

Article

Context-Sensitive Team Formation in Engineering Education: Structural Optimization of TREO-Based Assignment Across Sections with Contrasting Role Diversity Profiles

Yasmany García-Ramírez *  and Pablo Campoverde-Guerrero 

Department of Civil Engineering, Universidad Técnica Particular de Loja, Loja 110101, Ecuador

* Correspondence: ydgarcia1@utpl.edu.ec

Abstract

Effective team formation strategies must account for baseline behavioral diversity within student populations, yet most approaches apply uniform parameters regardless of section characteristics. This exploratory, quasi-experimental study proposes a context-sensitive framework for Team Role Experience and Orientation (TREO)-based team formation across two sections of a road geometric design course ($N = 77$ civil engineering undergraduates). We systematically evaluated 80 team formation configurations per section (16 thresholds $\times$ 5 group sizes) to maximize structural differentiation between role-balanced and role-redundant teams. Baseline diagnostics revealed substantial differences in role diversity, justifying independent optimization. Section A (role-redundant) achieved strong structural separation, whereas Section B (role-diverse) showed weak or negative separation across all configurations, reflecting limitations in achieving consistent complementarity contrasts. Post-project surveys assessed perceived role complementarity, team dynamics, and performance. Across both sections, perceptual differences between conditions were small and non-significant, indicating that structural role differentiation does not automatically translate into subjective team experience. However, perceptual trends varied systematically by baseline diversity, highlighting the moderating role of section composition. These findings emphasize the importance of diagnostic indices (R_{ef} , IDR) for selecting and interpreting team formation strategies and support the use of adaptive, context-sensitive approaches in engineering education.

Keywords: team formation optimization; TREO survey; engineering education; role complementarity; section diversity; context-sensitive pedagogy; civil engineering



Academic Editor: Jennifer Amos

Received: 20 January 2026

Revised: 18 February 2026

Accepted: 24 February 2026

Published: 26 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Collaborative learning has become a cornerstone of modern engineering education, reflecting the increasingly interdisciplinary and team-based nature of professional practice in fields such as civil engineering (Hadgraft & Smith, 2008; Van den Beemt et al., 2020). Research consistently demonstrates that well-structured teamwork enhances not only technical competencies but also critical soft skills including communication, leadership, and conflict resolution, all of which are essential for future engineers (Johri, 2023). However, the effectiveness of collaborative learning is highly dependent on how teams are formed, as poor group composition can lead to unequal workload distribution, interpersonal conflicts, and suboptimal learning outcomes (Opdecam & Everaert, 2018). Traditional approaches to team formation—such as random assignment, self-selection, or instructor intuition—often

fail to account for the cognitive and behavioral diversity necessary for high-performing teams, particularly in complex technical courses (Dzvonyar et al., 2018). This gap between the potential benefits of collaborative learning and the actual outcomes achieved underscores the need for evidence-based team formation strategies that systematically optimize group composition (Burris-Melville et al., 2024; Fiegel & Denatale, 2011; Karimi & Manteufel, 2020).

One promising framework for addressing this challenge is the Team Role Experience and Orientation (TREO) survey, which conceptualizes team effectiveness through the lens of six distinct behavioral roles: Organizer, Doer, Challenger, Innovator, Team Builder, and Connector (Mathieu et al., 2015). Unlike personality-based assessments, TREO focuses on task-related behaviors and role preferences, making it particularly suitable for project-based learning environments where concrete deliverables and deadlines are central (Fausett et al., 2021). The TREO framework posits that teams benefit from role complementarity—the presence of diverse roles that cover different aspects of team functioning—and suffer from role overlap or redundancy, where multiple members compete for the same functions (Fausett et al., 2021; García-Ramírez, 2025). Research on algorithmic team formation in educational settings has shown that intentional role balancing—whether based on personality traits, skill diversity, or behavioral preferences—can lead to improved team outcomes compared to random assignment or self-selection (Borrego et al., 2013; Oakley et al., 2004). However, most existing studies have focused on general academic contexts, and there is limited evidence regarding the effectiveness of TREO-based team formation in highly technical, applied disciplines such as civil engineering.

Civil engineering courses, particularly those centered on road geometric design, present unique challenges for team-based learning due to their integration of analytical rigor, regulatory constraints, and spatial reasoning (Castro, 2012; García-Ramírez, 2024). A typical road design project requires students to synthesize knowledge from traffic analysis, route alignment, horizontal and vertical geometric design, and consistency evaluation—tasks that demand both systematic planning and creative problem-solving (Garrido de la Torre et al., 2020; Kurtis Younkin, 2017; Pérez-Acebo, 2017). The multifaceted nature of these projects creates an ideal testing ground for role-based team formation strategies, as different phases of the design process may activate different TREO roles: Organizers may excel at project planning and timeline management, Doers at executing calculations and technical drawings, Challengers at critically reviewing design alternatives, Innovators at proposing creative solutions to constraints, Team Builders at maintaining group cohesion, and Connectors at integrating diverse contributions into a coherent final product.

Building upon this context, the effectiveness of team-based learning in civil engineering depends largely on how teams are composed and how individual strengths are aligned with project demands. Researchers have proposed several methodologies aimed at optimizing team composition and strengthening collaborative skills (Michalaka & Golub, 2016). Among these efforts, specialized teamwork skill scales have been developed to assess and enhance key competencies such as communication, cooperation, and leadership (Han, 2020; Paoletti et al., 2020). Furthermore, identifying distinct functional roles within teams has proven to be a critical factor in predicting successful group performance (Curşeu & Pluut, 2013). Despite this theoretical foundation, no prior research has systematically explored whether TREO-based team formation strategies can improve learning outcomes and team dynamics in civil engineering education, particularly within courses characterized by high technical complexity and real-world applicability.

A critical methodological consideration in implementing TREO-based team formation lies in the operational definition of strong roles. The TREO assessment generates continuous scores for each of the six roles; however, translating these scores into action-

able grouping criteria requires establishing a threshold (T) that determines when a role is considered strong or dominant for a given individual. Previous research (Salas et al., 2008) demonstrated that clearly defined leadership roles within student teams can improve project efficiency by 15%. Despite this, most TREO implementations have relied on thresholds without empirical validation—for example, classifying scores of 4.2–5.0 as High, 3.3–4.1 as Medium, and 1.0–3.2 as Low (The Group for Organizational Effectiveness, 2012). The selection of such thresholds directly affects the perceived diversity within the student population and, consequently, influences the optimization algorithm’s ability to generate balanced or overlapping teams (Boyce-Jacino & Harrington, 2022).

Another unresolved methodological issue concerns the optimal group size (N) that maximizes role complementarity while minimizing coordination costs. Competing theoretical perspectives suggest that smaller teams ($n_g = 3\text{--}4$) may promote accountability and communication—consistent with findings that report up to a 20% increase in engagement for teams of this size (Swanson et al., 2019)—whereas larger teams ($n_g = 5\text{--}7$) may enhance role diversity but risk social loafing. A systematic exploration of the $T \times n_g$ parameter space, supported by quantitative measures of role diversity such as Shannon’s effective diversity index (R_{ef}), could provide a more robust empirical foundation for TREO-based interventions and address a significant gap in the current literature.

An additional complication arises from the inherent diversity of student sections, which may vary substantially in their role distributions even when drawn from the same educational program. For instance, one class section might exhibit high role diversity with students distributed relatively evenly across all six TREO roles (high R_{ef}), while another section might show clustering around certain roles with limited representation of others (low R_{ef}). This variation has important implications for team formation: in a highly diverse section, it may be feasible to create groups with minimal role overlap and maximum complementarity (the “Balanced” condition), whereas in a more homogeneous section, any grouping strategy may inevitably produce some redundancy (the “Unbalanced” condition). Existing TREO studies have rarely accounted for these baseline differences between sections, raising the question of whether findings from one educational context can generalize to others. By analyzing two sections of the same course with potentially different role diversity profiles, and by independently optimizing team formation parameters ($T_A, n_{g,A}$ for Section A; $T_B, n_{g,B}$ for Section B), this study addresses the need for context-sensitive team formation strategies that adapt to the specific characteristics of each student population.

The primary objective of this study was to develop and validate a dual-optimization framework for TREO-based team formation in a road geometric design course for civil engineering undergraduates, and to empirically test whether algorithmically balanced teams (high complementarity, low overlap) outperform teams with intentional role redundancy (low complementarity, high overlap) in terms of perceived complementarity, team dynamics, and performance outcomes. To achieve this objective, we implemented a quasi-experimental design across two sections of the same course, a systematic exploration of the $T \times n$ parameter space, supported by quantitative measures of role diversity such as Shannon’s effective diversity index (R_{ef}), could provide a more robust empirical foundation for TREO-based interventions. These structural assessments are later operationalized through three group-level metrics—the Complementarity Index (CI), which captures the dispersion of role strengths across team members; the Overlap Index (OI), which measures role redundancy by quantifying how many members share the same primary role; and Role Diversity (RD), which counts the number of distinct primary roles represented within a team. Full formal definitions and computational procedures for these metrics are provided in Section 2.6.

This study makes several contributions to the literature on team-based learning in engineering education and to the practical application of TREO in educational settings. First, it provides empirical evidence for the effectiveness of role-based team formation in a highly technical, applied engineering course. Second, it introduces a dual-optimization methodology that independently tailors team formation parameters to the specific diversity profile of each student section, addressing the problem of one-size-fits-all approaches in heterogeneous educational environments. Third, it operationalizes and validates quantitative metrics of role diversity (Shannon's H, effective diversity, and the Diversity Index) that can serve as diagnostic tools for educators to assess whether their student populations are conducive to balanced team formation. Fourth, it demonstrates the importance of the Separation Score (S) as a criterion for evaluating the quality of team formation algorithms, offering a methodological advance over studies that rely solely on post hoc outcome measures. Finally, by collecting end-of-semester survey data on perceived complementarity, team dynamics, and performance across both balanced and unbalanced conditions in two independent sections, this study provides actionable insights for civil engineering educators seeking to implement evidence-based team formation strategies in their own courses, ultimately contributing to improved learning outcomes and more equitable collaborative experiences for engineering students.

2. Materials and Methods

This study was not preregistered and is therefore framed as exploratory rather than confirmatory, a choice that reflects the novelty of applying TREO-based team formation in civil engineering education, where no prior benchmarks exist for expected effect sizes, optimal parameters, or perceptual outcomes. The absence of preregistration introduces well-documented risks including analytical flexibility, selective reporting, and potential overfitting of optimization parameters to sample-specific characteristics (Nosek et al., 2018; Simmons et al., 2011).

To mitigate these risks while preserving the exploratory nature essential to methodological innovation, we implemented several safeguards: (1) Complete transparency: all 80 configurations per section are reported in Appendix A, including null and negative results, preventing selective reporting of favorable outcomes. (2) Algorithmic objectivity: the dual-optimization procedure was fully automated and criterion-based (maximizing Separation Score), eliminating researcher degrees of freedom in parameter selection. (3) Convergent validation: optimal configurations were independently validated through multiple compositional metrics (CI, OI, RD) rather than a single outcome. (4) Effect size emphasis: interpretation prioritizes effect sizes and confidence intervals over *p*-values, reducing dependence on arbitrary significance thresholds. (5) Explicit limitation acknowledgment: we clearly state that survey constructs and analytic strategies were refined iteratively in response to observed data patterns, and we call for independent replication. While these measures cannot fully eliminate the limitations inherent to exploratory research, they provide a transparent foundation for hypothesis generation and methodological refinement in future confirmatory studies.

2.1. Study Design and Context

This quasi-experimental study was conducted during the October 2024–February 2025 academic semester in the Road Construction I course, a core requirement for civil engineering undergraduates at Universidad Técnica Particular de Loja (UTPL), Ecuador. The course curriculum encompassed traffic analysis, route alignment, horizontal geometric design, and consistency evaluation of road projects (See details in Appendix A). Two sections were included in the study: Section A (N = 43 students) and Section B

(N = 34 students), taught by the same instructor using identical syllabi, learning objectives, and assessment criteria. Students in both sections were required to complete a semester-long road geometric design project in teams, simulating real-world professional practice. The project required integration of multiple technical competencies, including data analysis, CAD drafting, regulatory compliance (using Ecuadorian and AASHTO standards), and technical report writing.

No personally identifiable information was collected or stored during the study. All data were analyzed in an anonymized and aggregated form, in compliance with Ecuador's current data protection regulations. Under these conditions, formal approval from the institutional ethics committee was not required.

The sample size (N = 77 students across 18 teams) was necessarily modest, particularly for team-level analyses stratified by section, which limits the sensitivity of the study to detect subtle perceptual differences between conditions. While this scale was sufficient to validate the intended compositional manipulations at the structural level, it provides only a preliminary basis for examining downstream perceptual outcomes. Consequently, the absence of statistically robust differences in team perceptions should not be interpreted as evidence that compositional effects are absent, but rather that any such effects are likely modest and require larger samples to be reliably detected.

