



ANALYSIS OF SCORE CHANGES AND CEILING EFFECTS ON THE ASN INTEGRITY TRAINING TEST: IMPLICATIONS FOR THE MINISTRY OF TRANSPORTATION'S LEARNING EVALUATION

Rima Febrina¹, Afsuni Wijananto², Dyan Apriyani³, Siswantoyo⁴

Universitas Negeri Yogyakarta

E-mail: rimafebrina.2026@student.uny.ac.id**ABSTRACT**

Civil servant integrity training requires evaluation frameworks that distinguish recorded cognitive test performance from ethical workplace behavior. This study examines score changes and ceiling effects in the ASN Integrity training program and their implications for learning evaluation within Indonesia's Ministry of Transportation. A retrospective quantitative design analyzed Learning Management System (LMS) records from March 18 to September 3, 2026. The primary analysis involved 24,174 paired participant records after excluding 38 invalid time-sequence pairs. Data were processed in R using descriptive statistics, paired t-tests, and McNemar's test. Mean scores increased significantly from 90.25 to 95.27, representing a 5.02-point gain (95% CI: 4.85–5.19; $p < .001$). However, the median individual score difference was 0.00: 41.49% of participants improved, 41.10% remained unchanged, and 17.41% declined. The proportion of maximum scores (100) expanded from 43.18% at pretest to 57.30% at posttest ($p < .001$). High upper-end score concentration indicates a substantial ceiling effect that constrains the measurement of true learning gains. Furthermore, the lack of a control group and test-equating evidence precludes causal conclusions about training efficacy or the development of behavioral integrity. These findings highlight the need for Kemenhub Corporate University to refine question banks, implement Situational Judgment Tests (SJTs), and integrate cognitive assessments with workplace behavioral observations and Integrity Zone frameworks. Keywords: civil servant integrity; training evaluation; score change; ceiling effect; corporate university.

INTRODUCTION

Integrity is a crucial aspect in the performance of civil servant (ASN) duties, as their decisions are directly related to the exercise of authority and service to the public. Developing an integrity character must create consistency between knowledge of values, ethical considerations, and actions within the work environment. Regulation of the Minister of Administrative and Bureaucratic Reform Number 38 of 2017 categorizes integrity as a managerial competency encompassing honesty, responsibility for decisions, and behavior consistent with organizational ethics. Indicators include individual behavior and the creation of an ethical work environment (Ministry of Administrative and Bureaucratic Reform, 2017). Therefore, evaluation of integrity learning must consider the gap between mastery of the material and its application.

At the organizational level, the development of an Integrity Zone links improved governance with a clean, accountable, and quality bureaucracy in service delivery. Regulation of the Minister of Administrative and Bureaucratic Reform Number 90 of 2021, amended by Regulation Number 5 of 2024, provides the framework for this

change (Ministry of Administrative and Bureaucratic Reform, 2021, 2024). Integrity training can be part of human resource development and work culture. However, successful training alone is not sufficient to demonstrate consistent ASN behavior or the success of a work unit in developing an Integrity Zone. Accurate evaluation information is essential for the implementation of training achievements in organizational practice.

Public sector studies demonstrate the need to connect training to the workplace. Perlman et al. (2021) found that government ethics training places more emphasis on compliance than on values and the public interest. Demir et al. (2023) outlined policy support, enforcement, incentives, and ethics knowledge as part of an organization's ethics infrastructure. Both studies reinforce the importance of placing integrity learning within the context of character development and governance and distinguishing between evidence of knowledge and evidence of behavior.

At the Ministry of Transportation, strengthening evaluation is related to the development of the Ministry of Transportation Corporate University (Corpu). Article 49 of Law Number 20 of 2023 directs competency development through continuous learning integrated with the work and management of civil servants (ASN). The Corpu's implementation function at the Ministry of Transportation is stated in Article 726 of PM 4 of 2025. The Ministry of Transportation Corpu Blueprint also emphasizes the use of learning analytics and evaluation, encompassing implementation and organizational impact. This institutional direction requires the use of learning management system (LMS) data accompanied by adequate measurement interpretation. The ASN Integrity training provides pretest and posttest data that can be used to analyze score changes. However, average increases alone are not sufficient to clearly analyze variations in individual changes in training participants, especially when many participants have reached the upper limit of the scale on scores obtained from pre-posttest data. Staus et al. (2021) explain that ceiling effects can limit measurable changes in evaluations.

