Tracking IT Professionals' Emotion, Cognition, and Motivation Development While Deriving Unit Tests Systematically

Muhammad Ridho Kurniawan Pratama[1*], Vina Ardelia Effendy[2], Fuad Mumtas[3], Prasetyo Wibowo Yunanto[4]

[1,2,3,4]Information Systems and Technology, Faculty of Engineering, Universitas Negeri Jakarta, Jakarta, DKI Jakarta, Indonesia
E-mail: [1*]muhammadridho@unj.ac.id, [2]vinaardelia@unj.ac.id, [3]fuadmumtas@unj.ac.id, [4]prasetyo.wy@unj.ac.id

(Received: 8 Sep 2026, revised: 13 Oct 2026, accepted: 16 Oct 2026)

Abstract

Developers now face growing pressure to master a systematic approach for writing unit tests. Limited evidence shows how their emotions, cognition, and motivation develop as they learn this process. This study followed 29 IT professionals (26 completed all waves) in a graduate-level Information Systems Requirements Analysis and Design course over eight weeks, and data were collected across four waves: baseline or pre-training (T1), post-training (T2), early adoption (T3), and final synthesis (T4). A mixed-methods design was used. This study quantitatively measured positive and negative emotions, interest and enjoyment, effort and importance, self-efficacy, and knowledge mastery. Reflective journals were analyzed thematically. Four of the six quantitative scales showed a significant change based on repeated-measures ANOVA. Positive emotions and interests developed in a staircase pattern and stabilized during the T2-T3 transition. Self-efficacy and knowledge mastery rose across the three measured waves (T1, T3, T4). Negative emotions and effort/importance showed non-significant changes. Thematic analysis of the journals produced five themes. Compared with the quantitative results, one theme on protocol direction matched closely in both direction and trend. The testing reconceptualization theme matched only in direction. The quantitative justifications and deeper system understanding themes could not be verified, since their corresponding scales were not measured at T2. An ambivalent theme (challenging and enjoyable) ran opposite to its scale. Because the study had no comparison group, these changes are descriptive rather than causal estimates of the protocol's effect, but they indicate when follow-up support is most useful.

Keywords: Developer Experience, Unit Test Derivation, Longitudinal Mixed-Methods, Self-Efficacy, Cognition-Motivation Development.

I. INTRODUCTION

Software quality is typically assessed using characteristics such as functional suitability, performance, interaction capability, reliability, and security [1], which are verified during testing [2]. Most developers, however, still rely on manually designed tests, which often leave coverage gaps and miss test cases for specific scenarios compared with systematic or AI-assisted approaches [3], [4], [5]. Systematic approaches such as Test-Driven Development improve code quality and requirement coverage [6], [7], [8], [9], [10], but ad hoc approaches let defects surface late, when they are more expensive to fix [6].

The human mental activity dimension is categorized into cognitive, affective, and conative [11]. Specifically, conation captures the volitional and self-regulatory drive behind goal-directed behavior [12]. Systematic testing plausibly influences all those dimensions, consistent with growing evidence linking programmers' emotional state to their cognitive performance during development tasks [13]. On the cognitive dimension, testing activity can develop critical thinking and problem-solving [14], [15]. Related to the affective dimension, task design and feedback can shape emotional awareness and motivation [16], [17]. Broader evidence shows that measurable emotional responses in learners are reliably triggered by programming tasks [18]. For the conative dimension, trust in the testing process shapes persistence, which can be eroded by excessive cognitive load and test quality [19], [20].

A systematic protocol for deriving unit tests [21] was developed, integrating test-first principles, use-case-driven testing, and requirement traceability to tests. It addressed a gap in the literature, which had focused mainly on system-level test derivation, and observations from experienced

DOI: 10.34148/teknika.v15i3.1578

professionals in a graduate-level Information Systems Requirements Analysis and Design course, whose unit test plans were unstructured, inconsistent, and largely limited to positive-flow scenarios with no clear traceability to requirements [21].

Human factors are comparatively understudied in software testing relative to other areas of software engineering [22], though recent work has begun to trace how developers' testing behavior emerges from the combination of technical, organizational, and social conditions [23]. Despite the protocol's existence, empirical research on professionals' experiences and competence development during its mastery remains limited. This study asks three questions. RQ1: How do affect, intrinsic motivation, self-efficacy, and protocol knowledge change across four waves? RQ2: Which themes emerge in the journals, and how does their prevalence change? RQ3: Where do the themes and scale trajectories converge or diverge?

The study contributes a longitudinal account of how practitioners' affect, motivation, and cognition develop while learning a systematic unit-test derivation protocol, and a joint display comparing scale and theme trajectories at each transition.

II. RESEARCH METHODS

The study used a four-wave longitudinal mixed-methods design over eight weeks: T1 (baseline), T2 (post-training), T3 (early adoption), and T4 (final synthesis). This design captures change that cross-sectional studies cannot.

Quantitative and qualitative data were collected to enable triangulation and enhance understanding of the observed patterns. This approach follows the established guidance for designing mixed-methods studies in software engineering [24].

Participants were students of the Information Systems Requirements Analysis and Design graduate-level course. Most of the students worked in the IT sector; hence the term IT professionals, although the sample also included full-time students and non-IT employees. They were recruited by inviting the entire class (25–30 students in 5–6 teams) through in-class briefings, a study explanation, and a Q&A session.