2.2. TREO Assessment and Role Profiling

At the beginning of the semester (Week 1), all students completed the Team Role Experience and Orientation (TREO) survey, a validated psychometric instrument designed to assess preferences and strengths across six behavioral roles. The TREO survey consists of 36 items (6 items per role) rated on a 5-point Likert scale, where 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, and 5 = Strongly Agree. Scores range from 1.0 (no affinity for the role) to 5.0 (strong identification with the role). A sample item for the Doer role reads: "I am the person in the group who makes sure things actually get done".

The six TREO roles represent distinct behavioral orientations in team settings. The Organizer is oriented toward planning, scheduling, and maintaining structure. The Doer focuses on execution, acting on tasks and delivering concrete outputs. The Challenger questions assumptions, critically evaluates ideas, and pushes the team to consider alternatives. The Innovator generates creative ideas and novel approaches to problems. The Team Builder attends to interpersonal dynamics, cohesion, and morale. Finally, the Connector integrates diverse contributions, facilitates communication across the group, and ensures that the work of individual members is synthesized into a coherent whole (Fausett et al., 2021; Mathieu et al., 2015). Understanding these role definitions is essential for interpreting the compositional metrics and algorithmic procedures described in the following sections.

Raw TREO scores were standardized into z-scores using the grand mean and standard deviation across both sections combined (N = 77 students) to ensure comparability and to facilitate algorithmic optimization. Specifically, for each student j and each role r , the standardized score was computed as:

$$z_{rj} = \frac{(X_{rj} - \mu_r)}{\sigma_r} \quad (1)$$

where X_{rj} denotes the raw TREO score of student j for role r , and μ_r and σ_r represent the mean and standard deviation of raw scores for role r calculated across the full sample. This standardization was necessary to remove scale effects and inter-role variance, allowing meaningful comparison of relative role strengths across students and sections, and preventing roles with higher raw score dispersion from disproportionately influencing the

optimization process. For each student, we then computed mean z-scores for each role. Each student's Primary Role was determined as the role with the highest raw score (not z-score) to preserve the original assessment structure, and the Primary Role Score was defined as the maximum raw score across all six roles.

2.3. Role Diversity Diagnostics

To assess the baseline role diversity of each section and inform team formation feasibility, we computed Shannon's diversity index (H) and its exponential transformation, the effective number of roles (Ref). Shannon's index was calculated as:

$$H = -\sum(p_i \cdot \ln(p_i)) \quad (2)$$

where p_i represents the proportion of students in each Primary Role category. The effective diversity was computed as:

$$R_{\text{ef}} = e^H \quad (3)$$

This metric provides an intuitive interpretation as the "equivalent number of equally abundant roles" in the population. We then computed the Role Diversity Index as:

$$\text{IDR} = \frac{R_{\text{ref}}}{n_{\text{ref}}} \quad (4)$$

where $n_{\text{ref}} = 4$ served as the reference group size for diagnostic purposes. Following theoretical guidelines, sections with $\text{IDR} \geq 1.2$ were classified as "Diverse" (expected positive Separation Scores), $0.8 \leq \text{IDR} < 1.2$ as "Moderate" (risk of low or zero separation), and $\text{IDR} < 0.8$ as "Redundant" (expected negative separation or complementarity collapse). This diagnostic framework provided a priori predictions about the feasibility of creating distinct balanced and unbalanced conditions in each section.

2.4. Dual-Optimization Framework: Threshold and Group Size Selection

A critical methodological innovation of this study was the systematic exploration of the two-dimensional parameter space formed by the role strength threshold (T) and group size (n_g). The threshold T operationally defined a "strong role" as any TREO raw score exceeding T, and each student's number of strong roles ($n_{\text{StrongRoles}}$) was computed as the count of roles with raw scores $> T$. We tested 16 threshold values: $T \in \{0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00, 3.25, 3.50, 3.75, 4.00\}$, spanning from very lenient (most students have multiple strong roles) to very stringent (only top-scoring roles qualify). We tested 5 group sizes: $n_g \in \{3, 4, 5, 6, 7\}$, representing the range commonly used in project-based engineering courses. For each of the 80 configurations ($16 T \times 5 n_g$) per section, we executed the complete team formation algorithm (described below) and computed group-level metrics. The dual-optimization approach recognized that different sections may require different optimal parameters due to their distinct role diversity profiles, and therefore we independently optimized ($T_A, n_{g,A}$) for Section A and ($T_B, n_{g,B}$) for Section B.

2.5. Team Formation Algorithm: Balanced and Unbalanced Conditions

For each (T, n_g) configuration, students in each section were assigned to one of two experimental conditions: balanced or unbalanced. The balanced condition aimed to maximize role complementarity and minimize role overlap, while the unbalanced condition intentionally maximized role redundancy. To ensure comparable sample sizes, we implemented a 50/50 allocation strategy: approximately half of each section's students (rounded down to the nearest complete group) were assigned to balanced condition, and the remainder to unbalanced condition. All algorithms used a fixed random seed (seed = 42) to ensure

reproducibility. For the Balanced condition, we implemented a greedy optimization algorithm that constructed groups iteratively. The first student in each group was selected as the highest-ranked by $n_{\text{StrongRoles}}$ and PrimaryRoleScore . Subsequent members were selected to maximize a balance score, computed as the sum of absolute differences between the candidate's z-scores and the current group mean z-scores across all six roles. Critically, candidates whose Primary Role already existed in the partial group received an infinite penalty (adjusted balance score = $-\infty$), strongly encouraging role diversity. Groups were formed sequentially until all available students for the Balanced condition were assigned, with any remainder students added to the last group.

A key methodological issue arises from the sequential allocation procedure. The balanced condition was formed first from the full pool of students, whereas the unbalanced condition was constructed from the remaining students, introducing a systematic selection bias. Students with more distinctive role profiles (e.g., higher $n_{\text{StrongRoles}}$, higher PrimaryRoleScore , or rarer primary roles) were more likely to be assigned to balanced groups, potentially leaving less differentiated profiles in unbalanced groups. Although this sequence was necessary to ensure sufficient role diversity for constructing complementary teams, it confounds the experimental manipulation: observed differences may reflect both intended compositional structure and unintended differences in individual role differentiation.

We addressed this confound through multiple strategies: (i) Empirical verification: We tested whether mean $n_{\text{StrongRoles}}$ differed between conditions within sections, finding no significant differences (Section A: $p = 0.89$; Section B: $p = 0.81$), suggesting comparable individual role strength across conditions. (ii) Transparent reporting: We explicitly acknowledge this limitation when interpreting null perceptual effects, recognizing that any perceptual advantages of balanced composition may have been attenuated by unintended individual-level differences favoring the unbalanced condition. (iii) Methodological recommendation: We recommend that future studies randomize condition assignment prior to applying grouping algorithms, accepting reduced role coverage in the Balanced condition to eliminate selection bias. This alternative approach would enable cleaner causal inference, though at the cost of weaker compositional manipulation—a trade-off that merits systematic investigation. The present design prioritizes strong structural differentiation to validate the algorithm's technical capacity, accepting confounding as a limitation that subsequent research should resolve through randomized assignment protocols.

For the unbalanced condition, students were sorted by Primary Role, then by $n_{\text{StrongRoles}}$ (descending), then by PrimaryRoleScore (descending). Groups were formed by taking consecutive students from this sorted list, ensuring that students with the same Primary Role were clustered together, thereby maximizing role redundancy. Both algorithms handled remainder students (when total students were not evenly divisible by n_g) by appending them to the last formed group, ensuring all students participated. Importantly, the balanced algorithm used students not yet assigned to any condition, while the unbalanced algorithm used the remaining students after balanced assignment, preventing student overlap between conditions. This sequential allocation (balanced first, then unbalanced from remaining pool) was applied consistently across all 80 configurations for both sections.

For groups with fewer than six members (the most common case, given the optimal group size of $n_g = 4$), full role coverage across all six TREO dimensions was not required or expected. The Balanced condition algorithm prioritized maximizing the diversity of Primary Roles represented in each group (up to $\min(n_g, 6)$ distinct roles) while minimizing role overlap; role omissions were therefore a structural consequence of the group size rather than a deliberate selection. For groups with more than six members—which only occurred as a result of remainder students being appended to the last formed group (e.g., one group

of five in Section A Unbalanced)—role repetition was handled by the algorithm’s infinite penalty for role overlap in the Balanced condition, meaning that a seventh member sharing a Primary Role with an existing member received a maximally penalized balance score and was assigned to this larger group only when no alternative placement was feasible. In the Unbalanced condition, role repetition was intentional by design: students were sorted by Primary Role to create role-redundant clusters, and a group of five simply clustered one additional student sharing the dominant role.

2.6. Group-Level Metrics

For each formed group, we computed three quantitative metrics to characterize team composition. The Complementarity Index (CI) was calculated as the sum of the ranges (maximum minus minimum) across all six role z-scores within the group:

$$CI = \sum_{\text{roles}} (\max(z_{\text{role}}) - \min(z_{\text{role}})) \quad (5)$$

Higher CI values indicate greater dispersion in role strengths within the group, reflecting higher complementarity among members.

The Overlap Index (OI) quantified role redundancy by counting the number of “excess” members sharing the same Primary Role and was computed as:

$$OI = \frac{(\max(\text{count}_{\text{primary_role}}) - 1)}{(n_g - 1)} \quad (6)$$

where $\max(\text{count}_{\text{primary_role}})$ denotes the frequency of the most common Primary Role in the group, while n_g denotes group size (number of members per team). The OI ranges from 0 (all members have different Primary Roles) to 1 (all members share the same Primary Role).

Role Diversity (RD) captures the structural breadth of functional roles represented within a team and was computed as the number of unique Primary Roles present among group members, irrespective of role strength or frequency:

$$RD = |\text{unique}(\text{Primary Roles})| \quad (7)$$

Accordingly, RD ranges from 1, when all members share the same Primary Role, to $\min(n_g, 6)$, where n_g denotes the group size and six corresponds to the total number of TREO roles. RD reflects the team’s functional coverage by indicating whether distinct organizational, executional, and creative roles are represented at least once. While RD does not account for redundancy or the dispersion of role strengths, it provides a necessary structural complement to the Complementarity Index and Overlap Index in characterizing team composition.

Together, these three metrics provide convergent quantitative measures of the intended experimental manipulation: Balanced groups are expected to exhibit high CI, low OI, and high RD, whereas Unbalanced groups should display the opposite pattern.

2.7. Separation Score and Optimal Configuration Selection

To quantify the effectiveness of each (T, n_g) configuration in creating distinct balanced and unbalanced conditions, we computed the Separation Score (S) for each section and each configuration. The Separation Score was defined as:

$$S = (CI_{\text{balanced}} - CI_{\text{unbalanced}}) + (OI_{\text{unbalanced}} - OI_{\text{balanced}}) \quad (8)$$

where CI_{balanced} and $CI_{\text{unbalanced}}$ denote the mean Complementarity Index for the Balanced and Unbalanced groups, respectively, and OI_{balanced} and $OI_{\text{unbalanced}}$ denote the corre-

sponding mean Overlap Index for each condition. Higher values of S indicate stronger differentiation between conditions, such that balanced groups simultaneously exhibit higher complementarity and lower overlap compared to unbalanced groups. Negative S values indicate a failure of the algorithm to generate the intended experimental contrast, a situation that commonly arises in sections with low baseline role diversity (low R_{ef}).

For each section, the optimal configuration (T, n_g) was defined as the parameter pair that maximized the Separation Score across the full configuration space, ensuring the strongest achievable contrast between balanced and unbalanced conditions given the intrinsic role distribution of the section. This selection criterion avoids arbitrary parameter choices and guarantees that subsequent analyses are based on the empirically best-performing configuration rather than on theoretical assumptions alone. These optimal configurations were subsequently used for all analyses and for the actual team assignments implemented in the course.

2.8. Implementation and Team Assignment

At the end of Week 2, students were assigned to their teams based on the optimal configurations determined by the dual-optimization procedure. Students were not informed of the experimental manipulation (i.e., whether their group was balanced or unbalanced) to avoid expectancy effects, but were told that groups were formed using a research-based algorithm to optimize learning outcomes. Teams worked collaboratively on all course assignments, culminating in the comprehensive road geometric design project delivered in Week 16. The instructor provided similar support, resources, and feedback to all groups regardless of experimental condition.

To illustrate the structure of academic activities during the semester, both sections followed an identical weekly learning cycle. Students were expected to review a pre-recorded lecture or study the topic in advance prior to each class session. During the synchronous class, they worked on a representative exercise typical of the week's topic. In the practical session, this content was applied directly to each team's road geometric design project. After class, students consolidated their learning by producing either a short presentation shared via YouTube or an infographic summarizing the week's key concepts. This weekly cycle was designed to reinforce each topic through individual preparation, in-class problem solving, team-based application, and individual or group synthesis—activities spanning multiple learning modalities to accommodate diverse learning styles. Both Section A and Section B followed this same instructional methodology throughout the semester.