The evaluation issue addressed here is how to interpret changes in integrity training scores when room for improvement is limited and test package equivalency information is unavailable. This focus provides an empirical basis for improving evaluation, without considering test scores as a comprehensive measure of integrity character. The contribution of this study lies in the joint analysis of mean changes, maximum score transitions, and completeness of paired data sets on a large scale. This approach allows for distinguishing test achievement in terms of measurability and the need for follow-up. Changes in test results should be considered as one source of evidence in evaluating integrity learning, while implementation in tasks requires a separate measure. This study aims to analyze changes in pretest-posttest scores, changes in the proportion of maximum scores, and their implications for evaluating Kemenhub Corpu learning. Research questions include: what is the direction and magnitude of score changes; how does the status of maximum scores change within the same participant; and what are the implications of these patterns for follow-up integrity learning. Two statistical hypotheses are tested: no difference in mean pre-posttest scores and no difference in the proportion of paired maximum scores, each against a two-way alternative. Organizational implications are discussed based on the findings and literature, rather than being tested as behavioral changes in this study.

METHOD

This study employed a quantitative approach through a retrospective analysis of secondary data with a paired pre-posttest structure. The data source was the LMS export of the ASN Berintegritas training attended by participants within the Ministry of Transportation. The unit of analysis was the participant with a recorded score pair. All recordings that met the criteria were used without resampling, and the scope was limited to the participants in the available data export, not all ASNs at the Ministry of Transportation. The data analysis period was March 18 to September 3, 2026.

The data preparation process included separating participant rows from the overall summary, filtering for completion status, matching participant identities across pretests and posttests, and checking completion times. Of the 25,029 participants who completed the pretest, 24,212 had a completed posttest, and 817 did not have a posttest match. Thirty-eight pairs were excluded from the main analysis because the posttest start time preceded the pretest finish time. This time criterion served as an operational constraint to maintain the measurement sequence.

The primary analysis used 24,174 complete pairs; unavailable posttest scores were not imputed. The primary variable was the total pretest and posttest scores on a 0–100 scale, using the Grade/100.00 value in the LMS export. Difference was defined as the posttest score minus the pretest score and categorized as increased, constant, or decreased. Maximum score status was divided into two categories: 100 and less than 100. Maximum score concentration was used to examine indications of ceiling effects, without establishing a universal threshold or interpreting 100 as perfect integrity. The item bank and item identities were not available; the administrators stated that the order and selection of items were randomized and the package may change in subsequent experiments. Therefore, the analysis was limited to recorded scores, with no IRT estimates or claims of package equivalence.

Analyses were performed using R through RStudio. Descriptive statistics included mean, standard deviation, median, interquartile range, minimum-maximum value, and frequency of changes and maximum scores. A two-tailed paired t-test tested the mean difference against zero, along with 95% confidence intervals. McNemar's test with continuity correction tested the equality of proportions of maximum scores based on two opposing types of changes. The significance level was set at 0.05. Analyses based on initial scores less than 100 and equal to 100 were exploratory. Descriptive comparisons of initial scores of paired participants with those who only took the pretest were used to assess the possibility of selection due to incomplete pairs. Processing was carried out in stages to allow for tracking of changes in the number of recordings and variable definitions.

The `pre_raw` and `post_raw` objects store the imported data, the pairs contain the matching results, and `data_analysis` is used for the main calculations. The entire row represents the system summary and is therefore not treated as a participant. The completed status is used to distinguish completed tests from those in progress. This separation prevents summary rows and incomplete tests from influencing the average or number of participants.

