across gender and age was interpreted descriptively at the observed-score level and was not treated as formal measurement invariance, which would require structured multigroup modeling procedures (Rodríguez-Cancino & Concha-Salgado, 2023).

The earlier manuscript also reported confirmatory factor analysis, bifactor interpretation, network psychometrics, and gender-related structural comparisons. These analyses remain potentially relevant to the broader psychometric evaluation of the IST, but the present revision does not present them as definitive findings because they were not recalculated directly from the cleaned participant-level dataset used here. Accordingly, the current manuscript prioritizes descriptive and comparative findings that can be reported directly and defensibly from the re-audited archival data, while treating broader structural claims as provisional until reanalysis is completed on the same analytic sample.

3. Result and Discussion

Descriptive Statistics and Sample Characteristics

The final analytic sample consisted of 2,222 participants with complete data across the four IST domains. Participants were retained only when records could be matched confidently across all four domains and met the prespecified screening criteria. Mean total scores were 10.50 for WA, 10.38 for AN, 7.88 for ZR, and 8.77 for FA. The descriptive pattern indicates that average performance was highest in the verbal domain and lowest in the numerical domain in the pooled sample.

The sample was not drawn from a single homogeneous intake. Rather, it combined three operational cohorts with different score profiles. In general, the I 2025 cohort obtained higher mean scores across domains than the II 2025 cohort, with III 2025 generally falling between those groups. Accordingly, the pooled descriptive statistics should be interpreted as reflecting a heterogeneous defense-related archival sample rather than a single uniform testing cohort.

Table 1. Demographic characteristics of the final analytic sample

Variable	Value
Final N	2,222
Male	1,543
Female	679
Age Mean (SD)	24.10 (7.24)
Age Median	21
Age Range	15–56

Table 2. Descriptive statistics of IST domain scores

Domain	Mean	SD
WA	10.50	3.12
AN	10.38	3.82
ZR	7.88	3.22
FA	8.77	2.95

Reliability of IST Domain Scores

Using the cleaned archival dataset and dichotomous item scoring, the four IST domains showed varying levels of internal consistency. The estimated coefficients were as follows: WA alpha = .64, omega = .68; AN alpha = .74, omega = .74; ZR alpha = .62, omega = .60; and FA alpha = .52, omega = .51. These findings indicate that the analytical domain showed the strongest internal consistency in the present dataset, whereas the figural domain showed the weakest. The verbal and numerical domains fell in the low-to-moderate range.

The estimated standard errors of measurement were 1.87 for WA, 1.95 for AN, 1.98 for ZR, and 2.04 for FA. These values indicate a nontrivial range of observed-score error, particularly in the lower-reliability domains, and reinforce the need to interpret observed differences with caution in line with contemporary reliability guidance (Malkewitz et al., 2023).

These coefficients differ from those reported in the earlier draft of the manuscript. In the present revision, the values derived directly from the re-audited raw dataset are prioritized. Their interpretation remains limited to score consistency and measurement precision and should not be treated as evidence of theoretical centrality, fairness, or broader construct superiority.

Table 3. Reliability estimates and standard errors of measurement

Domain	Cronbach's α	McDonald's ω	SEM
WA	.64	.68	1.87
AN	.74	.74	1.95
ZR	.62	.60	1.98
FA	.52	.51	2.04

Gender Differences

Gender differences were examined using independent-samples t-tests. In contrast to the earlier version of the manuscript, the cleaned archival dataset indicated statistically significant gender differences in three of the four domains. Female participants scored significantly higher than male participants on WA, AN, and ZR, whereas no statistically significant gender difference was observed on FA.

These results directly address the reviewer's request to report t-values in addition to p-values. More importantly, they require correction of the earlier statement that no significant gender mean differences were found. A more accurate conclusion is that statistically significant gender differences emerged in three domains, although the effect sizes were small by conventional standards. Because the present comparison was conducted at the observed-score level, it should not be interpreted as a test of DIF or proof that the instrument is free from item-level bias.