The following instruments are used to measure each dimension: the affective dimension for both positive (PA) and negative emotions (NA) are adapted from the short-form Positive and Negative Affect Schedule [25] and likert scale specified by Watson et al. [26] were used; the conative dimension is evaluated using the eight items Self-Efficacy Scale (SE) which is constructed by following Bandura's guidance [27] and the Effort/Importance (EI) and Interest/Enjoyment (IE) subscales adapted from Intrinsic Motivation Inventory [28]; and the cognitive dimension is measured with a twelve multiple choice questions knowledge quiz (KP) around the testing in general and unit test specific. The quantitative instruments were supplemented with journal entries to capture participants' feelings and thoughts qualitatively. The authors forward-translated the instruments into Bahasa Indonesia, back-translated into English to ensure the item meaning was kept, and reworded them for the unit test derivation context. The knowledge quiz was written by the authors to cover the protocol and general unit-testing concepts (Appendix 6). A pilot with five non-participants checked item clarity and readability, which led to removing one EI item.

T1 (week 1, baseline) was conducted to assess six scales: PA, NA, KP, SE, IE, and EI before training. T2 (week 1, same day) consisted of a one-hour training session that included explaining concepts, demonstration, and guided practice. After the training session (T2), PA, NA, IE, and EI were re-administered, and the first journal entry was collected. SE and KP were not measured at T2 to reduce participant burden, because T2 took place right after training on the same day as T1. SE and KP therefore have three time points and cannot show change directly after training. Subsequently, on T3 (week 3) and T4 (week 8), the full quantitative assessments and journal entry were repeated, with T4's entry reflecting on the overall journey from beginning to end.

Both analysis techniques were used. First, the six scales were analyzed descriptively to capture means, standard deviations, and trajectory plots. Each scale was analyzed with a one-way repeated-measures ANOVA in JASP 0.98.1. Because the test requires a score from every participant at every wave, it was planned on complete cases only, while descriptive statistics used all available responses at each wave. Mauchly test of sphericity showed violations for IE ($p = 0.035$) and SE ($p = 0.006$), so the Greenhouse-Geisser correction was applied to all scales for consistency. Effect sizes are reported as generalized eta-squared (η^2_G) [29], and pairwise comparisons used the Holm correction with Cohen's d from JASP (significance at $p < 0.05$). Non-significant results are read as no detectable change at this sample size.

Beyond the quantitative measure, the reflective journals were analyzed with inductive thematic analysis [30]. The first author coded all entries: reading them twice per participant across waves, coding recurring phrases openly, and grouping codes into five themes, each assigned to a dimension. Prevalence was counted at the data-item level [30]: a participant counted once per theme per wave, and one entry could support several themes. With a single coder, there was no disagreement resolution or inter-rater check; credibility rests on a codebook with exemplar quotes kept as an audit trail and on verbatim quotes reported with translations.

The two strands were analyzed separately and then merged in a joint display, following a convergent design [24], [31]. Each theme was paired with its closest scales. A pair was strongly convergent if direction and shape agreed at both transitions (T2 to T3, T3 to T4), partially convergent if they agreed only from the first to the last wave, inversely related if they moved in opposite directions, and not yet verifiable if the scale was missing at a wave.

Formal approval from an institutional ethics committee was not obtained. The first author completed and signed the ethics self-assessment, which is mandatory under the faculty's

basic research scheme. Each questionnaire (T1 to T4) began with a study description and an explicit consent item. Participation was voluntary and could be withdrawn at any wave. Responses were pseudonymized, and quotes carry no identifying details.

III. RESULTS

A. Participants

29 participants were recruited to follow this four-wave study. Attrition occurred monotonically, resulting in 28 participants remaining at T3 and 26 at T4, yielding an 89.7% retention rate across all waves. The data from the 26 participants who completed all waves were used for inferential analysis.

Of the 29 participants, 17 were men, and 12 were women. At T1, 14 (48%) worked in IT or software roles, 6 (21%) were company employees with no stated role, 4 (14%) held government or other non-IT roles, and 5 (17%) were full-time students. Most had limited IT experience: 59% had less than five years, 34% had five to nine years, and only 7% had over ten years. This indicates that most of the sample were at junior-to-mid-level seniority. On software testing experience, 79% had prior testing experience for both manual and automated testing. Meanwhile, 21% were new to testing.

B. Instrument Reliability

All scales, shown in Table 1, showed acceptable reliability across all waves, exceeding the 0.7 threshold [32].

Table 1. Instrument Reliability (Cronbach's α)

Scale	T1	T2	T3	T4
Positive Emotion (PA, 4-item)	0.919	0.896	0.880	0.873
Negative Emotion (NA)	0.930	0.874	0.933	0.932
Interest/Enjoyment (IE)	0.923	0.937	0.969	0.964
Effort/Importance (EI, 3-item)	0.776	0.781	0.795	0.723
Self-Efficacy (SE, 8-item)	0.885	not measured	0.901	0.950

PA and NA measure the affective dimension; IE and EI reflect conative intrinsic motivation (interest/enjoyment and effort/importance); SE indicates conative self-efficacy; and the Knowledge Quiz (KP) assesses the cognitive dimension.