Table 1. Stages of data processing and analysis using RStudio

Stage	Process and output	Purpose of the examination
1. Check exports	Reviewing column names, score and time formats, and table(Status, useNA = "ifany") in pre_raw and post_raw.	Differentiate between participant records, incomplete status, and summary lines.
2. Filter status	Maintain Finished; 25,029 pretests and 24,212 posttests were completed.	Using records with the same processing status.
3. Pairing the test	Matching participant identities: 24,212 pairs, 817 only pretest, and 0 only posttest.	Compare scores on the same participant and record incomplete pairs.
4. Check the time	Reading English dates in AM/PM format; calculating the distance from the start of the posttest to the end of the pretest. A total of 38 negatively spaced pairs were excluded.	Maintaining operational measurement order; remaining 24,174 pairs.
5. Forming variables	Using pre_score and post_score; generate difference, direction of change, max_pre, and max_post.	Define the direction of the difference as posttest minus pretest and a maximum limit of 100.
6. Compile a summary	Produces descriptive tables, changes, initial score groups, participation, as well as histograms and maximum proportion graphs.	Reading distribution, consistency of quantity, and concentration at the upper limit.
7. Testing changes	t.test with paired = TRUE; mcnemar.test with correct = TRUE on transition table without total rows/columns.	Testing the mean difference and maximum status change while maintaining the pair structure.
8. Save output	Stores object checkpoints in RDS, CSV tables, test output, and analysis result images.	Provides a processing trail for review and manuscript preparation.

Source: processed author data.

The magnitude of change is also expressed as Cohen's *d*, standardized by the standard deviation of the pairwise differences (*dz*). The value is derived from the *t* statistic divided by the square root of the number of pairs. This measure complements the difference in points and its confidence interval; *dz* is not interpreted as a percentage increase, trait change, or causal effect of the training. Test results are reported along with the distribution of the change because statistical significance in large numbers of participants does not demonstrate practical significance. With a large number of pairs, a *t* test is used to estimate the observed mean score difference; the distribution of the difference is still reported through the median, direction of change, and histogram because the nesting of scores limits the interpretation of the mean. Inference assumes independence among participants and does not account for work unit clustering. The *p* value is not treated as a basis for probabilistic generalizations to

the entire ASN or as evidence of causality. The absence of a comparison group and evidence of package equivalence limits conclusions to the score changes in the analyzed data.

RESULTS AND DISCUSSION

The main analysis included 24,174 participants with test pairs that met the criteria. Table 2 shows higher baseline scores and a narrower posttest distribution.

Table 2. Descriptive statistics of pretest and posttest scores

Statistics	Pretest	Posttest
Number of participants	24,174	24,174
Mean $\pm$ standard deviation	90.25 $\pm$ 13.83	95.27 $\pm$ 6.74
Median	93.33	100.00
Interquartile range	13.33	6.67
Minimum–maximum	0–100	0–100

Source: processed by the author.

The mean score increased from 90.25 to 95.27, while the distribution of posttest scores was narrower (Table 2). A paired t-test revealed a mean difference of 5.02 points (95% CI: 4.85–5.19; $t(24,173) = 57.515$; $p < 0.001$). Thus, the null hypothesis of no difference in mean scores was rejected. The standardized measure of change was $d_z \approx 0.37$, calculated from $t = 57.515$ and 24,174 pairs. This means the mean change was approximately 0.37 standard deviations of the individual differences. This presentation helps distinguish evidence of change from magnitude of change. The 5.02-point score remains a key measure of readability on the 0–100 test scale; there is no minimum meaningful change threshold for task performance, so practical success categories are not defined solely by effect size.

These changes did not occur across all participants. 10,031 participants (41.49%) experienced an increase, 9,935 (41.10%) achieved the same score, and 4,208 (17.41%) experienced a decrease. The median score difference was 0.00, with a standard deviation of 13.57 and a range of –86.67 to 100.00. These findings indicate that the average increase occurred alongside considerable individual variation in change. Pair completion reached 96.74% of the 25,029 completed pretests; 817 participants (3.26%) did not have a completed posttest in export. Table 3 shows that the pretest mean for the unpaired group was 5.16 points lower, with a wider spread than for the paired group. The median for both groups was the same, at 93.33. Thus, the similarity in medians does not imply identical initial score profiles for the two groups.

Table 3. Pretest profile according to completeness of test pairs

Group	n	Average	Elementary School	Median	Score 100 (%)
Have a test partner	24,212	90.26	13.83	93.33	43.21
Only pretest completed	817	85.10	19.57	93.33	36.23

Note: processed by the author.