Table 4. Gender differences in IST domain scores

Domain	Male Mean	Female Mean	t	p	Cohen's d
WA	10.21	11.15	-7.46	< .001	-0.31
AN	10.00	11.23	-7.55	< .001	-0.33
ZR	7.54	8.67	-7.96	< .001	-0.36
FA	8.83	8.63	1.47	.143	0.07

Age Correlations

Pearson correlations were used to examine the associations between age and the four IST domain scores. In contrast to the pattern stated in the earlier draft, the cleaned dataset showed positive associations between age and all four domains: WA r = .322, AN r = .358, ZR r = .273, and FA r = .138, all p < .001.

Thus, the earlier claim that age was negatively associated with numerical reasoning was not supported by the re-audited archival dataset. Instead, older participants tended to obtain somewhat higher scores across all domains, with the strongest associations observed in analytical and verbal performance. Because these are observational correlations within an operational sample, they should be interpreted cautiously and not as direct developmental conclusions in a strict lifespan sense. More generally, age-related cognitive patterns are known to vary across domains and contexts, which supports a cautious interpretation of cross-sectional associations such as those reported here (Murman, 2015).

Table 5. Correlations between age and IST domain scores

Domain	r	p
WA	.322	< .001
AN	.358	< .001
ZR	.273	< .001
FA	.138	< .001

Discussion

The present study re-evaluated the psychometric characteristics of the IST in an Indonesian defense-selection context using de-identified archival operational testing data. The revised manuscript now reports participant characteristics, demographic composition, gender comparisons, age associations, and reliability estimates with substantially greater transparency. Overall, the findings suggest that the IST can yield interpretable domain scores in this context, but they also indicate that several claims in the earlier draft were overstated and required correction.

The revised reporting of the sample addresses a major weakness noted by the reviewer. The earlier manuscript did not clearly specify the number of participants or how the final dataset was constructed. The present revision clarifies that the final analytic sample consisted of participants with complete and matchable data across all four domains drawn from more than one operational context.

The reliability findings present a more mixed picture than the earlier draft suggested. The analytical domain showed the strongest internal consistency, whereas the figural domain showed the weakest. This pattern reinforces the need for domain-specific interpretation rather than blanket endorsement of the battery as a whole. Reliability coefficients indicate score consistency, not theoretical primacy (Cronbach, 1951; McDonald, 1999).

The gender results require substantial revision of the manuscript's fairness narrative. The re-audited archival dataset does not support the earlier statement that no significant gender mean differences were found. Instead, statistically significant differences were observed in verbal, analytical, and numerical scores, all favoring female participants, while figural scores did not differ significantly. At the same time, the effect sizes were small, suggesting that the differences, although statistically robust in a large sample, were limited in practical magnitude. This revised interpretation is scientifically stronger than the earlier all-or-none conclusion and aligns better with the reviewer’s request for fuller statistical reporting.

More broadly, the present findings underscore that fairness cannot be inferred from mean comparisons alone. The absence of large group differences does not establish measurement equivalence, and the presence of small differences does not automatically imply bias. Stronger claims regarding fairness would require item-level differential item functioning analyses and formal measurement invariance testing. The revised manuscript therefore narrows its claims to subgroup comparability at the observed-score level rather than definitive fairness. This is consistent with open-access methodological literature showing that DIF analysis is essential when fairness is at issue and that total-score comparisons alone can be misleading (Martinková et al., 2017; Liu et al., 2019).

The age findings also changed substantially relative to the earlier draft. Rather than showing a negative association with numerical reasoning, age was positively associated with all four domains in the cleaned dataset. The strongest correlations were observed for analytical and verbal scores, followed by numerical and then figural scores. Several interpretations are possible. Older participants in this archival sample may also have differed in educational background, training exposure, or selection pathway. Thus, age may partly index cohort composition rather than pure developmental change. For that reason, these results should not be overinterpreted as contradicting broader fluid–crystallized distinctions. Instead, they indicate that within this specific operational sample, older participants tended to obtain somewhat higher observed scores across domains, while broader age-related cognition findings remain domain-sensitive (Murman, 2015).