One item on the EI subscale ('I didn't try very hard at this activity') was eliminated after the pilot due to its negative correlation with the scale (pilot α = 0.540), so all four waves used the same three EI items. The PA item 'Alert' was retained through T1 and T2 (5-item α = 0.765 and 0.739, respectively) before being dropped after a reliability review between T2 and T3. Its Indonesian rendering, *"Waspada"*, also connotes wariness, and its corrected item-total correlation was negative at both waves (r = -0.04 and -0.17; other items 0.60 to 0.87). The decision used item reliability only. The 4-item PA scoring was subsequently applied retroactively to T1 and T2 to maintain a consistent dependent variable across all four waves (Table 1), yielding improved reliability across waves.

C. Descriptive Statistics

Table 2 reports mean scores and standard deviations for each scale at every wave and Figure 1 plots the same trajectories.

Table 2. Each Scale's Descriptive Statistics

Scale	T1	T2	T3	T4
PA	3.388 ± 0.890	3.897 ± 0.852	3.830 ± 0.819	4.288 ± 0.635
NA	2.352 ± 1.035	2.000 ± 0.733	2.264 ± 0.929	2.208 ± 1.090
IE	4.621 ± 1.141	5.086 ± 1.053	5.080 ± 1.126	5.827 ± 1.065
EI	5.391 ± 1.244	5.230 ± 1.035	5.143 ± 1.028	5.308 ± 1.289
SE	5.530 ± 1.755	not measured	6.897 ± 1.569	7.856 ± 1.288
KP	8.690 ± 2.123	not measured	9.464 ± 2.808	10.580 ± 1.983

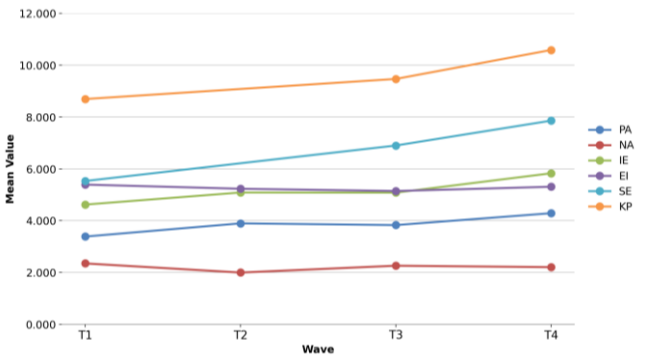


Figure 1. Trajectory Chart of the Scale's Mean Value

Descriptive differences alone cannot separate real change from sampling variation, so each scale was tested inferentially.

D. Inferential Testing

As seen in Table 3, a repeated-measures ANOVA (with the Greenhouse-Geisser correction) was conducted on each of the six scales independently, with time wave as the subject factor. Repeated-measures factor levels were applied: four levels for PA, IE, and EI (denoting four waves, T1 to T4, completely), and three levels for SE and KP (denoting three waves, T1, T3, and T4). Four out of six scales changed

DOI: 10.34148/teknika.v15i3.1578
TEKNIKA, Volume 15(3), November 2026, pp. 553-562
ISSN 2549-8037, EISSN 2549-8045

significantly, NA and EI did not (η^2_G = 0.0145 and 0.0036). Holm-corrected post hoc tests revealed the following patterns. Cohen's d is reported as an absolute value; all significant differences were increases.

PA showed significant differences across all wave pairs except T2 to T3 (p = 0.544), with the strongest effect observed from T1 to T4 (d = 1.177; p < 0.001). A marginal effect on IE was observed only from T1 to T2 (p = 0.074), with a confirmed change evident only from T1 to T4 (d = 1.207; p < 0.001). SE was the only scale showing significant differences across all wave pairs, including T1 to T3 (d = 0.889) and T1 to T4 (d = 1.483; both p < 0.001). KP was significant only across the entire T1 to T4 period (d = 0.768; p = 0.002), although T1 to T3 (p = 0.145) and T3 to T4 (p = 0.067) did not reach significance individually; the trend remained consistently upward.

E. Thematic Analysis Findings

Participants also wrote about their experience in reflective journals at T2, T3, and T4, from which five themes emerged inductively (Table 4, Figure 2).

Theme A, procedural clarity, moved from *"ternyata jauh lebih terstruktur dan sistematis, jadi lebih tau what-to-do/to-do-list selanjutnya apa saja"* ("much more structured and systematic, so I know better what to do next"; P27, T2) to *"sudah tidak ngawang-ngawang lagi"* ("I'm no longer just guessing blindly"; P5, T4).

Table 3. Inferential Test Results

Scale	F (df)	p	η^2_G	Effect Size [29], [41]
PA	F(2.38, 59.61) = 14.33	< 0.001	0.1543	Medium
NA	F(2.38, 59.58) = 1.02	0.379	0.0145	Negligible
IE	F(2.28, 57.01) = 12.18	< 0.001	0.1617	Medium
EI	F(2.46, 61.52) = 0.19	0.866	0.0036	Negligible
SE	F(1.49, 37.24) = 29.87	< 0.001	0.2785	Large
KP	F(1.97, 49.17) = 7.40	0.002	0.0944	Small

Note: Greenhouse-Geisser corrected. η^2_G = generalized eta-squared. Pairwise d values in the text are absolute; all significant differences are increases

Table 4. Qualitative Themes

#	Theme	Dimension	T2 (n=29)	T3 (n=28)	T4 (n=26)	Trend
A	The Protocol Gives Clear Direction for Creating Test Cases	Conative	17 (59%)	15 (54%)	20 (77%)	Strengthening
B	Testing as More Than Just Input-Output Checking	Cognitive	10 (34%)	13 (46%)	17 (65%)	Consistently rising
C	Quantitative Justification for Test Case Completeness	Cognitive	5 (17%)	12 (43%)	16 (62%)	Consistently rising
D	Challenging and Enjoyable at the Same Time	Affective	14 (48%)	13 (46%)	8 (31%)	Fading
E	Deeper Understanding of the System	Cognitive	8 (28%)	11 (39%)	15 (58%)	Consistently rising