This difference suggests that pair completeness is related to the initial score profile in the data, but does not explain why the posttest is unavailable. Because posttest scores were not available for this group, the direction of the unobserved change cannot be determined. A total of 10,438 participants achieved a score of 100 on the pretest and 13,851 on the posttest. The proportion increased from 43.18% to 57.30%, or 14.12

percentage points. Figure 1 shows the proportion of participants with maximum scores on the pretest and posttest (n = 24,174).

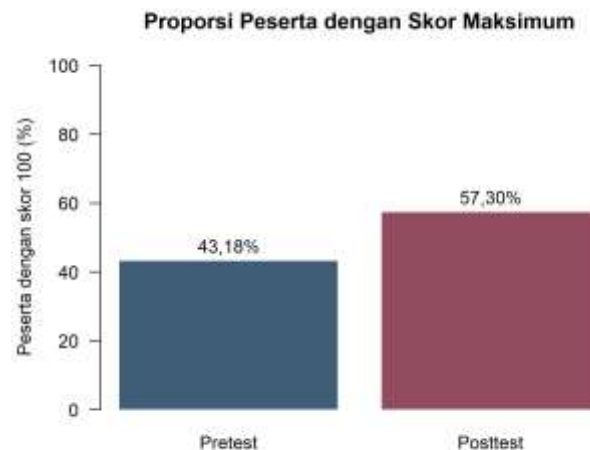


Figure 1. Proportion of participants with maximum scores on the pretest and posttest (n = 24,174).

Table 4. Transition of maximum score status in the same participant

Pretest score	Posttest <100	Posttest 100	Amount
<100	7,447	6,289	13,736
100	2,876	7,562	10,438
Amount	10,323	13,851	24,174

Source: LMS export processing results with R, 2026. The numbers in the cells indicate the number of participants.

6,289 participants moved from a score <100 to 100, compared to the 2,876 participants who moved in the opposite direction (Table 4). McNemar's test with continuity correction revealed a significant difference in the proportion of maximum scores ($\chi^2(1) = 1,270.2$; $p < 0.001$). This test assesses the change in the maximum score status, not the magnitude of the increase in competence. The two cells that changed status included 9,165 participants: 6,289 moved toward the maximum score and 2,876 left it. The difference between the two, 3,413 participants, corresponds to the increase in the number of people with a score of 100. Conversely, 7,562 participants remained at 100 and 7,447 remained below it. The latter group can experience increases, decreases, or a constant score as long as it does not cross the 100 threshold. Therefore, the status transition table complements but does not replace the tabulation of the direction of score change.

In the 13,736 participants with a pretest score <100, the mean score changed from 82.85 to 93.60, a mean difference of 10.76 points; 73.03% experienced an increase. In the 10,438 participants with a pretest score of 100, the posttest mean was 97.47, a difference of -2.53 points; 72.45% remained at 100 and 27.55% experienced a decrease. This summary by initial score is exploratory. Differences were calculated before rounding the mean.

Table 5. Direction of score changes according to pretest group

Early score	n	Increase (%)	Fixed (%)	Decrease (%)	Average difference
Less than 100	13,736	10,031(73.03)	2,373(17.28)	1,332(9.70)	10.76
100	10,438	0(0.00)	7,562(72.45)	2,876(27.55)	-2.53
All participants	24,174	10,031(41.49)	9,935(41.10)	4,208(17.41)	5.02

Source: processed by the author.

Of the 9,935 participants who did not experience a change in their scores, 7,562 (76.11%) had reached their maximum since the pretest. Thus, most scores remained at a threshold with no room for improvement. Meanwhile, 6,289 of the 13,736 participants with a pretest score of less than 100 (45.78%) reached their maximum on the posttest. This grouping reflects a pattern of change, not a comparison of effectiveness between equal groups: the initial score group of 100 can mathematically only remain the same or decrease.

The research findings indicate an uneven increase in recorded scores. Figure 2 shows a clustering of differences around zero, accompanied by changes in both positive and negative directions. The mean line is at +5.02 points, while the median difference remains at zero. The number of participants without change is known from the individual tabulations (41.10%), not from the height of a single histogram bar, as each bar represents a score interval. The large number of participants makes the mean estimate relatively precise, but the very small p-value does not yet establish the significance of the change for the performance of civil servant duties. Its practical significance needs to be assessed against learning objectives and achievement standards.