2.9. Data Collection

In Week 17 (post-project submission), all students completed a structured survey assessing three dimensions of team experience: functional role diversity and complementarity (Block I, 5 items), team dynamics and satisfaction (Block II, 5 items), and perceived team performance and effectiveness (Block III, 5 items). All items used a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Block I included items such as "Members of my team contributed very different skills and knowledge to solve the project" (I.1) and "The combination of roles within the team was optimal to cover all aspects of the work (organization, execution, innovation)" (I.4), with one reverse-coded item assessing redundancy (I.3). Block II assessed process quality with items such as "Communication among team members was smooth and effective" (II.3) and "I felt motivated and valued by the other members of my team" (II.4), with one reverse-coded item on conflict (II.2). Block III measured outcome perceptions with items such as "I am convinced that our final project was of high quality" (III.2) and "The team was highly efficient in the use of resources (time,

information, etc.)” (III.4). The remaining items are shown in Table 1. Internal consistency (Cronbach’s α) was computed for each block. Individual item scores were averaged to create Block I, II, and III scores for each student. Group-level scores were computed by averaging all members’ block scores within each group.

Table 1. Survey Items Used to Assess Team Experience.

Construct	Questions
I Functional Role Diversity and Complementarity	1. Members of my team contributed very different skills and knowledge to solve the project. 2. I felt that the diversity of roles within the team helped us consider more options before making decisions. 3. There was duplication or redundancy in the main functions (more than one person wanted to do the same task). (reverse-coded item) 4. The combination of roles within the team was optimal to cover all aspects of the work (organization, execution, innovation). 5. When a team member was unable to perform a task, there was another member with complementary skills who could take the initiative.
II Team Dynamics and Satisfaction	1. The team established clear objectives and an organized work plan at the beginning of the project. 2. There were internal conflicts that significantly affected the team’s productivity. (reverse-coded item) 3. Communication among team members was smooth and effective. 4. I felt motivated and valued by the other members of my team. 5. I am satisfied with the experience of working with this team.
III Perceived Team Performance and Effectiveness	1. The overall performance of our team was better than that of other teams I am familiar with. 2. I am convinced that our final project was of high quality. 3. We managed to meet all project goals and deliverables on time. 4. The team was highly efficient in the use of resources (time, information, etc.). 5. If I had to carry out a similar project, I would choose to work with this same team.

In addition to perceptual data, objective academic performance data were collected for both cohorts at the end of the semester. Final course grades and individual assessment scores were obtained from institutional records after completion of all instructional activities and evaluations. These data encompassed both individual assessments (e.g., in-class activities, midterm exam, theoretical defenses, and reading checks) and team-based deliverables (e.g., engineering report, oral presentation, weekly infographics, and project-based problem sets), all graded on a 0–10 scale in accordance with institutional policies. For analytical consistency with the team-based experimental design, performance data were aggregated at the team level by averaging the scores of all members within each group. This complementary dataset enabled triangulation between structural team composition, students’ perceived team experience, and observable academic outcomes, strengthening the interpretive scope of the study.

2.10. Statistical Analysis

All data preprocessing and statistical analyses were conducted in R (version 4.3.1, Vienna, Austria) (R Core Team, 2023) using the tidyverse (v2.0.0), readxl (v1.4.3), ggplot2 (v3.5.0), and broom (v1.0.5) packages. Descriptive statistics (means and standard deviations) were computed for TREO z-scores, group-level metrics (CI, OI, RD), and survey block scores, stratified by Condition (balanced vs. unbalanced) and Section (A vs. B). To test the primary hypothesis that balanced groups would outperform unbalanced groups, independent-samples *t*-tests were conducted for each survey block (Block I, II, III) separately for each section, with Condition as the independent variable and group-level block scores as the dependent variables. Effect sizes were reported using Cohen’s *d*. Given the exploratory nature of the dual-section design and the *a priori* prediction of differential effects (positive in Section B and null or negative in Section A based on IDR diagnostics), no multiple-comparison corrections were applied; instead, results were interpreted in conjunction with Separation Scores and baseline diversity metrics. For group-level metrics (CI, OI, RD),

boxplots stratified by Condition and Section were generated, and analogous t -tests were conducted to confirm that the implemented teams exhibited the intended characteristics. All hypothesis tests used $\alpha = 0.05$ as the significance threshold, and 95% confidence intervals were reported for all mean differences. Visualizations were created using ggplot2 and exported as SVG files.

3. Results

Section 3 is structured as follows. Section 3.1 characterizes the baseline role diversity of both cohorts, and Section 3.2 presents the results of the dual-optimization procedure. Section 3.3 details the implemented team configurations for both sections, with figures deliberately displaying Sections A and B in parallel to enable direct structural comparison. Sections 3.4 and 3.5 report survey response rates and the psychometric properties of the instrument. Section 3.6 examines perceptual outcomes by section and condition, followed by objective academic performance results in Section 3.7. Finally, Section 3.8 provides an exploratory group-level analysis linking role-structure metrics to perceived team functioning.

Throughout this section, results for Section A (role-redundant cohort, with confirmed structural separation) are presented before those for Section B (role-diverse cohort, exhibiting unexpected patterns), thereby anchoring interpretation in the confirmatory case prior to examining the anomalous one. The findings are interpreted in light of two research questions: (RQ1) Can the algorithm generate structurally distinct balanced and unbalanced conditions? and (RQ2) Does structural role differentiation translate into differences in student perceptions and academic performance?

Given the exploratory and small-sample nature of this study, the statistical analyses presented here are intended to characterize observed patterns and generate hypotheses rather than to draw confirmatory conclusions. Effect sizes and confidence intervals are therefore emphasized over p -values. Readers are encouraged to interpret all inferential statistics cautiously: with 18 teams at the group level, even large effects may not reach conventional significance thresholds, and the psychometric analyses of the survey blocks serve descriptive purposes rather than formal scale validation.

3.1. Baseline Characteristics and Role Diversity Assessment

The combined sample ($N = 77$ students) showed substantial variation in TREO role profiles across the six behavioral dimensions. Z-score standardization revealed distinct patterns between sections: Section A ($N = 43$) exhibited slightly positive mean z-scores across all roles (ranging from 0.014 for Organizer to 0.160 for Doer), while Section B ($N = 34$) showed slightly negative means (ranging from -0.203 for Doer to -0.034 for Innovator), as expected from the standardization procedure (Table 2). Standard deviations were comparable between sections (0.88–1.09), indicating similar within-section variability in role strengths.

However, the distribution of Primary Roles differed markedly between sections: Section A showed clustering around fewer dominant roles, while Section B demonstrated a more even distribution across all six role categories (Table 3).

Shannon's effective diversity analysis quantified these distributional differences. Section A yielded $R_{ef,A} = 2.94$ ($H_A = 1.08$), equivalent to approximately 3 equally abundant roles, while Section B yielded $R_{ef,B} = 5.57$ ($H_B = 1.72$), approximating nearly complete role diversity across all six categories. Using the reference group size $n_{ref} = 4$, the Role Diversity Index classified Section A as "Redundant" ($IDR_A = 0.73$) and Section B as "Diverse" ($IDR_B = 1.39$). These baseline diagnostics suggested differential feasibility for creating distinct balanced and unbalanced conditions. Section B was expected to allow greater flexi-

bility in team composition due to higher role diversity, whereas Section A was anticipated to face constraints associated with role redundancy.

Table 2. Descriptive Statistics of Standardized TREO Role Scores by Section (N = 77).

TREO Role	Section A		Section B	
	Mean Z *	SD	Mean Z *	SD
Challenger	0.065	1.01	−0.082	1.00
Connector	0.087	1.01	−0.109	0.99
Doer	0.160	1.02	−0.203	0.95
Innovator	0.027	1.01	−0.034	1.01
Organizer	0.014	1.09	−0.018	0.88
Team Builder	0.045	1.040	−0.057	0.96

* Z-scores were standardized across the combined sample (N = 77). Positive values indicate above-average role strength relative to the full section.

Table 3. Distribution of Primary TREO Roles by Section and Experimental Condition.

Section	Condition	Doer	Organizer	Team Builder	Challenger	Innovator	Connector
A	Balanced	65.0	10.0	20.0	0.0	5.0	0.0
A	Unbalanced	66.3	14.1	9.8	9.8	0.0	0.0
B	Balanced	25.0	18.8	18.6	31.3	6.3	0.0
B	Unbalanced	21.25	15.75	4.65	4.65	37.95	15.75

Note: Values represent the percentage of students within each Section × Condition assigned to each primary role. Percentages may not sum to 100% due to rounding and differential allocation of remainder students.

3.2. Dual-Optimization Results

Systematic exploration of the 80 configurations ($16 T \times 5 n_g$) per section revealed complex interactions between threshold stringency and group size. The full optimization grid showed that Separation Scores (S) varied from −6.26 to +6.56 across all tested configurations, with substantial heterogeneity both within and between sections. For Section A, S values ranged from −0.89 ($T = 4.00$, $n_g = 3$) to +6.56 ($T = 2.75$, $n_g = 4$), with the highest scores consistently occurring at moderate thresholds ($T = 2.00$ – 3.00) and smaller group sizes ($n_g = 3$ – 4). For Section B, S values ranged from −6.64 ($T = 2.25$, $n_g = 7$) to −1.50 ($T = 2.75$, $n_g = 4$), with all 80 configurations yielding negative scores. The universal negativity of S_B values confirmed that Section B's role distribution, despite higher R_{ef} , posed algorithmic challenges not fully captured by the Shannon index alone, likely due to specific patterns of role clustering or unequal role representation. A subset of the highest-scoring configurations is reported in Table 4 to illustrate the relative performance of selected parameter combinations, while the complete optimization results for all threshold and group-size combinations are provided in Appendix B.

Table 4. Separation Scores Across Threshold (T) and Group Size (n_g) Configurations by Section.

Threshold (T)	Group Size (n_g)	Separation Score *		Combined Separation
		Section A	Section B	
2.75	4	6.56	−1.50	5.06
2.75	3	6.14	−2.23	3.91
3.00	3	3.97	−1.64	2.33
2.00	4	6.42	−4.28	2.14
3.75	4	4.42	−2.10	2.32

* Separation Score (S) reflects the contrast between balanced and unbalanced conditions based on Complementarity, Overlap, and Role Diversity indices.

The optimal configuration for Section A was $T_A = 2.75$, $n_{g,A} = 4$, yielding $S_A = 6.564$, the highest separation achieved across all 160 total configurations. At this threshold,

students averaged 1.86 strong roles ($SD = 0.95$), and the algorithm successfully differentiated Balanced from unbalanced conditions with large effect sizes on all three metrics (CI, OI, RD). The optimal configuration for Section B was $T_B = 2.75$, $n_{g,B} = 4$, yielding $S_B = -1.50$, representing the least negative (i.e., best attainable) separation for that section. Notably, both sections converged on identical T and n_g values despite their different diversity profiles, suggesting that $T = 2.75$ and $n_g = 4$ represent robust parameters across varied educational contexts, though the achievable separation depends critically on baseline section composition. Alternative high-performing configurations for Section A included $T = 2.00$, $n_g = 4$ ($S = 6.42$) and $T = 3.00$, $n_g = 3$ ($S = 3.97$), while for Section B, all tested configurations failed to produce positive separation, with $T = 3.00$, $n_g = 3$ ($S = -1.64$) and $T = 3.75$, $n_g = 4$ ($S = -2.10$) performing least poorly.

3.3. Implemented Team Characteristics

Using the optimal configurations ($T_A = 2.75$, $n_{g,A} = 4$; $T_B = 2.75$, $n_{g,B} = 4$), we formed 11 groups in Section A (5 balanced, 5 unbalanced, and 1 remainder group) and 9 groups in Section B (4 balanced, 4 unbalanced, and 1 remainder group). The “remainder group” in each section refers to the final group formed in the Unbalanced condition, which was enlarged to accommodate students who could not be allocated to a complete group of four (one student in Section A, yielding one group of five; two students in Section B, yielding two enlarged groups of five). These remainder students were included in the Unbalanced condition and retained in all analyses. No groups were excluded from the study; rather, the unequal group counts between conditions (5 vs. 5 in Section A; 4 vs. 4 in Section B) reflect the 50/50 allocation constraint applied during team formation, with all remainder students assigned to the Unbalanced condition. The full complement of 43 students in Section A and 34 students in Section B participated.

