UDC 37.014.6:005.963

A METHODOLOGICAL FRAMEWORK FOR ORGANIZING PILOT-EXPERIMENTAL RESEARCH ON SOFT-COMPETENCY DEVELOPMENT AMONG LEADERSHIP PERSONNEL IN IN-SERVICE PROFESSIONAL DEVELOPMENT

Akhrorova Umida Khamidullayevna
Doctor of Philosophy (PhD) in Pedagogical Sciences,
Associate Professor, Department of Social Sciences,
University of Business and Science
E-mail: umiaxrorhamid@gmail.com
ORCID ID: 0009-0008-9086-4879

ABSTRACT. In-service professional development initiatives for leadership personnel increasingly seek to strengthen communication, emotional intelligence, and adaptive self-management; however, the reported gains are not always examined through a research design that can support a credible causal interpretation. To address this methodological gap, the article develops and substantiates an integrated procedure for pilot-experimental (quasi-experimental) investigation of soft-competency development in in-service settings, where administrative arrangements and ethical considerations commonly make full randomization impracticable. The study employs a theoretical-conceptual approach that combines comparative-typological analysis of methodological scholarship with structural modeling, bringing together classical quasi-experimental design principles, multi-level training evaluation, and psychometric reliability into one reproducible procedure. The resulting framework proposes a nonequivalent control-group design incorporating pretest, post-test, and follow-up measurements, with groups formed through matching; a triangulated assessment battery integrating self-report, 360-degree feedback, situational-judgement tasks, and structured expert observation; clearly defined reliability and validity procedures, including internal-consistency and test–retest assessment; a sequence consisting of baseline diagnosis, intervention, immediate post-intervention assessment, and delayed follow-up; a statistical plan that combines parametric and non-parametric comparisons within and across groups and requires effect-size reporting; together with an explicit correspondence between threats to internal validity and the safeguards used to reduce them. The article further specifies the ethical procedures, including a waiting-list control design, required for responsible implementation. The framework provides corporate universities, in-service training institutions, and human-resource development researchers with a reproducible and empirically defensible basis for evaluating soft-competency interventions, while also identifying priorities for future multi-site and cross-cultural validation.

Keywords: *pilot-experimental methodology; quasi-experimental design; soft competencies; leadership personnel; in-service professional development; training evaluation; reliability and validity; control group design*

INTRODUCTION. The expansion of in-service professional development (advanced training, retraining, and continuing professional education) as the principal mechanism for developing the soft competencies of leadership personnel has not been matched by an equivalent expansion in the rigor with which such programs are evaluated. Institutions frequently report that a training program "improved communication skills" or "strengthened leadership competencies" without specifying a control condition, a validated instrument, or a statistical procedure capable of distinguishing genuine development from measurement error, maturation, or the general effect of participating in any structured activity (Campbell & Stanley, 1963). This is not a peripheral methodological nicety: without a defensible pilot-experimental methodology, institutions cannot determine which components of a program are responsible for observed change, cannot compare competing program designs on a common evidentiary basis, and cannot justify continued investment in soft-competency development to organizational decision-makers on grounds stronger than anecdote (Hamzah et al., 2025).

The methodological difficulty is compounded by the specific conditions of in-service professional development for leadership personnel, which differ from both the controlled laboratory setting in which classical experimental design was elaborated and the degree-granting higher-education setting in which much educational-measurement methodology has been developed and tested. Participants cannot typically be randomly assigned to conditions, because enrollment in in-service training follows administrative scheduling, role-based eligibility, or voluntary registration rather than a randomization protocol; withdrawing a leader from participation purely for research purposes raises both organizational and ethical objections; and the "soft" nature of the competencies under study — communicative, socio-emotional, adaptive — makes measurement more vulnerable to social-desirability bias and rater subjectivity than the measurement of technical knowledge (Robles, 2012; Klaus, 2008). These conditions mean that the classical fully randomized controlled trial is rarely available, and institutions instead require a rigorously specified quasi-experimental alternative (Campbell & Stanley, 1963) adapted specifically to this setting.