High initial scores limit the change that the test can capture. Participants with a pretest score of 100 have no room to improve their score, although they may gain deeper understanding. Staus et al. (2021) showed that initial scores near the upper limit can potentially bias changes in learning outcomes. The large proportion of maximum scores indicates measurement limitations at the upper end of the scale, not evidence that all participants have fully mastered integrity competencies. The narrowing of the standard deviation on the posttest should also be interpreted in conjunction with the accumulation of scores around 100. This pattern may reflect increasingly uniform scores, but some uniformity may be due to the test not providing scores above the upper limit.

Simulations by Šimkovic and Träuble (2019) suggest that ceiling effects can influence variance estimates and their uncertainty. This study does not directly justify testing these data; its relevance lies in reminding us that observed score statistics do not necessarily reflect changes in latent ability. High initial scores may be related to prior knowledge or the participants' familiarity with the material. Another possibility is that the test's demands were relatively easy for some participants. Both of these explanations cannot be confirmed because this study did not provide information on prior training experience, the scope of material mastered, or item difficulty parameters. Therefore, the available data are insufficient to determine whether the test

was of poor quality or whether participants had mastered all the material from the start.

The decline in scores among some participants also has more than one possible explanation. In pre-posttests, items are selected from a question bank and the package may change when the test is re-administered (each participant may receive a different set of items). Without item identification and evidence of package equivalence, score differences may reflect variations in test difficulty. Furthermore, grouping based on extreme initial scores may indicate regression toward the mean due to measurement variation (Barnett et al., 2005). The magnitude of each mechanism's contribution was not estimated in this study. Therefore, the decline in scores cannot be interpreted as a decline in training integrity or failure.

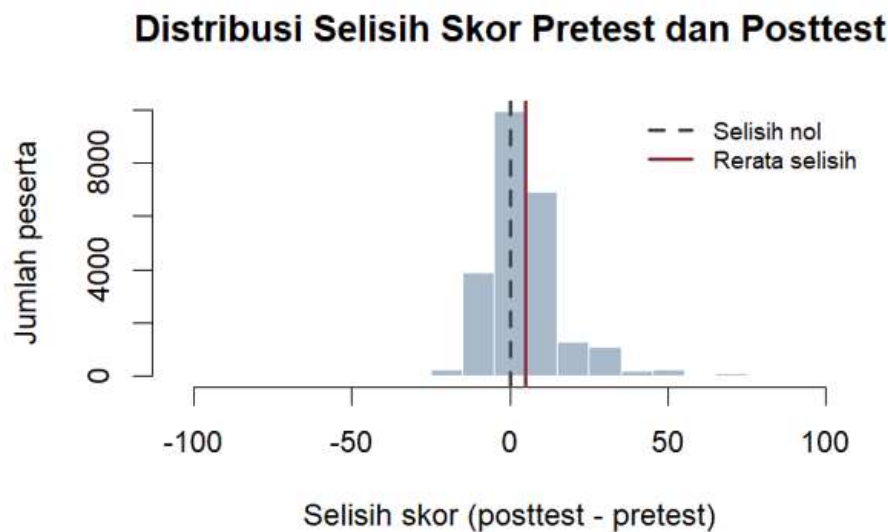


Figure 2. Distribution of the difference between posttest and pretest scores in the main analysis participants (n = 24,174).

High test scores do not diminish the need for character development based on integrity. Civil servants must exercise authority related to the public interest; therefore, the ability to choose correct answers must be followed up with consistent decisions and actions. Regulation of the Minister of Administrative and Bureaucratic Reform Number 38 of 2017 places integrity as a managerial competency encompassing honesty, alignment of behavior with ethics, and responsibility for decisions. Indicators range from individual behavior to team development and the creation of an ethical work environment (Ministry of Administrative and Bureaucratic Reform, 2017, Appendix II). This framework emphasizes the need to evaluate behavior in the workplace, beyond the cognitive achievements of training.