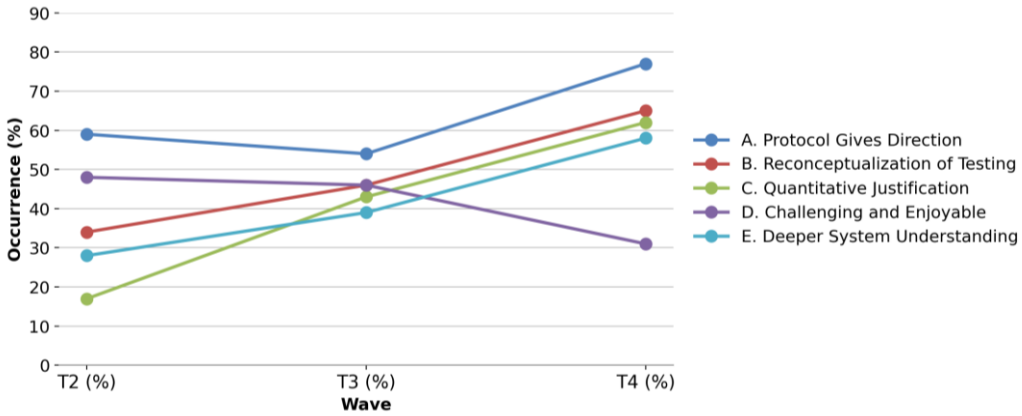


Figure 2. Trajectory of Qualitative Theme Occurrence by Wave

Theme B captured a shift from treating testing as a check on isolated outputs to confirming the underlying logic is correctly implemented to safeguard code quality from the beginning of development, as shown in this quote: *"pengujian unit justru merupakan fondasi penting untuk menjaga kualitas kode sejak awal"* ("unit testing is actually an important foundation for maintaining code quality from the start"; P12, T2).

Theme C, the discovery of quantitative justification, grew fastest of all five themes, depicted with this quote: "*cyclomatic complexity juga ngebantu banget buat nentuin minimal test case, jadi lebih jelas harus ngetes apa aja dan nggak asal coba-coba*" ("cyclomatic complexity really helps a lot in determining the minimum test cases, not just randomly guessing"; P12, T3).

Theme D was the only theme to fade, moving from *"pusing tapi asik"* ("confusing but fun"; P17, T2) to "*jika sudah tau caranya, itu akan sangat mudah dalam membuatnya*" ("once you know how, it becomes very easy"; P20, T4).

Theme E reflected a deepening system understanding as represented by this quote: *"proses seperti membuat CFG, menghitung cyclomatic complexity, dan menentukan independent path membantu saya melihat bagaimana suatu program berjalan"* ("building a CFG, calculating cyclomatic complexity, and determining independent paths helped me to see how a program actually runs"; P3, T4).

IV. DISCUSSIONS

A. Two Developmental Trajectories

A repeated-measures ANOVA finding classifies the constructs by trajectory shape rather than dimension segregation. One group exhibits a staircase pattern. The pattern depicts a notable increase from T1 to T2, a plateau from T2 to T3, followed by another significant rise from T3 to T4, with scores consistently staying above the T1 level. Both PA (affective) and IE (conative) are part of this group and behave nearly identically, including a plateau from T2 to T3 ($p = 0.544$ and $p = 0.847$). Decreased novelty [33], temporary mismatches between skills and task demands [34], depleted self-regulatory resources [35], and changes in perceived challenge independent of actual performance [36] are possible, unmeasured explanations for the plateau.

A second group rose at both measured transitions. Those constructs are SE (conative, with the strongest effect at $\eta^2_G = 0.2785$) and KP (cognitive). Neither was measured at T2, so a post-training plateau cannot be ruled out. This aligns with education research indicating that learners' self-perceptions of competence tend to mirror their actual mastery as their skills improve [37], [38], [39]. This provides a reasonable explanation for why SE's trajectory more closely resembles KP's than IE's, its other conative construct.

NA (affective) and EI (conative), the two scales that remained statistically insignificant, support this interpretation: despite stemming from different dimensions, they both show a flat trajectory.

B. Three Developmental Phases

There are three phases labeled as the orientation phase (T1 to T2), the productive struggle phase (T2 to T3), and the internalization phase (T3 to T4). The labels are interpretive, not tested phases.

In the orientation phase, PA increased significantly right after training ($d = 0.589$; $p = 0.005$), while the rise in IE was only marginal ($p = 0.074$). This aligns with Theme A and Theme B, which already account for 59% and 34% by T2. Procedural clarity and testing reconceptualization increase rapidly, whereas enjoyment of the activity increases significantly only in the subsequent wave.

During the next productive struggle phase, PA and IE are shown to be stabilized ($p = 0.544$ and $p = 0.847$). SE continues to increase significantly. KP also trends upward, although the changes from T1 to T3 aren't statistically significant ($p = 0.145$). Theme D remains high at 46%, and Theme C starts emerging strongly at 43%, indicating that affect remains steady while competence continues to develop.