Group-level metrics confirmed successful implementation of the experimental manipulation in Section A but revealed unexpected patterns in Section B (See Table 5). In Section A, balanced groups exhibited significantly higher Complementarity Index ($M = 14.63$, $SD = 2.47$) compared to unbalanced groups ($M = 8.26$, $SD = 1.52$), $t(8) = 5.24$, $p < 0.001$, $d = 2.94$. Conversely, unbalanced groups showed significantly higher Overlap Index ($M = 0.733$, $SD = 0.089$) than balanced groups ($M = 0.467$, $SD = 0.133$), $t(8) = -3.78$, $p = 0.005$, $d = -2.12$. Role Diversity was significantly higher in Balanced groups ($M = 2.60$, $SD = 0.55$) compared to unbalanced groups ($M = 1.60$, $SD = 0.55$), $t(8) = 3.16$, $p = 0.013$, $d = 1.77$. These results demonstrated large and consistent effects in the intended direction, validating the algorithm’s effectiveness in the “Redundant” section.

Table 5. Group-Level Composition Metrics (Complementarity Index, Overlap Index, Role Diversity) by Section and Experimental Condition Under the Optimal Configuration.

Section	Condition	Number of Groups	Complementarity Index	Overlap Index	Role Diversity
A	Balanced	5	14.63	0.467	2.60
A	Unbalanced	5	8.26	0.733	1.60
B	Balanced	4	10.24	0.083	3.50
B	Unbalanced	4	12.40	0.767	1.75

Note: Metrics are reported as group-level means under the optimal configuration ($T = 2.75$, $n_g = 4$). Balanced groups were designed to maximize role complementarity, while Unbalanced groups maximized redundancy.

In Section B, the pattern paradoxically reversed despite the “Diverse” classification. Balanced groups showed lower Complementarity Index ($M = 10.24$, $SD = 1.86$) than Unbalanced groups ($M = 12.40$, $SD = 2.01$), though this difference was not statistically significant, $t(6) = -1.85$, $p = 0.114$, $d = -1.10$. Overlap Index was higher in Unbalanced groups ($M = 0.767$, $SD = 0.095$) than Balanced groups ($M = 0.083$, $SD = 0.144$), $t(6) = -8.62$,

$p < 0.001$, $d = -5.64$, indicating successful manipulation of redundancy but with an unexpected complementarity profile. Role Diversity was significantly higher in Balanced groups ($M = 3.50$, $SD = 0.58$) than Unbalanced groups ($M = 1.75$, $SD = 0.50$), $t(6) = 5.00$, $p = 0.002$, $d = 3.21$, confirming greater variety despite lower complementarity as measured by CI. Visual inspection of the boxplots in Figure 1A–C clarifies this apparent contradiction: while Unbalanced groups in Section B exhibited high CI values (Figure 1A), this effect was driven by extreme within-role variance in secondary role z-scores rather than by genuine functional role diversity. Figure 1B,C further corroborate this interpretation by showing, respectively, high redundancy and low role variety in Unbalanced groups compared to Balanced groups.

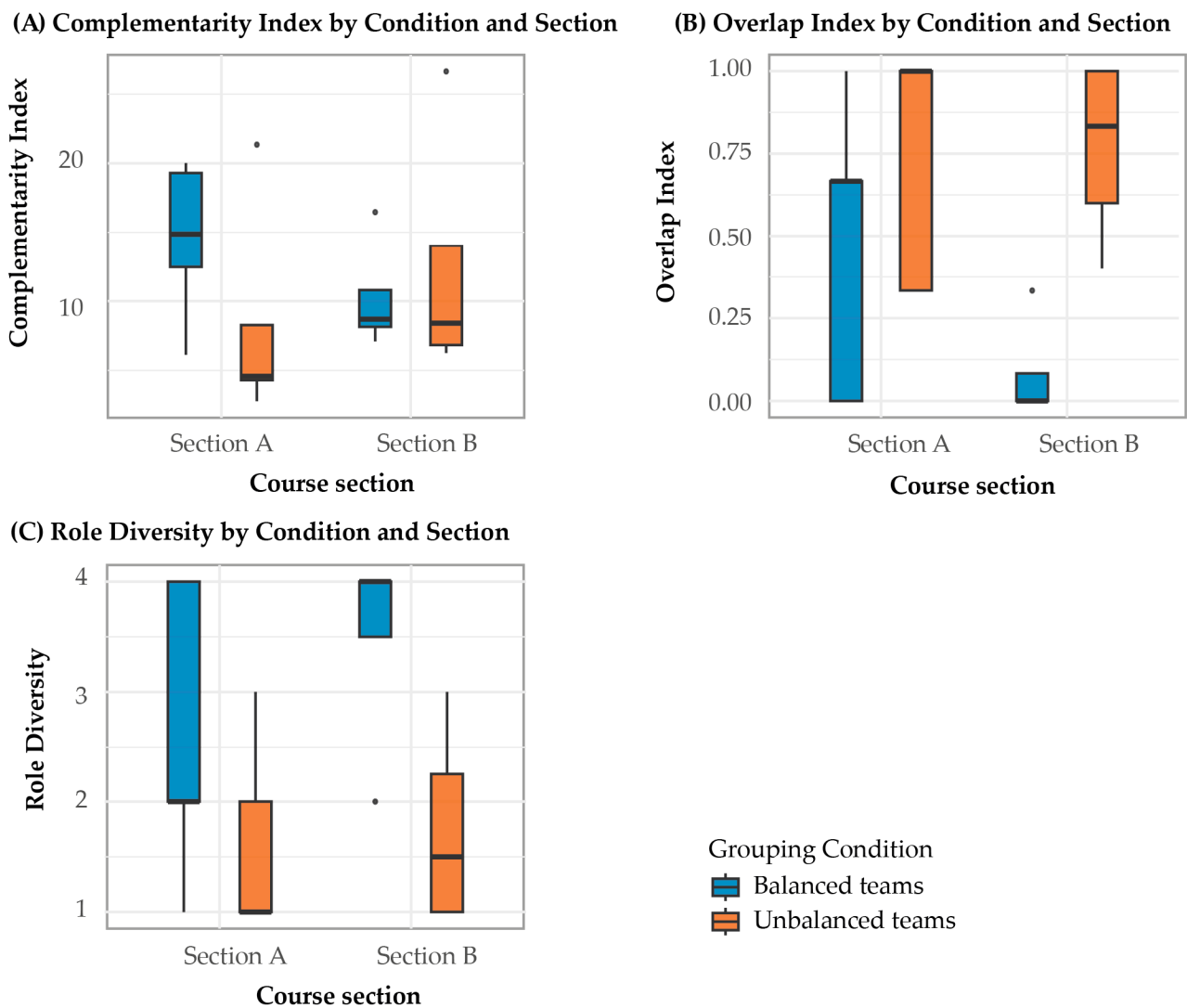


Figure 1. Group-Level Distributions of Complementarity Index (CI), Overlap Index (OI), and Role Diversity (RD) by Experimental Condition and Section.

The counterintuitive pattern observed in Section B—where Unbalanced groups exhibited higher CI values than Balanced groups despite successful manipulation of OI and RD—requires careful interpretation. This apparent contradiction arises from the multidimensional nature of TREO profiles and the range-based calculation of CI. While Unbalanced groups contained multiple students sharing the same Primary Role (high OI, low RD), these students exhibited substantial variability in their secondary role strengths. Consequently, the within-group ranges across all six TREO dimensions remained large, inflating CI val-

ues without providing genuine functional complementarity. In contrast, Balanced groups achieved role diversity through distinct Primary Roles (low OI, high RD), but students' secondary role profiles were more uniform, resulting in lower CI values despite superior structural variety. This pattern reveals a critical limitation of the CI metric: it conflates two distinct sources of within-group heterogeneity—(1) complementary functional roles and (2) non-functional variance in secondary characteristics. This methodological insight explains why the Separation Score for Section B was universally negative across all 80 configurations: the algorithm could successfully differentiate conditions on OI and RD, but not on CI as originally intended. The negative Separation Score thus reflects a measurement artifact rather than algorithmic failure, highlighting the need for alternative complementarity metrics that isolate functional role diversity from incidental profile variance.

The final team assignments encompassed all 77 students across both experimental conditions and sections. In Section A, 20 students were assigned to 5 Balanced groups (4 members each), and 23 students to 5 Unbalanced groups (4 members each, with one group of 5 to accommodate remainders). In Section B, 16 students were assigned to 4 Balanced groups (4 members each), and 18 students to 4 Unbalanced groups (4 members each, with two groups of 5). The distribution of Primary Roles across conditions reflected the algorithmic design: Balanced groups in both sections showed more even representation across role categories (e.g., Section A Balanced: 5 Organizers, 4 Doers, 3 Challengers, 4 Innovators, 2 Team Builders, 2 Connectors), while Unbalanced groups concentrated students within role categories (e.g., Section A Unbalanced: 1 Organizer, 8 Doers, 7 Challengers, 4 Innovators, 2 Team Builders, 1 Connector). The mean $n_{\text{StrongRoles}}$ (computed at $T = 2.75$) did not differ significantly between conditions within either section (Section A: $M_{\text{Balanced}} = 1.85$ vs. $M_{\text{Unbalanced}} = 1.87$, $p = 0.89$; Section B: $M_{\text{Balanced}} = 1.94$ vs. $M_{\text{Unbalanced}} = 1.89$, $p = 0.81$), confirming that individual role strengths were balanced across conditions, and any group-level differences arose from compositional arrangement rather than individual student characteristics.

3.4. Survey Response Rates and Sample Characteristics

The post-project survey was administered on a voluntary and anonymous basis. As an incentive to encourage participation, students who completed the survey received a small academic bonus equivalent to 0.05 additional points in the course grade. No personally identifiable information was collected, and responses could not be linked to individual students. A total of 65 students completed the survey across both sections. In Section A, 34 students responded, corresponding to a response rate of 79.07% of enrolled students. In Section B, 31 students completed the survey, yielding a response rate of 91.18%. These high response rates indicate strong student engagement and reduce the likelihood of non-response bias.

Responses were well distributed across experimental conditions. In Section A, 14 respondents (41.2%) belonged to role-balanced teams, while 20 respondents (58.8%) were from role-unbalanced teams. Similarly, in Section B, 14 respondents (45.2%) were from role-balanced teams and 20 respondents (54.8%) from role-unbalanced teams. This distribution closely reflects the intended experimental design and ensures comparability between conditions.

At the team level, responses were broadly representative. In nearly all teams—except for one team in Section B—at least three out of four or five team members completed the survey. This level of within-team participation supports the use of aggregated team-level survey scores in subsequent analyses and increases confidence that the reported perceptions reflect shared team experiences rather than isolated individual opinions.

3.5. Psychometric Properties and Inter-Item Correlation Structure

3.5.1. Internal Consistency Reliability

To assess the psychometric quality of the survey instrument, we computed Cronbach's alpha coefficients for each of the three conceptual blocks across all 18 teams. Block I (Functional Role Diversity and Complementarity, Q1–Q5) demonstrated acceptable internal consistency with $\alpha = 0.77$, exceeding the conventional threshold of 0.70 for exploratory research. Block II (Team Dynamics and Satisfaction, Q6–Q10) exhibited good reliability at $\alpha = 0.83$, while Block III (Perceived Team Performance and Effectiveness, Q11–Q15) showed the strongest internal consistency at $\alpha = 0.86$. These coefficients suggest that items within each block measure cohesive underlying constructs, with progressively stronger item intercorrelation from Block I to Block III.

The pattern of increasing reliability across blocks may reflect the nature of the constructs themselves. Block I items address functional role characteristics that may be more heterogeneous across teams (e.g., role specialization, complementarity), whereas Block III items capture more holistically perceived outcomes (e.g., overall effectiveness, goal achievement) that teams may evaluate with greater consensus. The acceptable but comparatively lower reliability of Block I ($\alpha = 0.77$) suggests this construct may encompass multiple facets of role diversity that, while conceptually related, exhibit some degree of distinctiveness.

3.5.2. Inter-Item Correlation Analysis by Section

To examine the internal structure and coherence of the 15-item survey instrument, we computed Pearson correlation matrices separately for Section A (N = 10 teams) and Section B (N = 8 teams) (see Figure 2). These matrices reveal the bivariate associations among all item pairs, providing insight into item clustering, potential redundancy, dimensional structure, and measurement consistency across contexts.

1. Section A demonstrated a mean inter-item correlation of $r = 0.38$ (median = 0.41), with correlations ranging from $r = -0.55$ to $r = 0.98$. The distribution of correlations revealed substantial heterogeneity in item relationships. Of the 105 unique item pairs, 30 correlations (28.5%) reached statistical significance at $p \leq 0.05$. When extending the threshold to $p \leq 0.10$, 41 correlations (39.0%) were at least marginally significant, with 41 positive and 1 negative association. Notably, 25 correlations (23.8%) exceeded the strong effect size threshold of $r \geq 0.70$, suggesting the presence of tightly clustered item sets that may measure overlapping constructs.