Research by Perlman et al. (2021) found that in the government sector, training emphasizes compliance over values and the public interest. This context may be similar to the training conditions for civil servants in Indonesia. This finding does not prove that the training in this study only teaches compliance, as the complete question bank from the Corruption Eradication Commission (KPK) is not available. As a follow-up, training can utilize reflection on work dilemmas, case studies, and assignment simulations to justify choices, as well as coaching and mentoring, including for participants who have achieved the maximum score.

Character building also requires organizational support. Demir et al. (2023) identified ethical knowledge, rule enforcement, incentives, and policy support as part of the local government's ethical infrastructure. Building on this foundation, fostering integrity behavior in the Ministry of Transportation can be designed through leadership role models, ethical dialogue, mentoring, and consistent follow-up. The relationship between this support and participant behavior has not been tested in this study's data. Reddick et al. (2025) tested a multidimensional framework of ethical leadership at the managerial level and positioned leadership behavior as a researchable component of public organizations. Relevant to follow-up training is the need to involve superiors as role models and as a forum for ethical consultation.

The Integrity Zone framework links governance and human resource changes with accountability, oversight, and service quality. Training can support changes in mindset and work culture, but posttest scores cannot be the sole measure of success in establishing an Integrity Zone or achieving a Corruption-Free Zone/Clean and Serving Bureaucracy Zone. The implications of the research findings are the need to link integrity learning with practice and evaluation in work units. Within the framework of competency development and the Corpu mandate, these findings provide a basis for improving the evaluation information used by organizations. Score increases can help identify achievement on tests, while ceiling effects indicate the need for further assessments that better distinguish participants' mastery at the upper end of the scale. Thus, LMS data can be used to determine the need for follow-up learning, while still considering the limitations of the instrument.

This direction aligns with the Ministry of Transportation Corpu Blueprint, which positions 2025–2026 as the foundation-building phase, directs the use of learning analytics, and differentiates learning evaluation from workplace implementation and organizational impact. Operationally, the ceiling effect provides a rationale for reviewing test/question quality levels, clarifying assessment blueprints, and documenting item identification and trial history. Examples for developing material in the question bank should include cases involving conflicts of interest in licensing services or pressure to bypass safety inspection procedures. Febrina and Ekowati (2026) also highlighted the need for contextualization of material and interactions in the Ministry of Transportation's CPNS Basic Training, although the population and outcome measures differed. Hughes et al. (2020) demonstrated a positive relationship between a supportive work environment and the transfer and sustainability of training outcomes. Follow-up evaluations can combine ethical situation assessments, supervisor observations, and evidence of procedure implementation. At the organizational level, how conflict of interest and service complaints are followed up can be considered as performance indicators in accordance with the work unit's duties and functions.

Limitations in interpretation in this study stem from the use of pretest-posttest archives without a comparison group, so score changes cannot be causally attributed to training. The equivalence of the packages and the position of scores in the trial history have not been verified. Differences in baseline scores among the 817 unpaired participants suggest potential selection, while the direction of their influence on the estimated change is unknown. Paired-test inference assumes independence among participants and does not account for the possibility of clustering within work units.

The absence of behavioral data and qualitative evidence also limits the explanation of mechanisms. Conclusions are therefore limited to recorded score changes and the maximum score concentration in the data that meets the analysis criteria.

CONCLUSIONS

Based on an analysis of 24,174 pairs of test data from the ASN Integrity Training, statistically there was a 5.02-point increase in the average score and a 14.12-point increase in the proportion of those with the maximum score (100). However, the median score difference remained at 0.00, and 41.10% of participants experienced no change in their scores. The high proportion of participants who had achieved a score of 100 since the beginning of the program (43.18% in the pretest) resulted in a significant ceiling effect. This effect limits the LMS's ability to capture knowledge gains and differentiate levels of mastery among high-ability participants. The increase in scores detected in the LMS system should be interpreted cautiously as a change in the test instrument's performance in a retrospective sample, rather than as direct evidence of the causal effectiveness of the training or the development of integrity in the workplace. The research design without a control group and the random selection of questions without a test-equating procedure resulted in a decrease in scores in 17.41% of participants, while an increase in the other group cannot be interpreted as an indication of changes in individual integrity. The main implication of this finding supports strengthening the evaluation function in the foundational development phase of the Ministry of Transportation Corporate University:

1. Short-Term Measurement Improvements: The Ministry of Transportation Corpu needs to conduct audits and documentation of question items in the LMS question bank, record the history of test trials, and apply test package equating techniques to ensure measurement standardization.
2. Advanced Assessment Instrument Development: Learning evaluation needs to transform from simply multiple choice based on cognitive compliance to the development of Situational Judgment Test (SJT) and work dilemma simulations in the context of licensing services and safety supervision.