The revised manuscript also adopts a more cautious stance toward structural claims. Broader conclusions concerning bifactor superiority, network centrality, and factorial invariance should not be presented as settled facts unless they can be linked directly to the same cleaned analytic sample used for the descriptive and subgroup analyses reported here. In practical terms, the present paper should be read as a psychometric re-evaluation rather than as a final structural validation. This caution is consistent with open-access guidance on network accuracy

and with recent work emphasizing structured invariance testing across sex and age when fairness claims are made (Epskamp et al., 2018; Rodríguez-Cancino & Concha-Salgado, 2023).

The study has several limitations. Although the sample is sizable, it is archival and heterogeneous, which complicates simple causal or developmental interpretation. The present revision does not rerun the full confirmatory factor, network, or invariance analyses directly on the same cleaned analytic sample. The study also does not yet include differential item functioning analyses, which are essential for a stronger fairness evaluation.

Finally, institutional authorization regarding research use of the de-identified archival data should be finalized before formal journal submission. From a psychometric fairness perspective, future item-level DIF work remains particularly important (Martinková et al., 2017; Liu et al., 2019).

Future research should rerun confirmatory factor analysis and formal model comparison directly on the cleaned participant-level dataset used in the revised manuscript, conduct measurement invariance testing and item-level DIF analysis, compare the IST with other cognitive measures used in Indonesia, and validate the instrument across additional operational cohorts. Available open-access IST-related work in Indonesia and abroad suggests that such follow-up analyses are both feasible and worthwhile for building stronger local evidence (Jokste et al., 2024; Tarigan & Fadillah, 2022).

4. Conclusion

This study showed that the IST can produce interpretable scores across verbal, analytical, numerical, and figural domains in an Indonesian defense-selection context. However, the strength of the evidence varies across psychometric aspects, so the findings support cautious interpretation rather than broad endorsement.

The revised manuscript also shows that demographic comparisons and score consistency must be interpreted carefully. The results do not support broad claims about the absence of demographic differences or definitive structural validity.

Overall, the study supports continued local validation, transparent reporting, and stronger fairness evaluation. The main contribution of this paper is to present a more accurate and methodologically defensible psychometric re-evaluation of the IST in this context.

5. References

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. Behavior Research Methods, 50(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Jokste, I., Trups-Kalne, I., Lubenko, J., Millere, I., & Kolesnikova, J. (2024). Adaptation of the Intelligence Structure Test, Latvian version: Psychometric properties. Frontiers in Psychology, 15, Article 1319983. https://doi.org/10.3389/fpsyg.2024.1319983

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. Frontiers in Psychology, 10, 1137. https://doi.org/10.3389/fpsyg.2019.01137

Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's alpha, McDonald's omega, and the greatest lower bound. MethodsX, 10, 102032. https://doi.org/10.1016/j.mex.2023.102032

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. CBE—Life Sciences Education, 16(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Murman, D. L. (2015). The impact of age on cognition. Seminars in Hearing, 36(3), 111–121. https://doi.org/10.1055/s-0035-1555115

Rodríguez-Cancino, M., & Concha-Salgado, A. (2023). WISC-V measurement invariance according to sex and age: Advancing the understanding of intergroup differences in cognitive performance. Journal of Intelligence, 11(9), 180. https://doi.org/10.3390/jintelligence11090180

Tarigan, M., & Fadillah, F. (2022). Quality analysis of Intelligence Structure Test 2000 Revision (IST 2000R) items in Indonesian. Psyche: Jurnal Psikologi, 9(1), 103–116. https://doi.org/10.15575/psy.v9i1.14977