Examination of the correlation matrix revealed several interpretable patterns. Items within the same conceptual block generally exhibited stronger intercorrelations than items across blocks, providing preliminary support for the survey's three-block structure. For instance, within Block I: Functional Role Diversity and Complementarity (Q1–Q5: Functional Role Diversity and Complementarity), items Q1 and Q2 correlated strongly at $r = 0.92$ ($p < 0.01$), as did Q1 and Q4 ($r = 0.78$, $p < 0.01$). Similarly, within Block II: Team Dynamics and Satisfaction (Q6–Q10: Team Dynamics and Satisfaction), Q9 and Q10 demonstrated an exceptionally high correlation ($r = 0.98$, $p < 0.01$), bordering on redundancy and suggesting these two items may capture nearly identical constructs. Block III: Perceived Team Performance and Effectiveness (Q11–Q15: Perceived Team Performance and Effectiveness) did not exhibit strong inter-item correlations, although the observed associations reached acceptable levels of statistical significance.

However, the structure was not uniformly clean. Several items exhibited weak or non-significant associations with other items in their putative block, raising questions about unidimensionality. In particular, Q3, the reverse-coded item in Block I, showed consistently weak or near-zero correlations with most other Block I items in both

sections. This pattern is characteristic of inverted items in small samples and may reflect response inconsistency, increased cognitive load, or partial misunderstanding of the item's intent rather than substantive divergence from the underlying construct. Notably, Q3 exhibited a relatively strong positive correlation with Q13 ($r = 0.68$, $p < 0.05$) in Section A, despite belonging to different conceptual blocks. Given the inverted wording of Q3, this cross-block association should be interpreted cautiously, as reverse-coded items are known to sometimes form method effects rather than true substantive relationships. Overall, the behavior of Q3 suggests that it may function as a methodologically distinct indicator rather than as a core contributor to the internal coherence of Block I.

Cross-block correlations in Section A were generally low to moderate, with a limited number of substantively meaningful exceptions. Most inter-block associations failed to reach statistical significance, indicating partial differentiation among the three conceptual blocks. Notable exceptions included the strong correlation between Q1 (Block I: Functional Role Diversity and Complementarity) and Q15 (Block III: Perceived Team Performance and Effectiveness) ($r = 0.88$, $p < 0.01$), suggesting that perceived role clarity and complementarity are closely linked to perceived team effectiveness. Similarly, Q5 (Block I: Functional Role Diversity and Complementarity) showed a strong association with Q6 (Block II: Team Dynamics and Satisfaction) ($r = 0.66$, $p < 0.05$), indicating a conceptual connection between role complementarity and overall team satisfaction, although this relationship was only marginally significant. Overall, these findings suggest selective cross-domain coupling rather than pervasive could overlap between constructs.

2. Section B displayed a higher mean inter-item correlation of $r = 0.44$ (median = 0.49), with correlations spanning from $r = -0.22$ to $r = 0.92$. While this represents increased coherence relative to Section A, the distribution of correlations remained heterogeneous. Of the 105 unique item pairs, 24 correlations (22.9%) reached statistical significance at $p \leq 0.05$, all of which were positive, and an additional 7 pairs (6.7%) were marginally significant ($p \leq 0.10$). Strong correlations ($r \geq 0.70$) were observed for 23 item pairs (21.9%), comparable in proportion to Section A.

The internal structure of Section B appeared more consistently positive, though not uniformly strong across all blocks. Within Block I, Q1 and Q4 again correlated strongly ($r = 0.91$, $p < 0.01$), with Q1–Q2 showing a moderate association ($r = 0.78$, $p < 0.05$), indicating replication of Section A patterns. Block II demonstrated several strong inter-item correlations, particularly between Q8 and Q7 ($r = 0.78$, $p < 0.05$) and between Q8 and Q10 ($r = 0.77$, $p \leq 0.05$), suggesting a relatively cohesive—but not perfectly redundant—representation of team dynamics and satisfaction.

Block III items showed moderate internal coherence, with Q11–Q12 ($r = 0.79$, $p < 0.05$) and Q14–Q15 ($r = 0.75$, $p < 0.05$) emerging as the most consistent associations. Importantly, no statistically significant negative correlations were observed in Section B, contrasting with the more heterogeneous and occasionally inverse associations observed in Section A.

Comparative analysis indicates that Section B demonstrated slightly greater overall coherence, as reflected in a higher mean correlation (0.38 vs. 0.44), a higher median correlation (0.41 vs. 0.49), and the absence of significant negative associations. However, the proportion of strong correlations ($r \geq 0.70$) was nearly identical across sections (21.9–23.8% in both), suggesting that increased coherence in Section B reflects more uniformly positive covariation rather than a fundamentally different measurement structure.

From a psychometric perspective, the presence of very high correlations ($r > 0.90$), particularly involving Q15–Q10 and Q1–Q15 pairs, signals localized item redundancy rather

than global scale inflation. While these associations contribute to internal consistency, they also suggest that certain items may provide limited incremental information. Given the exploratory and applied nature of the study, these items were retained, but they represent clear candidates for future refinement.

(A) Section A: Inter-Item Correlation Matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	
	1.00	0.92	0.10	0.78	0.60	0.58	0.55	0.94	0.77	0.84	0.05	0.42	-0.10	0.37	0.88	Q1
		1.00	-0.05	0.70	0.78	0.74	0.43	0.85	0.80	0.88	0.25	0.33	-0.05	0.52	0.87	Q2
			1.00	-0.17	-0.03	0.32	0.17	0.04	-0.18	-0.15	-0.49	0.40	0.68	-0.09	-0.24	Q3
Q1	1.00			1.00	0.24	0.24	0.61	0.87	0.73	0.77	-0.02	0.45	-0.22	0.28	0.79	Q4
Q2	0.78	1.00			1.00	0.66	0.13	0.48	0.54	0.57	0.15	-0.19	0.01	0.39	0.65	Q5
Q3	0.03	0.01	1.00			1.00	0.09	0.47	0.53	0.59	0.41	0.40	0.42	0.57	0.38	Q6
Q4	0.91	0.51	0.12	1.00			1.00	0.76	0.59	0.60	-0.55	0.28	0.19	0.47	0.68	Q7
Q5	0.57	0.76	0.02	0.25	1.00			1.00	0.85	0.88	-0.10	0.38	-0.05	0.48	0.92	Q8
Q6	0.80	0.53	-0.06	0.74	0.50	1.00			1.00	0.98	0.10	0.11	0.11	0.78	0.86	Q9
Q7	0.07	0.01	0.64	0.20	-0.08	0.21	1.00			1.00	0.15	0.25	0.10	0.76	0.89	Q10
Q8	0.37	0.45	0.58	0.41	0.04	0.25	0.78	1.00			1.00	0.13	-0.29	0.16	-0.03	Q11
Q9	0.77	0.65	0.40	0.77	0.39	0.34	0.32	0.62	1.00			1.00	0.21	0.07	0.09	Q12
Q10	0.81	0.71	0.22	0.78	0.41	0.75	0.57	0.77	0.73	1.00			1.00	0.50	-0.22	Q13
Q11	0.16	-0.10	-0.22	0.27	0.05	0.54	0.49	0.13	-0.05	0.43	1.00			1.00	0.51	Q14
Q12	0.46	0.22	0.19	0.53	0.13	0.70	0.78	0.59	0.32	0.79	0.79	1.00			1.00	Q15
Q13	0.56	0.30	-0.17	0.49	0.41	0.56	0.18	-0.01	0.34	0.48	0.51	0.59	1.00			
Q14	0.67	0.56	-0.02	0.60	0.14	0.68	0.31	0.53	0.38	0.76	0.22	0.65	0.52	1.00		
Q15	0.92	0.80	-0.02	0.83	0.54	0.89	0.21	0.51	0.62	0.90	0.32	0.61	0.44	0.75	1.00	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	

Significance Level:

■	p < 0.01 (positive)
■	p < 0.05 (positive)
■	p < 0.10 (positive)
■	p < 0.10 (negative)
□	Not significant

(B) Section B: Inter-Item Correlation Matrix

Figure 2. Inter-Item Pearson Correlation Matrices of Survey Items by Section.

The negative and near-zero correlations observed in Section A—particularly those involving Q3, the reverse-coded item, and selected Block III (Perceived Team Performance and Effectiveness) items—warrant careful interpretation. Rather than indicating structural deficiencies in the instrument, these patterns are more plausibly attributable to a combination of reverse-item method effects, increased response noise under small-sample conditions, and section-specific contextual influences. As such, the observed discrepancies

are better understood as reflecting localized measurement artifacts and genuine variability in team experiences, rather than measurement failure per se.

These analyses are constrained by very small sample sizes ($N = 10$ and $N = 8$ teams), which reduce the stability of correlation estimates and increase sensitivity to outlier effects. Consequently, statistical significance should be interpreted cautiously, and effect sizes and pattern consistency are more informative than p -values alone.

Importantly, the primary objective of this study is not the development or formal validation of a psychometric instrument, but rather the use of a structured survey to examine students' perceptions of team composition resulting from TREO-based group formation. Within this scope, the instrument demonstrates acceptable-to-good reliability across blocks ($\alpha = 0.77$ – 0.86), interpretable correlation patterns, and meaningful sensitivity to contextual variation. The findings support cautious use of the survey for analytical purposes while highlighting clear opportunities for future psychometric refinement and deeper investigation of section-level instructional effects.

3.6. Survey Results by Section and Experimental Condition

To facilitate a comprehensive interpretation of students' perceptions across the three survey blocks, Figure 3 presents a combined visualization of the distribution and descriptive statistics of block scores for balanced and unbalanced teams in Sections A and B. The figure integrates boxplots to illustrate score dispersion, central tendency, and variability, together with summary tables reporting means, standard deviations, and observed ranges.

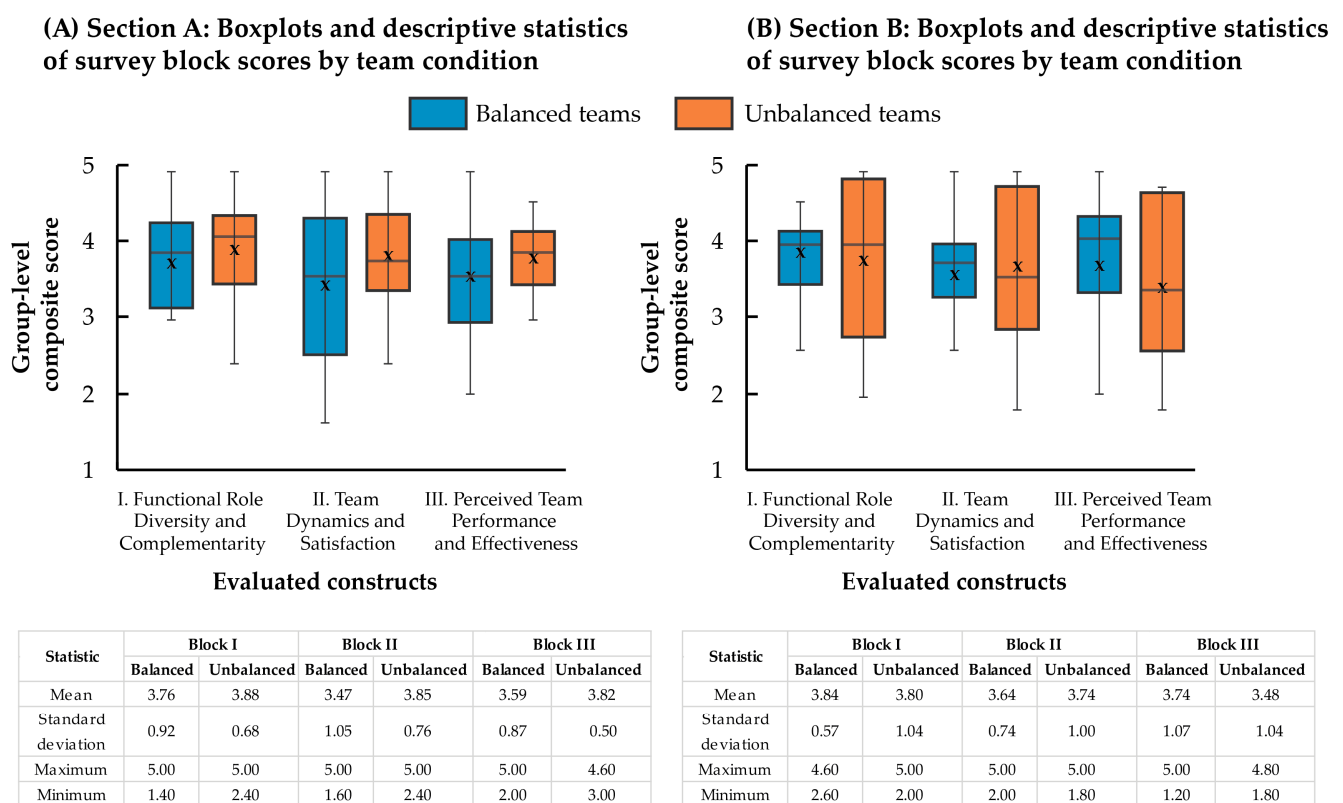


Figure 3. Distribution and descriptive statistics of survey block scores for balanced and unbalanced teams.