In the last internalization phase, the widespread rebound was observed in three constructs. Those constructs are PA ($d = 0.481$; $p = 0.008$), IE ($d = 0.700$; $p = 0.002$), and SE ($d = 0.594$; $p < 0.001$). KP also increases marginally ($p = 0.067$). These results are consistent with gradual knowledge growth from T1 to T4. Theme D falls from 46% to 31% (−15 pp), while Themes A, B, C, and E all reach their peak levels. One plausible reason behind this late recovery is that interest itself becomes a resource. Once the interest is sustained, it can replenish the effort and self-regulation needed to persist even under demanding conditions [35]. That pattern is consistent with evidence that sustained situational interest supports ongoing self-regulated effort among adult online learners [40].

C. Quantitative and Qualitative Convergence

The five qualitative themes are compared with their corresponding quantitative scales, as shown in Table 5. This compares both the direction of the change and the shape of the trajectory at each transition.

Theme A aligns closely with IE: the minor T2 to T3 decrease (−5 points, 17 to 15 mentions) is small enough to reflect attrition and wave-to-wave variation in what participants chose to write about, and is consistent with IE's stable trend during this period ($p = 0.847$). Both then show a significant rise from T3 to T4.

Generally, the overall trend of Theme B follows PA and IE. Theme B rose by 12 points whilst PA and IE remained stable from T2 to T3. Despite remaining unchanged in affect and motivation, the findings indicate that participants began discussing reconceptualization of testing.

DOI: 10.34148/teknika.v15i3.1578
TEKNIKA, Volume 15(3), November 2026, pp. 553-562
ISSN 2549-8037, EISSN 2549-8045

Table 5. Convergence of Quantitative and Qualitative Data

Theme	Change T2 to T3	Change T3 to T4	Comparison Quantitative Scale	Shape Fit	Conclusion
A. Protocol Gives Direction	−5 pp	+23 pp	IE	Flat (p = 0.847) then a significant jump (p = 0.002)	Strongly convergent
B. Reconceptualization of Testing	+12 pp	+19 pp	PA, IE	Both scales flat (p = 0.544 and p = 0.847) right when the theme rose 12 points	Partially convergent, only when viewed from the beginning to the end of the study
C. Quantitative Justification	+26 pp	+19 pp	SE, KP	Cannot be checked because both scales were not measured at T2	Not yet fully verifiable
D. Challenging and Enjoyable	−2 pp	−15 pp	IE	Both flat at T2 to T3, then at T3 to T4 Theme D drops sharply exactly when IE rises sharply	Inversely related
E. Deeper System Understanding	+11 pp	+19 pp	SE, KP	Cannot be checked because both scales were not measured at T2	Not yet fully verifiable

Subsequently, Theme C and Theme E shared the same limitation. The absence of the SE and KP in T2 means the convergence between the T2 and T3 qualitative data cannot be compared. Their convergence therefore remains unconfirmed.

Theme D shows a different pattern compared to the others. The relationship between the theme and IE is opposite. Both stay flat from T2 to T3, then sharply diverge from T3 to T4, with Theme D decreasing from 46% to 31% while IE increases significantly (p = 0.002). This fits with Theme D being coded as a single ambivalent theme (challenging and enjoyable) that often appears in the same sentence and is not separated during coding. As participants enjoyed the process more, they were less likely to mention the challenging aspect, leading to a decline in Theme D as IE rises. Only the enjoyable aspect is supported here. The challenging side that linked to NA lacks support in the data since NA did not change significantly throughout (η^2_G = 0.0145; p = 0.379). The fading of Theme D likely reflects decreased mention of challenge rather than a real reduction in negative feelings.

Overall, only Theme A shows a clear convergence in both direction and curve shape. Theme B shows convergence only from beginning to end. Verification of Theme C and Theme E is incomplete due to the T2 measurement gap. Theme D moves in the opposite direction to IE, supporting its enjoyable aspect but not its challenging aspect.

D. Implications of the Findings

PA rose significantly right after the one-hour session (T1 to T2), although without a comparison group, the session cannot be named as the cause. IE rose significantly only from T3 to T4 (p = 0.002), after independent practice, suggesting that one session is not enough for interest to consolidate.

SE was the only scale to rise significantly at every measured transition, with the largest effect (η^2_G = 0.2785). Theme C, the protocol as a quantitative basis for test completeness, may contribute, as 79% of participants had testing experience; this link was not tested and remains a hypothesis.

E. Limitations

Several limitations are acknowledged in this study. First, statistical power to detect small-to-medium effects at individual wave transitions could be limited (e.g., on KP) due to inferential tests that only rely on 26 complete data points. Second, the observed significant changes cannot be fully separated from external factors because of the absence of a control group. Third, SE and KP were not measured at T2, a design limitation rather than a minor gap: their three-point trajectories cannot reveal change directly after training, and Themes C and E cannot be checked at T2 to T3. Fourth, EI had the lowest reliability (α = 0.723 to 0.795); with three items, it is the least precise scale, so its null result is the weakest finding. Fifth, measurement equivalence was not tested: PA and NA referred to "today" at T1 and T2 but to "the last few weeks" at T3 and T4, and participants' understanding of the constructs may have shifted. Invariance testing needs a far larger sample; the stable reliabilities in Table 1 are consistent with, but do not establish, the equivalence. Sixth, one author coded all journals without a second coder. Seventh, the sample came from a single course and included full-time students, which limits generalizability. Eighth, the study had no formal ethics committee review, although the safeguards in Section II were applied. Finally, all data came from participants' self-reports.