Cognitive test results need to be followed up with non-test mechanisms in work units, including coaching/mentoring, exemplary commitment from superiors, and integrated monitoring with Integrity Zone (WBK/WBBM) indicators. Future research is recommended to use a comparative experimental design and measure the transfer of ethical behavior in real-world work environments.

REFERENCES

- Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, 50(2), 341-358. <https://doi.org/10.1111/j.1744-6570.1997.tb00911.x>
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215-220. <https://doi.org/10.1093/ije/dyh299>
- Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, 36(4), 1065-1105. <https://doi.org/10.1177/0149206309352880>

- Demir, T., Reddick, C. G., & Perlman, B. J. (2023). In search of ethics infrastructure in U.S. local governments: Building blocks or dead end? *Administration & Society*, 55(10), 1866–1892. <https://doi.org/10.1177/00953997231190571>
- Febrina, R., & Ekowati, A. (2026). Facilitator competencies gap analysis in the implementation of distance learning model for CPNS training in Ministry of Transportation. *Kontingensi: Jurnal Ilmiah Manajemen*, 14(1), 378–387.
- Hughes, A. M., Zajac, S., Woods, A. L., & Salas, E. (2020). The role of work environment in training sustainment: A meta-analysis. *Human Factors*, 62(1), 166–183. <https://doi.org/10.1177/0018720819845988>
- Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi. (2017). Peraturan Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Nomor 38 Tahun 2017 tentang Standar Kompetensi Jabatan Aparatur Sipil Negara. <https://peraturan.bpk.go.id/Download/123418/PERMENPAN%20NOMOR%2038%20TAHUN%202017.pdf>
- Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi. (2021). Peraturan Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Nomor 90 Tahun 2021 tentang Pembangunan dan Evaluasi Zona Integritas Menuju Wilayah Bebas dari Korupsi dan Wilayah Birokrasi Bersih dan Melayani di Instansi Pemerintah. <https://peraturan.bpk.go.id/Details/202223/permen-pan-rb-no-90-tahun-2021>
- Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi. (2024). Peraturan Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Nomor 5 Tahun 2024 tentang Perubahan atas Peraturan Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Nomor 90 Tahun 2021 tentang Pembangunan dan Evaluasi Zona Integritas Menuju Wilayah Bebas dari Korupsi dan Wilayah Birokrasi Bersih dan Melayani di Instansi Pemerintah. <https://peraturan.go.id/id/permenpanrb-no-5-tahun-2024>
- Kementerian Perhubungan. (2025). Peraturan Menteri Perhubungan Republik Indonesia Nomor PM 4 Tahun 2025 tentang Organisasi dan Tata Kerja Kementerian Perhubungan. <https://peraturan.go.id/id/permenhub-no-4-tahun-2025>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Perlman, B. J., Reddick, C. G., Demir, T., & Ogilby, S. M. (2021). What do local governments teach about in ethics training? Compliance versus integrity. *Journal of Public Affairs Education*, 27(4), 490–507. <https://doi.org/10.1080/15236803.2021.1972742>
- Reddick, C. G., Demir, T., & Perlman, B. J. (2025). A categorical framework of ethical leadership for public organizations: Testing a multi-dimensional model of local government managers. *Administration & Society*, 57(1), 3–33. <https://doi.org/10.1177/00953997241291833>
- Republik Indonesia. (2023). Undang-Undang Republik Indonesia Nomor 20 Tahun 2023 tentang Aparatur Sipil Negara. <https://peraturan.go.id/id/uu-no-20-tahun-2023>

- Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. PLOS ONE, 14(8), e0220889. <https://doi.org/10.1371/journal.pone.0220889>
- Staus, N. L., O'Connell, K., & Storksdieck, M. (2021). Addressing the ceiling effect when assessing STEM out-of-school time experiences. Frontiers in Education, 6, 690431. <https://doi.org/10.3389/feduc.2021.690431>