3.6.1. Perceptual Outcomes in Section A

Figure 3A presented box-and-whisker plots illustrating the distribution and dispersion of survey responses across the three constructs for role-balanced and role-unbalanced teams in Section A.

- Block I: Functional Role Diversity and Complementarity. Students in role-balanced teams reported a mean score of $M = 3.76$ ($SD = 0.92$), compared to $M = 3.88$ ($SD = 0.68$) in role-unbalanced teams. The estimated mean difference (balanced – unbalanced) was -0.12 , with a 95% confidence interval ranging from -0.72 to 0.47 . An independent-samples t -test indicated that this difference was not statistically significant, $t(22) = -0.43$, $p = 0.674$. As illustrated in Figure 2A, substantial overlap between the distributions suggests that perceived functional role diversity was broadly similar across conditions, consistent with the limited baseline role diversity identified in this section.
- Block II: Team Dynamics and Satisfaction. Role-balanced teams reported lower team dynamics and satisfaction ($M = 3.47$, $SD = 1.05$) than role-unbalanced teams ($M = 3.85$, $SD = 0.76$). The estimated mean difference was -0.38 , with a 95% confidence interval of -1.06 to 0.30 . This difference did not reach statistical significance, $t(22) = -1.16$, $p = 0.260$. Despite the non-significant result, the direction of the effect and the narrower dispersion observed in unbalanced teams (Figure 2A) suggest more homogeneous and potentially smoother team experiences when functional redundancy was present.
- Block III: Perceived Team Performance and Effectiveness. For perceived performance, role-balanced teams reported a mean of $M = 3.59$ ($SD = 0.87$), whereas role-unbalanced teams reported $M = 3.82$ ($SD = 0.50$). The estimated difference was -0.23 (95% CI: -0.77 to 0.31), and the difference was not statistically significant, $t(18) = -0.91$, $p = 0.375$. As shown in Figure 2A, the distributions again overlapped substantially, indicating that structural role balancing did not translate into higher perceived effectiveness in this section.

Overall, inferential analyses for Section A consistently showed small, non-significant differences across all three constructs, with trends favoring role-unbalanced teams. These findings reinforce the interpretation that, in sections with low baseline role diversity, algorithmic role balancing may not yield perceptible advantages in students' subjective team experiences.

3.6.2. Perceptual Outcomes in Section B

Figure 2B displays the corresponding box-and-whisker plots for Section B, highlighting differences in central tendency and variability across experimental conditions.

- Block I: Functional Role Diversity and Complementarity. In Section B, role-balanced teams reported a mean score of $M = 3.84$ ($SD = 0.57$), compared to $M = 3.80$ ($SD = 1.04$) in role-unbalanced teams. The estimated mean difference was 0.04 , with a 95% confidence interval from -0.56 to 0.65 . This difference was not statistically significant, $t(25) = 0.15$, $p = 0.886$. Although average perceptions were nearly identical, Figure 2B shows greater dispersion in the unbalanced condition, suggesting more heterogeneous experiences of role diversity.
- Block II: Team Dynamics and Satisfaction. For team dynamics, role-balanced teams reported slightly lower satisfaction ($M = 3.64$, $SD = 0.74$) than role-unbalanced teams ($M = 3.74$, $SD = 1.00$). The estimated difference was -0.10 (95% CI: -0.74 to 0.54), and this difference was not statistically significant, $t(28) = -0.31$, $p = 0.755$. As in Section A, unbalanced teams exhibited greater variability, indicating that redundancy may yield uneven but occasionally positive team experiences.
- Block III: Perceived Team Performance and Effectiveness. Perceived performance showed the opposite trend. Role-balanced teams reported higher perceived effectiveness ($M = 3.74$, $SD = 1.07$) than role-unbalanced teams ($M = 3.48$, $SD = 1.04$). The estimated mean difference was 0.26 , with a 95% confidence interval from -0.52 to 1.04 . This difference was not statistically significant, $t(27) = 0.68$, $p = 0.501$. Despite

the lack of statistical significance, the positive direction of the effect suggests that in a section with higher baseline role diversity, complementary role structures may be more readily translated into a perception of improved team effectiveness.

3.6.3. Integrative Interpretation

Across both sections, none of the survey constructs showed statistically significant differences between role-balanced and role-unbalanced teams. Confidence intervals were wide and consistently included zero, reflecting limited statistical power at the team level. Nevertheless, the direction and dispersion of effects differed systematically between sections, supporting the broader conclusion that the impact of algorithmic role balancing on subjective team experience is contingent on baseline role distributions and section-specific dynamics rather than on role structure alone.

3.7. Objective Performance Outcomes

To complement self-reported perceptions, we analyzed objective academic performance across nine assessment components spanning individual and team-based activities. Performance data were aggregated at the team level by averaging the scores of all members within each group, enabling direct comparison between balanced and unbalanced conditions while accounting for the nested structure of students within teams. All assessments used a 0–10 grading scale consistent with institutional standards at UTPL.

3.7.1. Assessment Structure and Performance Metrics

The course employed a mixed assessment model integrating individual accountability with team-based deliverables (Table 6). Individual assessments included in-class activities, a midterm exam, theoretical defense presentations, and reading comprehension checks. Team-based assessments comprised the engineering report for the road geometric design project, oral presentation of findings, weekly progress infographics, and weekly project-based problem sets. The final course grade synthesized both individual and team performance.

Table 6. Team-Level Academic Performance by Assessment Component, Section, and Experimental Condition.

Assessment Component	Type	Balanced A *	Unbalanced A *	Balanced B *	Unbalanced B *
In-Class Activities	Individual	7.98 (2.03)	8.68 (0.97)	9.11 (0.73)	8.84 (0.97)
Midterm Exam	Individual	1.55 (1.28)	1.80 (1.26)	1.92 (1.82)	1.23 (1.00)
Engineering Report	Team	6.73 (0.86)	6.30 (1.85)	6.78 (0.74)	6.39 (1.72)
Oral Presentation	Team	9.24 (0.68)	9.13 (0.52)	9.18 (0.45)	6.71 (1.82)
Theoretical Defense	Individual	6.73 (0.85)	7.11 (0.92)	7.97 (0.60)	7.39 (0.59)
Weekly Infographics	Team	9.42 (0.36)	8.49 (0.64)	9.30 (0.44)	7.79 (1.35)
Weekly Problem Sets	Team	9.23 (0.84)	9.40 (0.32)	9.43 (0.27)	8.66 (0.93)
Reading Checks	Individual	8.23 (1.92)	7.91 (2.01)	8.71 (1.30)	9.09 (0.77)
Final Grade	Mixed	6.03 (0.67)	6.06 (0.68)	6.51 (0.71)	5.82 (0.59)

* Values represent means with standard deviations in parentheses. Section A: Balanced N = 20 students, Unbalanced N = 20. Section B: Balanced N = 16, Unbalanced N = 16.

3.7.2. Statistical Comparisons and Implications

Independent-samples *t*-tests were conducted for each assessment component, stratified by section. Given the exploratory scope of the study, interpretation emphasizes consistency of effects and practical relevance alongside statistical significance.

- In Section A, no statistically significant differences were observed between Balanced and Unbalanced teams for most assessment components. Performance on individual assessments (in-class activities, midterm exam, theoretical defense, reading checks) showed small and inconsistent differences, with confidence intervals consistently

spanning zero. Final grades were virtually identical between conditions, indicating no overall academic advantage associated with team composition. One exception emerged for weekly infographics, where Balanced teams significantly outperformed Unbalanced teams, $t(31) = 5.79, p < 0.001$. This task-specific advantage suggests that role-balanced teams may coordinate more effectively on recurrent, visually structured collaborative outputs, even in cohorts with relatively low role diversity.

- In contrast, Section B exhibited a clear and consistent pattern favoring Balanced teams on several team-based assessments. Balanced teams significantly outperformed Unbalanced teams on oral presentations, $t(14) = 4.96, p < 0.001$, as well as on theoretical defense, weekly infographics, and weekly problem sets (p values ranging from 0.010 to 0.001). Balanced teams also achieved significantly higher final course grades, $t(30) = 3.03, p = 0.005$. Differences on individual assessments (in-class activities, midterm exam, reading checks) were smaller and not statistically significant, indicating that compositional effects primarily manifested in collaborative, performance-intensive tasks rather than individual knowledge checks.

The perceptual findings provide a nuanced complement to the objective team composition results. While the algorithm produced clear and, in some cases, large structural differences between Balanced and Unbalanced teams—particularly in Section A—these compositional contrasts did not translate into equally pronounced differences in students' subjective evaluations of team experience.

Across both sections, perceptions of functional role diversity, team dynamics, satisfaction, and perceived performance were broadly similar between conditions, with substantial overlap in score distributions. This pattern suggests that structural differentiation at the role-composition level is not automatically or immediately perceived by students, especially in short-term academic projects and in cohorts with constrained baseline role diversity. In Section A, where role redundancy limited the range of feasible team configurations, role-unbalanced teams tended to show slightly more homogeneous and, in some cases, marginally more favorable perceptions across all three survey blocks. These trends are consistent with the idea that functional redundancy may support smoother coordination and reduce friction in contexts where complementary roles are difficult to implement meaningfully.

In Section B, which exhibited substantially higher baseline role diversity, perceptual patterns differed in direction. Although differences were small and not statistically significant, role-balanced teams showed a modest tendency toward higher perceived team performance and effectiveness. This alignment between higher structural role diversity and more positive performance perceptions suggests that complementary role configurations may be more readily recognized and valued by students when sufficient diversity exists at baseline. At the same time, the greater dispersion of scores observed in role-unbalanced teams indicates that redundancy can generate uneven team experiences, producing both highly functional and less effective groups.

Importantly, the perceptual findings must be interpreted in light of the methodological insights derived from the compositional analyses. In Section B, the unexpected inflation of the Complementarity Index in role-unbalanced teams—driven by variability in secondary role profiles rather than functional role differentiation—highlights a potential disconnect between mathematically defined complementarity and experientially meaningful role differentiation. Students' perceptions appear to align more closely with visible and functionally distinct primary roles (captured by Role Diversity and Overlap indices) than with abstract multidimensional variance captured by CI.

Taken together, the integration of compositional and perceptual evidence suggests that algorithmic role balancing is most likely to influence students' subjective team experience under specific conditions: when baseline role diversity is sufficiently high, when role dis-

tinctions are functionally salient, and when complementary roles translate into observable differences in task allocation and interaction. In contexts where these conditions are not met, the benefits of role balancing may remain largely structural and methodologically informative, rather than perceptually salient.

3.8. Exploratory Group-Level Analysis of Role Structure Metrics and Team Perceptions

An exploratory group-level analysis was conducted to examine the relationship between objectively derived role structure metrics—Complementarity Index (CI), Overlap Index (OI), and Role Diversity (RD)—and students' perceptions of team functioning. Perceptions were captured through three survey blocks: Functional Role Diversity and Complementarity (Block I), Team Dynamics and Satisfaction (Block II), and Perceived Team Performance and Effectiveness (Block III). All analyses were performed at the team level (18 teams across two course sections) and should be interpreted as exploratory due to the limited sample size.

3.8.1. Global Patterns Across Sections

Descriptive statistics revealed substantial variability in both structural team composition and perceived outcomes (Table 7). CI values ranged widely, indicating marked heterogeneity in internal role complementarity across teams. OI spanned the full theoretical range from complete complementarity (0) to complete overlap (1), while RD varied from highly redundant teams dominated by a single role to teams approaching full role coverage. Survey responses showed moderate-to-high average ratings across all three blocks, with some dispersion at the team level.

Table 7. Descriptive Statistics of Objective Team Composition Metrics and Survey-Based Perception Scores (Global, N = 18 Teams).

Variable	Mean	SD	Min	Max
Complementarity Index (CI)	11.37	6.85	2.79	26.64
Overlap Index (OI)	0.52	0.41	0.00	1.00
Role Diversity (RD)	2.33	1.24	1.00	4.00
Block I: Functional Role Diversity and Complementarity	3.91	0.55	2.80	4.90
Block II: Team Dynamics and Satisfaction	3.69	0.72	2.50	5.00
Block III: Perceived Team Performance and Effectiveness	3.65	0.81	1.60	4.80

At the global level, Spearman correlations suggested modest but theoretically coherent patterns (Figure 4). CI exhibited a positive association with Block I ($\rho \approx 0.35$), indicating that teams with higher algorithmic complementarity tended to perceive greater functional role differentiation. Associations between CI and Blocks II and III were weaker, suggesting that structural complementarity alone does not necessarily translate into more favorable perceptions of team climate or performance.