V. CONCLUSIONS

Over eight weeks of learning and applying the protocol, four of six constructs changed significantly. With a single-group design, these changes cannot be attributed to the protocol alone. Those constructs are SE (large effects), PA

and IE (medium effects), and KP (small effects). The two others, NA and EI, showed non-significant changes.

The findings from repeated-measures ANOVA resulted in two distinct trajectory shapes across affective, conative, and cognitive dimensions. PA and IE constructs had a staircase pattern. PA and IE increased rapidly after training, stabilized during productive struggle, and finally rebounded during internalization. SE and KP increased at both measured transitions; their shape directly after training is unknown.

Only Theme A converged with its scale in both direction and shape. Theme B converged only from the first to the last wave; Themes C and E could not be verified, and Theme D ran opposite to IE. The relatively small sample used for inferential testing requires caution when interpreting transitions that did not achieve significance on their own.

In a training design, organizers should plan the follow-up support during the independent adoption phase. IE (T3 to T4) and KP (T1 to T4) rose significantly only weeks after the session, and PA and IE plateaued from T2 to T3 while SE kept rising, so support matters most during independent practice.

For future research, a full scale should be measured completely across all waves to address the convergence verification gap, as they do for Themes C and E with SE and KP. Dividing the ambivalent emotions (as in Theme D) into separate themes could facilitate more accurate convergence checks against NA and IE. Additionally, the professionals' experience and SE relationship can be analyzed by using cross-tabulation to clarify their resonance, which is currently interpretive. The statistical power increases at each transition (especially on KP) could be achieved by using a larger sample size.

ACKNOWLEDGEMENT

This research is funded by the Faculty of Engineering, Universitas Negeri Jakarta under the Faculty's Basic Research (*Penelitian Dasar Fakultas*) scheme with contract number T/030/7.FT-PPK/Kontrak-Penelitian/PT.01.03/III/2026.

APPENDIXES

Appendix 1. PA items (1-5 Likert scale)

T1: To what extent do you feel TODAY when you think about unit testing?
T2: To what extent do you feel TODAY after taking the training when systematically creating unit test scenarios?
T3 & T4: To what extent do you feel in the LAST FEW WEEKS when systematically creating unit test scenarios?

1. Inspired
2. Alert (T1 and T2 only; excluded from scoring)
3. Excited
4. Enthusiastic
5. Determined

Appendix 2. NA items (1-5 Likert scale)

T1: To what extent do you feel TODAY when you think about unit testing?
T2: To what extent do you feel TODAY after taking the training when systematically creating unit test scenarios?
T3 & T4: To what extent do you feel in the LAST FEW WEEKS when systematically creating unit test scenarios?

1. Afraid
2. Upset
3. Nervous
4. Scared
5. Distressed

Appendix 3. IE items (1-7 likert scale)

To what extent are the following statements true for you in applying a systematic method of deriving unit test scenarios?

1. I enjoyed this activity very much
2. Doing this activity was fun
3. I would describe this activity as very interesting
4. While doing this activity, I was thinking about how much I enjoyed it

Appendix 4. EI items (1-7 Likert scale)

To what extent are the following statements true for you in applying a systematic method of deriving unit test scenarios?

1. I put a lot of effort into this activity
2. It was important to me to do well at this activity
3. I tried very hard while doing this activity
4. I didn't try very hard at this activity (not administered on all waves)

Appendix 5. SE items (0-10 Likert scale)

To what extent are you confident in doing the following tasks when applying the systematic method for deriving unit test scenarios?

1. Creating a use case specification
2. Creating a sequence diagram
3. Creating an algorithm of the method based on the sequence diagram
4. Creating a control flow graph (CFG)
5. Calculating cyclomatic complexity
6. Identifying the basis path from an algorithm
7. Creating a unit test scenario
8. Creating a complete test case with an input sample and the expected output

Appendix 6. KP items

1. What is the main difference between a test case and a test scenario?
2. What is the primary purpose of unit testing?
3. What is meant by test coverage?
4. What is the primary function of a Control Flow Graph (CFG)?
5. In basis path testing, what is meant by an independent path?
6. Sequence diagrams in the context of testing are used to…
7. The basis path testing method was developed by…

8. The first step in the systematic method for deriving unit test scenarios is…
9. What is the recommended cyclomatic complexity value to ensure a module or program can be effectively tested?
10. The cyclomatic complexity value is directly proportional to the number of test cases required.
11. What is the minimum number of test cases required to achieve 100% unit test coverage if a program's cyclomatic complexity is 5?
12. Consider the following test case snippet:
    1. *user = new User("John", 25)*
    2. *result = user.isAdult()*
    3. *expect(result).toBe(true)*

    Parts 1, 2, and 3 represent the phases in the following order…

Appendix 7. Reflective Journal

- T2: Tell us about your experience after completing the training on systematically creating unit test scenarios. What were your impressions and what did you learn?
- T3: Tell us about your experience systematically creating unit test scenarios. What did you experience and learn so far?
- T4: Reflect on your overall experience over the past few weeks in systematically creating unit test scenarios. What were the most important lessons you learned?

REFERENCES

[1] ISO/IEC 25010:2023, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model*. Geneva: International Organization for Standardization, 2023.

[2] ISO/IEC/IEEE 29119-1:2022, *Software and systems engineering — Software testing — Part 1: General concepts*. Geneva: International Organization for Standardization, 2022.