Accordingly, the purpose of this article is to construct, substantiate, and operationalize such a methodology. It draws on three established but rarely integrated methodological traditions — quasi-experimental design theory (Campbell & Stanley, 1963), multi-level training-evaluation theory (Kirkpatrick & Kirkpatrick, 2006), and psychometric reliability theory (Cronbach, 1951) — and combines them into a single, replicable procedure for organizing pilot-experimental work on soft-competency development among leadership personnel in in-service professional development settings. The study is guided by three research questions: (RQ1) What quasi-experimental design is appropriate for testing soft-

competency interventions in in-service professional development, given the organizational and ethical constraints that preclude full randomization? (RQ2) What diagnostic instrumentation and reliability/validity procedures are required for soft-competency measurement to support a defensible causal inference in this setting? (RQ3) What statistical analysis plan and validity-threat safeguards are needed to interpret pilot-experimental results responsibly? The remainder of the article reviews the relevant methodological literature, describes the conceptual-analytical method used to construct the proposed methodology, presents the methodology itself, and discusses its implications, limitations, and directions for further validation.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK.

Competence diagnostics in leadership development. Any pilot-experimental methodology for soft-competency development presupposes a defensible account of what is being measured. Competence theory (Raven, 1984; Boyatzis, 1982; Spencer & Spencer, 1993) treats competence as an integrated set of knowledge, skill, motivation, and value, of which the visible, easily measured layer (knowledge, skill) is only a partial indicator of the deeper, harder-to-observe layer (self-concept, traits, motives) that most strongly predicts superior performance. Zimnyaya (2004) and Khutorskoy (2003) further specify a taxonomy in which competencies governing a person's relations with others form a distinct, measurable cluster. For pilot-experimental purposes, this literature establishes two requirements that recur throughout the methodology developed below: soft competencies must be operationalized into observable, ratable indicators before they can be diagnosed, and because their deeper determinants are not directly observable, diagnosis should not rely on a single measurement method.

Quasi-experimental design theory. Campbell and Stanley's (1963) classical treatment of experimental and quasi-experimental design remains the foundational reference for research conducted in field settings where full randomization is not achievable. Their central methodological contribution — of direct relevance to in-service professional development — is the nonequivalent control-group design, in which an experimental and a comparison group are both measured before and after an intervention but are not formed through random assignment, together with an explicit taxonomy of threats to internal validity that such a design must anticipate and control: history (events unrelated to the intervention that occur during the study period and could account for change), maturation (change that would have occurred regardless of the intervention), testing (the effect of taking a diagnostic instrument on subsequent scores, independent of any intervention), instrumentation (change attributable to the measuring instrument or the observers rather than to the participants), statistical regression (the tendency of extreme initial scores to move toward the mean on retesting), selection (systematic pre-existing differences between groups), and mortality or attrition (systematic loss of participants that alters group composition over time). Because randomization is typically unavailable in in-service leadership training, this taxonomy —

rather than the assumption of randomization itself — is what a defensible pilot-experimental methodology in this domain must be built around.

Multi-level training evaluation. Kirkpatrick and Kirkpatrick's (2006) four-level model — reaction, learning, behavior, and results — provides the evaluative structure within which diagnostic measurement should be organized. Reaction captures participants' immediate response to the intervention; learning captures the extent to which targeted competencies have measurably changed; behavior captures whether that change is observable in actual on-the-job conduct after the intervention has ended; and results captures whether the change is associated with organizationally meaningful outcomes. Applied to soft-competency development, this model implies that a pilot-experimental methodology confined to an immediate post-test measures learning at best, and says nothing about behavior or results — a limitation that recent training-transfer research shows is consequential, since a substantial proportion of soft-skill gains measured immediately after training do not persist into sustained workplace behavior unless the evaluation design, and the intervention itself, explicitly build in delayed, in-context reinforcement and measurement (Hamzah et al., 2025).