RD showed consistently positive correlations with all three blocks, with the strongest alignment again observed for Block I, reinforcing the conceptual correspondence between objective role variety and perceived functional diversity. In contrast, OI displayed weak to negative associations with perceived outcomes, most notably with Block III, suggesting that higher redundancy may be mildly detrimental to perceived effectiveness. Overall, these global patterns indicate that CI and RD capture partially distinct but relevant aspects of team functioning, with their strongest correspondence occurring for perceptions directly related to role differentiation rather than broader evaluative judgments.

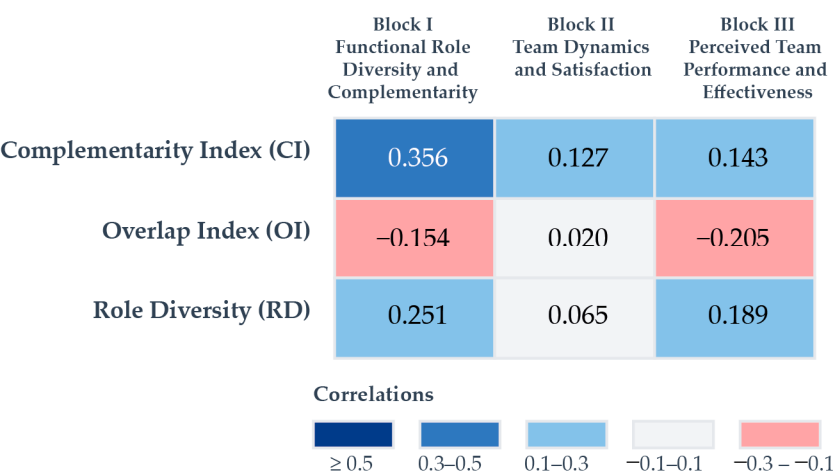


Figure 4. Global Spearman Correlation Matrix Between Team Composition Metrics and Survey Block Scores (N = 18 Teams).

3.8.2. Section-Level Contrasts

When analyses were disaggregated by section, markedly different patterns emerged (Table 8; Figure 5), highlighting the contextual sensitivity of role structure effects.

Table 8. Section-Level Descriptive Statistics of Team Composition Metrics and Perception Scores.

Variable	Section A (N = 10)		Section B (N = 8)	
	Mean	SD	Mean	SD
Complementarity Index (CI)	11.41	7.13	11.33	6.96
Overlap Index (OI)	0.60	0.41	0.42	0.43
Role Diversity (RD)	2.10	1.20	2.63	1.30
Block I: Functional Role Diversity and Complementarity	3.84	0.59	3.99	0.53
Block II: Team Dynamics and Satisfaction	3.64	0.76	3.74	0.71
Block III: Perceived Team Performance and Effectiveness	3.63	0.65	3.67	1.02

In Section A (N = 10 teams), RD exhibited strong positive correlations with all three survey blocks, particularly with Block I, indicating that teams with greater role variety consistently reported better functional differentiation, team dynamics, and perceived performance. CI also showed moderate positive associations, especially with Block I, suggesting close alignment between structural complementarity and students’ lived team experiences. Conversely, OI demonstrated strong negative relationships with Block I and moderate negative associations with Blocks II and III, implying that role redundancy was clearly perceived as dysfunctional in this context.

In Section B (N = 8 teams), the pattern was notably different. CI and RD showed weak or negative associations with several outcome blocks, while OI displayed moderate positive correlations with Block I and Block II. This reversal suggests that, within this section, role overlap was not necessarily experienced as inefficiency but may have been interpreted as shared competence, mutual backup, or functional flexibility. This interpretation is consistent with previously observed within-role variability and secondary role profiles in Section B.

Taken together, these contrasts underscore that the relationship between role structure and perceived team functioning is not uniform across sections and may depend on contextual factors such as task interpretation, coordination norms, or section-specific dynamics.

(A) Spearman Correlations by Section A (N = 10)

	Block I Functional Role Diversity and Complementarity	Block II Team Dynamics and Satisfaction	Block III Perceived Team Performance and Effectiveness
Complementarity Index (CI)	0.552	0.236	0.406
Overlap Index (OI)	−0.733	−0.273	−0.442
Role Diversity (RD)	0.818	0.455	0.527

(B) Spearman Correlations by Section B (N = 8)

	Block I Functional Role Diversity and Complementarity	Block II Team Dynamics and Satisfaction	Block III Perceived Team Performance and Effectiveness
Complementarity Index (CI)	−0.262	−0.310	−0.429
Overlap Index (OI)	0.548	0.524	0.119
Role Diversity (RD)	−0.357	−0.667	−0.310

Correlations

≥ 0.5 0.3–0.5 0.1–0.3 −0.3–−0.1 −0.5–−0.3 < −0.5

Figure 5. Section-Specific Spearman Correlation Matrices Between Team Composition Metrics and Survey Block Scores.

3.8.3. Balanced and Unbalanced Conditions

A complementary exploratory comparison between Balanced and Unbalanced conditions using non-parametric summaries is presented in Table 9. As intended by design, Balanced teams showed higher median CI and RD values and lower OI values, confirming the effectiveness of the role assignment procedure in generating structurally distinct team configurations.

Table 9. Non-Parametric Comparison of Balanced and Unbalanced Conditions for Team Composition Metrics and Perception Outcomes.

Variable	Balanced (N = 9) *		Unbalanced (N = 9) *	
	Median	IQR	Median	IQR
Complementarity Index (CI)	12.46	8.49–16.47	7.01	4.58–9.79
Overlap Index (OI)	0.00	0.00–0.67	1.00	0.40–1.00
Role Diversity (RD)	4.00	2.00–4.00	1.00	1.00–2.00
Block I: Functional Role Diversity and Complementarity	3.80	3.40–4.13	4.13	3.60–4.30
Block II: Team Dynamics and Satisfaction	3.60	2.80–3.87	3.70	3.35–4.30
Block III: Perceived Team Performance and Effectiveness	3.90	2.87–4.07	3.80	3.60–4.10

* Non-parametric summaries (median and IQR) used due to small sample sizes.

Despite these clear structural differences, survey outcomes were broadly comparable across conditions. Unbalanced teams occasionally reported slightly higher median scores for Blocks I and II, while Block III medians were nearly identical between conditions. Sub-

stantial overlap in interquartile ranges across all outcome measures indicates considerable within-condition variability.

This apparent disconnect between objective structure and subjective perception highlights an important point: perceived team quality is not determined solely by structural role metrics. Social dynamics, coordination practices, and individual adaptability may compensate for—or in some contexts even benefit from—role redundancy.

Overall, this exploratory team-level analysis suggests that role complementarity and role diversity are meaningfully, though modestly, associated with students' perceptions of team functioning, particularly for outcomes directly related to functional role differentiation. However, section-level analyses revealed pronounced heterogeneity in these relationships, emphasizing the importance of contextual and section-specific factors. Given the small number of teams and the exploratory nature of the analyses, these findings should be viewed as hypothesis-generating rather than confirmatory. Future research with larger samples and multilevel modeling frameworks will be necessary to more rigorously disentangle structural, contextual, and perceptual drivers of team effectiveness.

4. Discussion

This study investigated whether TREO-based algorithmic team formation could improve collaborative learning outcomes in civil engineering education, and whether optimal team formation parameters depend on baseline section characteristics. Our dual-optimization framework revealed three critical findings that challenge prevailing assumptions about role-based team formation: (1) baseline role diversity profiles fundamentally determine the feasibility and effectiveness of algorithmic team formation strategies, (2) higher baseline diversity does not guarantee superior team formation outcomes and may paradoxically complicate differentiation between balanced and overlapping conditions, and (3) structural role complementarity does not automatically translate into improved student perceptions of team functioning, particularly in sections with limited role diversity. These findings have significant implications for engineering educators seeking to implement evidence-based team formation practices and suggest that diagnostic assessment of section characteristics should precede any intervention.

4.1. Baseline Diversity

Contrary to intuitive expectations, Section B ($R_{ef} = 5.57$, $IDR = 1.39$, classified as “Diverse”) exhibited universally negative Separation Scores across all 80 tested configurations, indicating that the algorithm could not effectively differentiate role-balanced from role-unbalanced conditions. Meanwhile, Section A ($R_{ef} = 2.94$, $IDR = 0.73$, classified as “Redundant”) achieved strongly positive separation ($S_{max} = 6.564$), successfully creating distinct experimental conditions. This counterintuitive result challenges the assumption that higher baseline diversity facilitates better team formation outcomes (Borrego et al., 2013; Oakley et al., 2004).

Several mechanisms may explain this paradox. First, the Shannon index and its derivative metrics (R_{ef} , IDR) quantify the evenness of role distribution but do not capture configurational constraints imposed by specific patterns of role clustering. Section B's higher effective diversity likely reflected relatively balanced representation across multiple roles, but examination of the primary role distribution reveals uneven clustering within conditions: 37.95% of Unbalanced students were Innovators, while only 4.65% were Team Builders or Challengers. Such within-condition imbalances may have prevented the algorithm from assembling coherent role-balanced teams without exhausting certain role categories entirely, particularly given the 50/50 allocation constraint between conditions. In contrast, Section A's lower overall diversity, dominated by Doers (65–66% across condi-

tions), paradoxically provided predictability: the algorithm could reliably create overlap by clustering Doers and achieve balance by distributing the limited number of non-Doer roles strategically across groups.

Second, the complementarity paradox may reflect fundamental differences in how secondary role profiles distribute across sections. While primary role assignments capture the highest-scoring dimension for each student, TREO is inherently multidimensional, and students exhibit varying degrees of strength across non-primary roles. Section B students may have possessed more uniform secondary role profiles—that is, relatively flat z-score distributions across all six roles—making it difficult to create groups with genuinely distinct functional compositions even when primary roles differed. This interpretation is supported by the unexpected finding that Unbalanced groups in Section B exhibited higher Complementarity Index values ($M = 12.40$) than Balanced groups ($M = 10.24$), despite having significantly higher Overlap Index values (0.767 vs. 0.083) and lower Role Diversity (1.75 vs. 3.50). High CI values in Unbalanced groups suggest substantial within-role variance in secondary characteristics rather than genuine functional complementarity—students shared the same primary role but differed widely on non-dominant dimensions, inflating the range-based CI metric without providing the functional diversity intended by the Balanced condition.

This finding aligns with emerging research on multidimensional team composition, which suggests that simple diversity indices may inadequately capture the complex interplay between primary roles, secondary competencies, and situational role flexibility (Harrison & Klein, 2007; Meslec & Curşeu, 2015). Future research should investigate whether alternative metrics—such as profile similarity indices, Euclidean distances in six-dimensional role space, or conditional entropy measures—provide better predictive validity for team formation feasibility than Shannon-based approaches.

4.2. Structural Composition Versus Lived Experience

Despite successfully creating structurally distinct teams in Section A (large effect sizes for CI, OI, and RD), we observed no significant differences in student perceptions of team functioning across any of the three survey blocks (Functional Role Diversity, Team Dynamics, Performance). This pattern is consistent with findings from cooperative learning research showing that structural interventions do not automatically translate into improved outcomes without explicit process support (Johnson & Johnson, 2009). In fact, trends consistently favored role-unbalanced teams, with Unbalanced groups reporting numerically higher satisfaction (Block II: $M = 3.85$ vs. $M = 3.47$) and perceived performance (Block III: $M = 3.82$ vs. $M = 3.59$), though differences remained non-significant. This dissociation between objective team composition and subjective team experience represents a critical challenge for role-based team formation interventions.

Several theoretical frameworks may explain this perception gap. First, the functional substitutability hypothesis suggests that in highly structured engineering design tasks—such as road geometric design projects with clearly defined phases (traffic analysis, route alignment, horizontal/vertical design, consistency evaluation)—role diversity may be less critical than in open-ended creative tasks. When project requirements are well-specified and procedural, students with similar role profiles (e.g., multiple Doers) may distribute tasks sequentially or spatially rather than functionally, leveraging redundancy to provide mutual quality control, backup capacity, and workload balancing rather than experiencing it as dysfunctional competition (Mathieu et al., 2015). This interpretation aligns with findings from software engineering education showing that homogeneous teams sometimes outperform diverse teams on tightly constrained programming tasks where coordination costs outweigh complementarity benefits (Dzvonyar et al., 2018; Pieterse et al., 2006).

Second, the compensation hypothesis posits that students may adaptively adjust their behaviors to compensate for structural team limitations. In role-unbalanced teams, the absence of designated individuals for certain functions (e.g., Organizers in Doer-dominated groups) may have prompted emergent role differentiation, with some students voluntarily adopting secondary roles to fill functional gaps. Conversely, in algorithmically balanced teams, the explicit expectation of role complementarity may have created interpersonal friction when students' actual contributions did not align with their assessed TREO profiles—a phenomenon termed “role enforcement conflict” in organizational psychology (Salas et al., 2008). This phenomenon has been documented in organizational teams where explicit role assignments can create rigidity and reduce adaptive capacity compared to emergent role differentiation (Humphrey et al., 2009). Our survey included no items assessing actual role enactment versus anticipated roles, preventing direct evaluation of this mechanism, but future research should investigate whether algorithmically formed teams experience greater role rigidity and reduced adaptive flexibility compared to organically formed groups.