[3] H. A. Ramzan, S. Ramzan, and T. Kalsum, "Test-Driven Development (TDD) in Small Software Development Teams: Advantages and Challenges," in *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan: IEEE, Feb. 2024, pp. 1–5. doi: 10.1109/ICACS60934.2024.10473291.

[4] A. J. H. Simons, N. Griffiths, and C. Thomson, "Feedback-Based Specification, Coding and Testing with JWalk," in *Testing: Academic & Industrial Conference - Practice and Research Techniques (taic part 2008)*, IEEE, Aug. 2008, pp. 69–73. doi: 10.1109/TAIC-PART.2008.20.

[5] R. Habarneh, H. Fawareh, and G. Albesani, "Scenario-Driven Software Testing: Comparative Analysis of Manual and AI-Based Test Case Generation," in *2025 26th International Arab Conference on Information Technology (ACIT)*, IEEE, Dec. 2025, pp. 291–294. doi: 10.1109/ACIT68900.2025.11510542.

[6] V. Bakhtiary, T. J. Gandomani, and A. Salajegheh, "The effectiveness of test-driven development approach on software projects: A multi-case study," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2030–2037, Oct. 2020, doi: 10.11591/eei.v9i5.2533.

[7] B. Papis, K. Grochowski, K. Subzda, and K. Sijko, "Experimental Evaluation of Test-Driven Development With Interns Working on a Real Industrial Project," *IEEE Transactions on Software Engineering*, vol. 48, no. 5, pp. 1644–1664, May 2022, doi: 10.1109/TSE.2020.3027522.

[8] Z. Wang, X. Guo, and T. Tsuchiya, "Graph-Centric Approaches for Coverage Optimization in Software Requirement Testing," in *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, Jul. 2025, pp. 1270–1280. doi: 10.1109/COMPSAC65507.2025.00160.

[9] F. Taufiqurrahman, S. Widowati, and M. J. Alibasa, "The Impacts of Test Driven Development on Code Coverage," in *2022 1st International Conference on Software Engineering and Information Technology (ICoSEIT)*, IEEE, Nov. 2022, pp. 46–50. doi: 10.1109/ICoSEIT55604.2022.10030006.

[10] Md. S. Rahman, A. Kumar Saha, U. Chakraborty, H. T. Sujana, and S. M. A. Shafi, "Evaluating the impact of Test-Driven Development on Software Quality Enhancement," *International Journal of Mathematical Sciences and Computing*, vol. 10, no. 3, pp. 51–76, Sep. 2024, doi: 10.5815/ijmsc.2024.03.05.

[11] E. R. Hilgard, "The trilogy of mind: Cognition, affection, and conation," *J. Hist. Behav. Sci.*, vol. 16, no. 2, pp. 107–117, Apr. 1980, doi: 10.1002/1520-6696(198004)16:2<107::AID-JHBS2300160202>3.0.CO;2-Y.

[12] W. G. Huitt and S. C. Cain, "An overview of the conative domain. Educational Psychology Interactive," 2005. Accessed: Oct. 13, 2026. [Online]. Available: http://www.edpsycinteractive.org/brilstar/chapters/conative.pdf

[13] K. Kaur, A. Mishra, and P. Chopra, "Emotional Dynamics in Software Engineering: A Comprehensive Review of Cognitive Skills and Affective States," in *Proceedings of International Conference on Computing Systems and Intelligent Applications*, A. Jaiswal, S. Anand, A. E. Hassanien, and A. T. Azar, Eds., Singapore: Springer Nature Singapore, 2026, pp. 803–824.

[14] E. Enoiu, G. Tukseferi, and R. Feldt, "Towards a Model of Testers' Cognitive Processes: Software Testing as a Problem Solving Approach," in *2020 IEEE 20th International Conference on Software*

Quality, Reliability and Security Companion (QRS-C), IEEE, Dec. 2020, pp. 272–279. doi: 10.1109/QRS-C51114.2020.00053.

[15] E. P. Enoiu, G. Gay, J. Esber, and R. Feldt, "Understanding Problem Solving in Software Testing: An Exploration of Tester Routines and Behavior," in *(eds) Testing Software and Systems. ICTSS 2023. Lecture Notes in Computer Science*, vol. 14131 LNCS, Springer, Cham, 2023, pp. 143–159. doi: 10.1007/978-3-031-43240-8_10.

[16] L. Silva, M. Lima, and H. Madeira, "Emotional Awareness in Software Development: Evaluating Software Project Managers' Perspectives," *IEEE Access*, vol. 13, pp. 168268–168286, 2025, doi: 10.1109/ACCESS.2025.3613682.

[17] K. Madampe, R. Hoda, and J. Grundy, "The Emotional Roller Coaster of Responding to Requirements Changes in Software Engineering," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 1171–1187, Mar. 2023, doi: 10.1109/TSE.2022.3172925.

[18] N. Atman Uslu and A. Onan, "Exploring students' emotions towards programming: Analysing sentiments using concurrent conversion mixed methods," *Educ. Inf. Technol. (Dordr).*, vol. 30, no. 14, pp. 19879–19909, Sep. 2025, doi: 10.1007/s10639-025-13482-z.

[19] M. Gruber and G. Fraser, "A Survey on How Test Flakiness Affects Developers and What Support They Need To Address It," in *2022 IEEE Conference on Software Testing, Verification and Validation (ICST)*, IEEE, Apr. 2022, pp. 82–92. doi: 10.1109/ICST53961.2022.00020.