Reliability and measurement quality. Because soft-competency diagnosis relies heavily on self-report and observer rating, both of which are more susceptible to inconsistency than objective performance tests, the reliability of the diagnostic instrument is a precondition for any subsequent causal claim. Cronbach's (1951) coefficient alpha remains the standard index of a scale's internal consistency, and its systematic reporting — alongside test–retest reliability for the same instrument administered on separate occasions — is treated in the methodology below as a mandatory, not optional, component of instrument validation prior to use in a pilot-experimental study.

Recent empirical precedents. Recent applied research offers concrete precedent for the kind of design integration this article proposes. Lapian, Zulkifli, Razak, and Sidin (2022) implemented a quasi-experimental pretest–posttest design with an experimental and a comparison group of hospital nurses to test whether self-leadership training combined with emotional-intelligence mentoring reduced burnout, demonstrating the feasibility of a nonequivalent control-group design for a "soft," self-report-dependent outcome in a real organizational setting where randomization was not implemented. Hamzah et al. (2025), in a systematic scoping review of soft-skills training-transfer research, show that evaluation designs limited to immediate post-training measurement systematically overstate durable competency change relative to designs that include a delayed follow-up measurement — a finding that directly motivates the follow-up stage built into the methodology proposed in Section 4. Taken together, this literature demonstrates that the individual components required — nonequivalent control-group design, multi-level evaluation, reliability testing, and delayed follow-up — are each independently established, but existing studies typically apply only a subset of them; no widely used, explicitly integrated procedure specifies how all four should be combined for soft-competency development in leadership in-service training specifically. This is the gap the present article addresses.

MATERIALS AND METHODS. This article is a theoretical-methodological study: its object is not a single empirical dataset but the methodological literature itself, and its output is a proposed, internally justified research procedure rather than a report of new primary findings. The method employed combines comparative-typological analysis with structural modeling. The comparative-typological analysis proceeded through a narrative synthesis of the methodological literature on quasi-experimental design, training evaluation, and psychometric reliability (Campbell & Stanley, 1963; Cronbach, 1951; Kirkpatrick & Kirkpatrick, 2006), competence diagnostics (Raven, 1984; Boyatzis, 1982; Spencer & Spencer, 1993; Zimnyaya, 2004; Khutorskoy, 2003), and recent applied quasi-experimental and training-transfer studies (2022–2025) retrieved through conceptual database searches combining the terms "quasi-experimental," "soft skills," "leadership training," and "pilot study" in sources indexed by Scopus, Web of Science, and PubMed/MEDLINE, together with citation-tracing from the foundational methodological texts. Each source's treatment of design structure, sampling, instrumentation, reliability procedure, and statistical analysis was extracted and compared against the requirements implied by the three research questions stated in Section 1.

The structural modeling method was then used to integrate the extracted design elements into the single procedure presented in Section 4: for each stage of a pilot-experimental study (design selection, sampling, instrumentation, procedure, statistical analysis, and validity-threat control), the literature's most methodologically defensible option compatible with the organizational and ethical constraints of in-service leadership training (Section 1) was selected and specified in operational terms. It should be stated explicitly that the resulting methodology is offered as a validated synthesis of established methodological components rather than as a novel statistical technique; its contribution lies in the integration and operational specification of these components for a setting — corporate and organizational in-service leadership training — in which they have not previously been combined into a single explicit procedure. The methodology's own effectiveness in producing valid causal inferences, once applied in a specific institutional setting, remains an empirical question that the methodology is designed to make answerable, and Section 5.4 outlines the corresponding validation agenda.

RESULTS: The Proposed Pilot-Experimental Methodology. Design. The proposed framework adopts a nonequivalent control-group pretest–posttest–follow-up design (Campbell & Stanley, 1963), selected specifically because it does not require random assignment, which is typically unavailable in in-service professional development. An experimental group receives the soft-competency development intervention under evaluation; a comparison group, drawn from personnel enrolled in the same or a closely comparable professional development cycle but not receiving the intervention (or receiving the institution's standard, non-intervention program), is measured on the same schedule. Both groups are diagnosed before the intervention (baseline), immediately after it (post-test), and again after a defined interval, recommended at three to six months (delayed follow-up),

directly operationalizing the behavior level of Kirkpatrick and Kirkpatrick's (2006) model rather than relying on the learning level alone.

Sampling and group formation. Since participants cannot normally be assigned randomly, approximate group equivalence is established through a matching procedure. The methodology specifies that experimental and comparison groups be matched, prior to the intervention, on baseline diagnostic scores, managerial level, tenure in a leadership role, and, where organizationally relevant, functional area, with any residual baseline difference treated as a covariate in the subsequent statistical analysis rather than assumed away. A minimum of twenty to twenty-five participants per group is recommended as a practical lower bound for the non-parametric and parametric between-group tests specified in Section 4.6, while recognizing that many in-service training cohorts will be smaller, in which case the methodology directs that results be reported as an internally valid pilot with explicit statistical-power limitations rather than as an externally generalizable finding.

Diagnostic instrumentation. Following the argument in Section 2.1 that soft competencies should be assessed through more than one source of evidence, the framework uses a four-part triangulated diagnostic battery. A self-assessment questionnaire, using Likert-type items mapped onto the target competency components, captures the participant's own perception of change. A 360-degree feedback instrument, completed by the participant's supervisor, peers, and, where applicable, direct reports, provides an external, multi-source corrective to the self-report data and is particularly important given the social-desirability risk inherent in self-assessed soft skills. A situational judgement task presents realistic managerial scenarios and asks participants to select or rank response options, providing a performance-based measure less susceptible to self-report bias than questionnaire items. Structured expert observation, conducted by a trained rater using a standardized observation protocol during a real or simulated managerial task, provides a fourth, behaviorally anchored data source. Convergence across at least three of these four sources is treated as the criterion for concluding that an observed change reflects genuine competency development rather than an artifact of a single measurement method.

Reliability and validity procedures. Before the main pilot-experimental study begins, each component of the diagnostic battery is subjected to a separate validation procedure. Internal consistency is assessed using Cronbach's (1951) coefficient alpha, with a threshold of 0.70 treated as the minimum acceptable level for retaining a scale in its current form; subscales falling below this threshold are revised or removed prior to the main study. Test–retest reliability is assessed by administering the instrument to a small independent group on two occasions separated by approximately two weeks and computing the correlation between administrations, providing a check on temporal stability independent of any intervention effect. Content validity is established through review by a panel of at least three subject-matter experts (in pedagogy, organizational psychology, or human-resource development) who rate each item's relevance to the competency it is intended to measure; items with low panel agreement are revised. Where an instrument originates in a different

linguistic or cultural context, a forward–backward translation and small-sample cultural-adaptation pilot precede its use, since items validated in one organizational or national culture cannot be assumed to retain the same psychometric properties elsewhere.

Procedure and timeline. The procedure is organized into four consecutive stages. In the preparatory stage, groups are formed and matched, instruments are validated as described in Section 4.4, and informed consent is obtained. In the baseline stage, both groups complete the full diagnostic battery before the intervention begins. In the intervention stage, the experimental group undergoes the soft-competency development intervention under evaluation (for example, a structured training-mentoring-coaching sequence), while the comparison group continues its standard professional development activity without the intervention component being tested; immediately upon completion, both groups again complete the diagnostic battery (post-test). In the follow-up stage, three to six months after the intervention concludes, both groups are diagnosed a third time, and, where feasible, supplementary organizational indicators (for example, 360-degree feedback scores already collected through routine institutional processes, or supervisor-reported behavioral indicators) are reviewed to approximate Kirkpatrick and Kirkpatrick's (2006) results level.