[20] O. Parry, G. M. Kapfhammer, M. Hilton, and P. McMinn, "Surveying the developer experience of flaky tests," in *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*, New York, NY, USA: ACM, May 2022, pp. 253–262. doi: 10.1145/3510457.3513037.

[21] M. R. K. Pratama, D. Utama, and R. Fahmi, "Test-First Protocol for Deriving Unit Tests from Use Case Specifications," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 15, no. 1, pp. 72–80, Feb. 2026, doi: 10.22146/jnteti.v15i1.24073.

[22] N. Yahaya, N. S. M. Yusop, and M. M. Rosli, "Is Human Factors Being Studied in Software Testing?: Review of the Literature," in *2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICRAIE59459.2023.10468409.

[23] M. Swillus, R. Hoda, and A. Zaidman, "On the emergence of testing strategies: A socio-technical grounded theory," *Empir. Softw. Eng.*, vol. 31, no. 5, p. 147, Sep. 2026, doi: 10.1007/s10664-026-10828-0.

[24] M.-A. Storey, R. Hoda, A. Maciel Paz Milani, and M. T. Baldassarre, "Guiding principles for mixed methods research in software engineering," *Empir. Softw. Eng.*, vol. 30, no. 5, p. 138, Sep. 2025, doi: 10.1007/s10664-025-10629-x.

[25] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the Positive and Negative Affect Schedule: evaluation of factorial validity and invariance across demographic variables in a community sample," *Pers. Individ. Dif.*, vol. 27, no. 3, pp. 405–416, Sep. 1999, doi: 10.1016/S0191-8869(98)00251-7.

[26] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales.," *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, Jun. 1988, doi: 10.1037/0022-3514.54.6.1063.

[27] A. Bandura, "Guide for Creating Self-Efficacy Scale," in *Self-Efficacy Beliefs of Adolescents*, F. Pajares and T. Urdan, Eds., Emerald Publishing Limited, 2006, pp. 307–337. doi: 10.1108/978-1-60752-750-3.

[28] E. McAuley, T. Duncan, and V. V. Tammen, "Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis," *Res. Q. Exerc. Sport*, vol. 60, no. 1, pp. 48–58, Mar. 1989, doi: 10.1080/02701367.1989.10607413.

[29] R. Bakeman, "Recommended effect size statistics for repeated measures designs," *Behav. Res. Methods*, vol. 37, no. 3, pp. 379–384, Aug. 2005, doi: 10.3758/BF03192707.

[30] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

[31] M. D. Fetters, L. A. Curry, and J. W. Creswell, "Achieving Integration in Mixed Methods Designs—Principles and Practices," *Health Serv. Res.*, vol. 48, no. 6pt2, pp. 2134–2156, Dec. 2013, doi: 10.1111/1475-6773.12117.

[32] J. C. Nunnally, *Psychometric Theory*, 2nd ed. New York, NY: McGraw-Hill, 1978.

[33] G. M. Fastrich and K. Murayama, "Development of Interest and Role of Choice During Sequential Knowledge Acquisition," *AERA Open*, vol. 6, no. 2, Apr. 2020, doi: 10.1177/2332858420929981.

[34] C. Schrader and S. Kalyuga, "Expertise reversal effect in a pen-tablet-based learning environment: The role of learning-centered emotions in the interplay between learner expertise and task complexity," *British Journal of Educational Psychology*, vol. 93, no. S2, pp. 270–286, Aug. 2023, doi: 10.1111/bjep.12547.

[35] D. B. Thoman, J. L. Smith, and P. J. Silvia, "The Resource Replenishment Function of Interest," *Soc.*

Psychol. Personal. Sci., vol. 2, no. 6, pp. 592–599, Nov. 2011, doi: 10.1177/1948550611402521.

[36] A. Shum and L. K. Fryer, "Examining short-term self-efficacy change connections with performance, perceived difficulty, and interest," *Learn. Instr.*, vol. 101, p. 102223, Feb. 2026, doi: 10.1016/j.learninstruc.2025.102223.

[37] F. Núñez-Regueiro *et al.*, "I Feel Competent, Therefore I Am: Self-concept and Skill Interact at Different Speeds," *Educ. Psychol. Rev.*, vol. 38, no. 1, p. 39, Dec. 2026, doi: 10.1007/s10648-026-10131-w.

[38] S. Avdiaj and R. Paçarizi, "Teachers' Attitudes Towards Self-Efficacy in the Compilation of Tests According to Bloom's Taxonomy Framework," *Journal of Educational and Social Research*, vol. 14, no. 6, p. 223, Nov. 2024, doi: 10.36941/jesr-2024-0167.

[39] E. Dorsah, A. K. Ampah-Mensah, R. Asumadu, T. A. Achido, and C. Ayew-Sam, "Modeling Pre-service Management Teachers' Professional Knowledge and Self-efficacy in Teaching: An Implication Study for Quality Education," *International Journal of Educational Reform*, May 2026, doi: 10.1177/10567879261448917.

[40] R. Sutarmina, J. Costley, A. Gorbunova, and C. Lange, "Examining the relationship between self-regulated effort and maintained situational interest in an online learning environment," *Educ. Inf. Technol. (Dordr).*, vol. 30, no. 3, pp. 3443–3462, Feb. 2025, doi: 10.1007/s10639-024-12920-8.

[41] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988. doi: 10.4324/9780203771587.