Statistical analysis plan. The statistical plan separates within-group change from between-group differences and specifies parametric and non-parametric alternatives so that the choice can be made on the basis of the obtained data's distributional properties, checked using a normality test (for example, Shapiro–Wilk) prior to analysis. Within-group pre–post change is assessed using the paired-samples t-test where normality holds and the Wilcoxon signed-rank test otherwise. Between-group differences in the magnitude of change (typically operationalized as post-test score adjusted for baseline, or as a difference score) are assessed using the independent-samples t-test where normality and variance-homogeneity assumptions hold and the Mann-Whitney U test otherwise. Effect sizes (Cohen's d for parametric comparisons, or a rank-based equivalent for non-parametric comparisons) are reported alongside significance tests in every case, since statistical significance alone, particularly with the modest sample sizes typical of in-service training cohorts, does not establish the practical magnitude of an observed effect. Where baseline differences between groups remain after matching, analysis of covariance is used in preference to a simple between-group comparison, with the baseline score entered as the covariate.

Mapping validity threats to procedural safeguards. In accordance with Campbell and Stanley's (1963) framework, each identified threat to internal validity is linked to a specific procedural safeguard rather than being handled only descriptively. The threat of history is addressed by running the experimental and comparison groups concurrently, within the same organizational period, so that external events affect both groups equally. The threat of maturation is addressed by the presence of the comparison group itself, since naturally occurring change over time should appear in both groups and can be subtracted out. The threat of testing is addressed by using parallel or counterbalanced instrument versions across administrations where feasible, and, at minimum, by applying the same testing schedule to

both groups so that any testing effect is common to both. The threat of instrumentation is addressed by using standardized rating protocols and training all observers and 360-degree raters to a documented inter-rater reliability standard before data collection begins. The threat of statistical regression is addressed by the matching procedure in Section 4.2, which reduces reliance on extreme baseline scores as the sole basis for group assignment. The threat of selection is addressed by the baseline-equivalence matching itself and by treating residual baseline differences as covariates rather than assuming their absence. The threat of mortality is addressed by recording and statistically comparing the baseline characteristics of participants who withdraw against those who complete the study, so that differential attrition, if present, can be identified and reported rather than silently biasing the results.

Ethical procedures. The methodology requires informed, voluntary consent, with an explicit statement that non-participation or withdrawal carries no professional consequence; confidentiality of individual diagnostic results, with only aggregated or de-identified data used in institutional reporting; and, in place of a comparison group that receives nothing, a waiting-list design in which the comparison group is offered the same intervention after the follow-up measurement is complete, addressing the ethical concern that would otherwise arise from withholding a potentially beneficial developmental intervention from any participant indefinitely.

DISCUSSION. The proposed framework provides a direct methodological response to the three research questions introduced in Section 1. In answer to RQ1, it specifies the nonequivalent control-group pretest–posttest–follow-up design as the appropriate quasi-experimental alternative to randomization, matched rather than randomly assigned groups, and a validity-threat safeguard mapped to each element of Campbell and Stanley's (1963) taxonomy. In answer to RQ2, it specifies a four-instrument triangulated diagnostic battery together with mandatory internal-consistency, test–retest, and content-validity procedures prior to use. In answer to RQ3, it specifies a statistical analysis plan combining parametric and non-parametric within- and between-group tests, mandatory effect-size reporting, and covariate adjustment for residual baseline imbalance.

The methodology extends the applied precedent set by Lapian et al. (2022) by generalizing a nonequivalent control-group design, there applied to nurse burnout outcomes, to the broader class of soft-competency outcomes relevant to leadership personnel, and by adding the explicit reliability-validation sub-study (Section 4.4) that is often assumed rather than reported in applied quasi-experimental training research. It responds to the durability concern raised by Hamzah et al. (2025) by building a delayed follow-up measurement into the design itself (Section 4.5) rather than treating durability as a separate question to be investigated after the fact. And it operationalizes Kirkpatrick and Kirkpatrick's (2006) four-level framework concretely, rather than invoking it only as a general evaluative aspiration, by specifying which stage of the procedure corresponds to which evaluative level.

From a practical standpoint, the methodology offers corporate universities, in-service training institutes, and human-resource development researchers a procedure that can be

applied directly, with institution-specific adaptation limited mainly to the content of the intervention itself and the precise items of the diagnostic instruments, both of which remain outside the scope of a general-purpose methodology and are properly specified at the level of an individual empirical study.

Several limitations require explicit consideration. The methodology is quasi-experimental rather than randomized, and even with the safeguards specified in Section 4.7, causal inference from a nonequivalent control-group design remains inherently weaker than that available from full randomization; where organizational conditions permit randomization — for example, when demand for a program exceeds available places and places can be allocated by lottery — a randomized design should be preferred over the quasi-experimental alternative specified here. The recommended sample sizes (Section 4.2) support pilot-level rather than large-scale confirmatory inference, and results obtained through the methodology should be interpreted as internally valid within the institution in which they were obtained rather than as automatically generalizable to other organizational or cultural settings, a concern reinforced by the culture-dependence of soft-competency measurement noted in Section 4.4. Finally, as a theoretical-methodological article, this study has not yet applied the proposed methodology to a specific empirical case; doing so, including in Central Asian corporate and public-sector in-service training institutions where such methodology is currently underdeveloped, is the immediate priority for subsequent research, together with multi-site replication and, where feasible, comparison against a randomized variant of the same design.

CONCLUSION. This study has developed and substantiated an integrated methodology for organizing pilot-experimental work on the development of leadership personnel's soft competencies within in-service professional development, a setting in which such work has commonly been reported without a design capable of supporting a defensible causal claim. By combining a nonequivalent control-group pretest–posttest–follow-up design, matched rather than randomized group formation, a triangulated and psychometrically validated diagnostic battery, an explicit statistical analysis plan, and a systematic mapping of internal-validity threats to procedural safeguards, the methodology gives institutions a replicable, ethically grounded procedure for evaluating soft-competency interventions rather than relying on unverified claims of effectiveness. While the methodology itself now requires empirical application and validation across institutional and cultural settings, including within Central Asian corporate higher education and in-service training systems, it provides a theoretically grounded and immediately usable starting point for that work.

REFERENCES:

1. Boyatzis, R. E. (1982). The competent manager: A model for effective performance. John Wiley & Sons.
2. Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Rand McNally.
3. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555
4. Hamzah, H. A., Marcinko, A. J., Stephens, B., & Weick, M. (2025). Making soft skills "stick": A systematic scoping review and integrated training transfer framework grounded in behavioural science. European Journal of Work and Organizational Psychology. https://doi.org/10.1080/1359432X.2024.2376909
5. Khutorskoy, A. V. (2003). Key competences as a component of the personality-oriented paradigm of education. Narodnoye Obrazovaniye, (2), 58–64.
6. Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). Evaluating training programs: The four levels (3rd ed.). Berrett-Koehler.
7. Klaus, P. (2008). The hard truth about soft skills: Workplace lessons smart people wish they'd learned sooner. HarperCollins.
8. Lapian, L. G., Zulkifli, A., Razak, A., & Sidin, I. (2022). A quasi-experimental study: Can self-leadership training and emotional intelligence mentoring lower burnout rates in hospital nurses? Open Access Macedonian Journal of Medical Sciences, 10(E), 905–912. https://doi.org/10.3889/oamjms.2022.8756
9. Raven, J. (1984). Competence in modern society: Its identification, development and release. H. K. Lewis.
10. Robles, M. M. (2012). Executive perceptions of the top 10 soft skills needed in today's workplace. Business Communication Quarterly, 75(4), 453–465.
11. Spencer, L. M., & Spencer, S. M. (1993). Competence at work: Models for superior performance. John Wiley & Sons.
12. Zimnyaya, I. A. (2004). Key competences as a result-target basis of the competence-based approach in education. Research Center for the Quality of Specialist Training Problems.