Critically, the dissociation between structural composition and perceptual outcomes observed in survey data is not fully mirrored in objective performance metrics. Analysis of academic performance data revealed task-specific and section-specific patterns that provide important context for interpreting null perceptual findings. In Section A, despite no overall difference in final grades between conditions ($M_{\text{Balanced}} = 6.03$, $M_{\text{Unbalanced}} = 6.06$, $p > 0.05$), Balanced teams significantly outperformed Unbalanced teams on weekly infographics ($M_{\text{Balanced}} = 9.42$, $M_{\text{Unbalanced}} = 8.49$, $t(31) = 5.79$, $p < 0.001$), suggesting that role-balanced composition may enhance coordination on recurrent, visually structured collaborative outputs even in low-diversity sections. In Section B, Balanced teams demonstrated consistent advantages across multiple team-based assessments, including oral presentations ($M_{\text{Balanced}} = 9.18$, $M_{\text{Unbalanced}} = 6.71$, $t(14) = 4.96$, $p < 0.001$), weekly infographics, weekly problem sets, and final course grades ($M_{\text{Balanced}} = 6.51$, $M_{\text{Unbalanced}} = 5.82$, $t(30) = 3.03$, $p = 0.005$). Notably, these objective performance differences emerged primarily in collaborative, performance-intensive tasks rather than individual knowledge checks, indicating that compositional effects manifest most clearly in coordinated team deliverables.

The pattern of objective performance gains without corresponding perceptual differences suggests that students may lack metacognitive awareness of how team composition influences outcomes, or that performance improvements occur through mechanisms (e.g., reduced coordination errors, better task allocation) that operate below the threshold of conscious perception. This interpretation aligns with research on implicit coordination in teams, where structural factors influence performance through automatic processes not accessible to explicit reflection (Rico et al., 2008). Alternatively, ceiling effects in survey responses (Block means ranged 3.47–3.99 on a 5-point scale) may have compressed variance and obscured perceptual differences that exist but cannot be captured by coarse self-report measures. Future research should incorporate process-tracing methods (e.g., communication pattern analysis, task allocation tracking) to illuminate the mechanisms linking composition to performance, and should employ more sensitive measurement approaches including peer evaluations, behavioral observations, and timestamped interaction logs.

Third, the temporal dimension of team development may be critical. Tuckman's classic model of group development (forming, storming, norming, performing) suggests that teams require time to negotiate functional structures regardless of initial composition. Our intervention assigned students to teams in Week 2, but the semester-long project allowed 14–16 weeks of collaboration before the final survey. This extended timeframe may have permitted both Balanced and Unbalanced teams to converge toward functional equilibria through iterative adjustment, erasing initial compositional advantages. Research on temporary project teams in industry suggests that compositional effects are strongest

during early task phases and attenuate as teams develop shared mental models and coordination routines (Marks et al., 2001; Mathieu et al., 2015).

Engineering educators should therefore consider whether role-based team formation is most beneficial for short-duration projects where teams lack time for organic adaptation, rather than semester-long collaborations where developmental processes may override initial configurations.

4.3. Context-Sensitivity and the Failure of Universal Optimization

A central contribution of this study is demonstrating that optimal team formation parameters depend critically on section-specific characteristics, invalidating one-size-fits-all approaches. Both sections converged on identical optimal parameters ($T = 2.75$, $n_g = 4$), yet achieved radically different Separation Scores (6.564 vs. -1.502), illustrating that the same algorithm with the same settings produces fundamentally different outcomes depending on population input characteristics. This finding has important practical implications: educators cannot simply adopt threshold and group size recommendations from prior studies (e.g., The Group for Organizational Effectiveness, 2012) without first assessing whether their student population exhibits the role diversity profile for which those parameters were optimized.

Our proposed diagnostic framework—computing R_{ef} and IDR before team formation—provides a practical solution. Sections with $IDR \geq 1.2$ (“Diverse”) theoretically possess sufficient role variety to create non-overlapping balanced teams, though our Section B results suggest additional configurational constraints may interfere. Sections with $IDR < 0.8$ (“Redundant”) will struggle to achieve meaningful role balance regardless of algorithmic sophistication, and educators in such contexts should consider alternative team formation criteria (skill diversity, personality traits, prior academic performance) or accept that role-based differentiation may be infeasible. Indeed, research on team composition optimization has increasingly emphasized the need for multi-attribute approaches that integrate personality, skills, and experience rather than relying on single frameworks (Wilde, 2011). Sections with $0.8 \leq IDR < 1.2$ (“Moderate”) occupy an uncertain middle ground where success likely depends on specific patterns of role clustering beyond what Shannon indices capture. Research on team composition in engineering contexts has similarly found that aggregated diversity metrics often mask important configurational patterns that determine team effectiveness (Kozlowski & Ilgen, 2006).

However, our study also reveals limitations of the IDR framework. Section B’s classification as “Diverse” incorrectly predicted positive separation capacity, suggesting that effective diversity (R_{ef}) alone provides insufficient diagnostic information. Future research should develop enhanced diagnostic metrics incorporating not only role frequency distributions but also measures of within-role cohesion, cross-role correlation structures, and configurationally permissible team compositions given allocation constraints (e.g., 50/50 splits between conditions). Machine learning approaches—training classifiers on section characteristics to predict achievable Separation Scores—may provide more robust feasibility assessment than theory-driven indices alone.

4.4. Measurement Considerations and Methodological Reflections

The psychometric analysis of the survey instrument revealed important patterns that inform interpretation of null findings. Internal consistency was acceptable to good across all three blocks ($\alpha = 0.77$ – 0.86), and inter-item correlations generally supported the intended three-construct structure. However, Section A exhibited substantially more heterogeneous correlation patterns than Section B, including negative correlations absent in Section B, suggesting section-specific measurement properties. The reverse-coded item (Q3: role

redundancy) performed particularly poorly, showing weak or inconsistent associations with other Block I items and elevated correlations with items from other blocks. This pattern is characteristic of method effects in small samples and suggests that reverse-coded items may introduce measurement noise rather than controlling for acquiescence bias in team-level contexts where consensus perceptions are aggregated.

More fundamentally, the modest correlations between objective role structure metrics (CI, OI, RD) and subjective perceptions (particularly in Section A: $\rho \approx 0.35$ between CI and Block I) raise questions about construct validity. If the Complementarity Index truly captures functional role differentiation, and if Block I items genuinely assess students' perceptions of "very different skills and knowledge" and "optimal combination of roles," stronger convergence would be expected. The weak observed correlations suggest either (a) the CI metric inadequately operationalizes complementarity as experienced by students, (b) students lack reliable insight into their teams' objective compositional characteristics, or (c) perceived complementarity depends more on interpersonal dynamics and communication quality than on structural role differences. The third interpretation finds support in the stronger correlations observed between Block I and Block II (team dynamics) than between Block I and objective metrics, suggesting that students' perceptions of role diversity are shaped substantially by process quality rather than compositional structure per se.

The weak-to-moderate correlations between objective compositional metrics and subjective perceptions warrant deeper theoretical consideration. Three non-mutually exclusive mechanisms may explain this disconnect:

- First, the temporal mismatch hypothesis: Structural metrics (CI, OI, RD) were computed at team formation (Week 2) based on initial TREO profiles, whereas perceptions were assessed post-project (Week 17) after 14+ weeks of collaboration. During this interval, teams may have undergone substantial role renegotiation and emergent differentiation that altered functional structures independent of initial composition. Research on team development suggests that assigned roles often converge with emergent roles only after prolonged interaction (Stewart et al., 2005), implying that Week 2 structural metrics may have limited predictive validity for Week 17 experiences. Future studies should assess role enactment longitudinally through repeated measurements to capture dynamic role evolution.
- Second, the construct misalignment hypothesis: The CI metric operationalizes complementarity as mathematical variance in z-scores across six dimensions, but students may conceptualize "complementarity" as functional differentiation in task allocation—a distinct construct not perfectly captured by multidimensional profile distances. Qualitative interviews with students could clarify how they subjectively define role diversity and whether these lay theories align with psychometric operationalizations. Alternative metrics such as task-based complementarity indices (e.g., proportion of unique task assignments per member) may demonstrate stronger convergence with perceptions.
- Third, the adaptation and compensation hypothesis: Teams possess substantial capacity to adapt structures and compensate for compositional limitations through behavioral adjustments, communication norms, and leadership emergence (LePine et al., 2008). In role-unbalanced teams, explicit discussion of functional gaps may prompt voluntary differentiation, while in algorithmically balanced teams, misalignment between assigned and preferred roles may create friction that degrades perceived complementarity. These dynamic processes would attenuate correlations between static compositional metrics and dynamic perceptual experiences. Process-oriented research examining how teams negotiate roles over time—and whether this negoti-

ation differs between balanced and unbalanced conditions—could illuminate these compensatory mechanisms.

Importantly, the modest correlations observed here do not invalidate structural metrics but rather highlight that composition represents an input factor whose effects are mediated by complex emergent processes (Ilgen et al., 2005). The IMOI (Input-Mediator-Output-Input) framework suggests that compositional inputs rarely produce direct effects on outcomes, but instead influence team processes (e.g., communication quality, conflict resolution) which in turn shape perceptions and performance. Our study assessed inputs (composition) and outcomes (perceptions, performance) but did not measure mediating processes—a critical gap that future research should address through process measures including communication network analysis, conflict frequency tracking, and real-time coordination assessments.

4.5. Implications for Engineering Education Practice

Despite null perceptual findings, this study offers several actionable insights for civil engineering educators. First, the dual-optimization framework itself—systematic exploration of $T \times n_g$ parameter space, computation of Separation Scores, selection of configurations maximizing experimental contrast—provides a rigorous methodological template for implementing and evaluating team formation interventions. Educators adapting this approach to their own courses should collect TREO data, compute R_{ef} and IDR as diagnostic checks, then empirically optimize parameters rather than relying on theoretical defaults.

Second, our findings suggest that role-based team formation may be most appropriate for specific educational contexts rather than as a universal best practice. Short-duration projects (≤ 4 weeks) may benefit more from compositional optimization than semester-long collaborations where teams self-organize. Highly ambiguous, open-ended design challenges may activate role complementarity more strongly than well-structured problems with procedural solutions. Research on problem types in engineering education supports this distinction, showing that ill-structured problems benefit more from diverse perspectives than well-structured technical exercises (Jonassen, 2000). Courses emphasizing interpersonal skill development (communication, conflict resolution, leadership) may prioritize compositional diversity even absent immediate performance benefits, while courses focused purely on technical deliverables may tolerate or even prefer homogeneous teams to minimize coordination costs. This finding aligns with recent meta-analytic evidence showing that diversity effects are highly context-dependent and moderated by task type and team processes (Bell et al., 2011).

Third, educators should carefully consider the trade-offs between algorithmic assignment and student autonomy. While our study used researcher-determined team assignments to enable experimental manipulation, self-selected teams may exhibit higher motivation and commitment despite suboptimal compositional balance (Chapman et al., 2006; Opdecam & Everaert, 2018). Hybrid approaches—allowing students to express preferences while algorithmically optimizing within constraint sets—may balance compositional goals with student agency.

Finally, transparency in team formation rationale deserves consideration. Our study did not inform students whether their teams were Balanced or Unbalanced to avoid expectancy effects, but post hoc debriefing about the role-based formation process could enhance students' metacognitive awareness of team composition dynamics and promote more intentional role differentiation behaviors. Educational interventions combining algorithmic formation with explicit instruction on TREO roles and complementarity strategies may yield stronger effects than compositional manipulation alone. Indeed, meta-analytic evidence suggests that team

training interventions produce stronger effects when combined with structural composition strategies rather than when applied independently (Salas et al., 2008).

4.6. Limitations and Future Directions

Several limitations qualify the scope of our findings. First, the sequential allocation procedure likely introduced selection bias favoring Balanced groups, as students with more distinctive role profiles were assigned first; although mean $n_{\text{StrongRoles}}$ did not differ between conditions, unobserved individual differences may have attenuated the intended experimental contrast. Second, the modest sample size (77 students across 18 teams) limited statistical power at the team level, resulting in wide confidence intervals that preclude definitive conclusions about small-to-moderate perceptual effects and underscore the need for replication with larger, multi-institutional samples. Third, the single-course, single-institution context constrains generalizability, as the highly structured, regulation-driven nature of a road geometric design course may not reflect dynamics in more open-ended or multidisciplinary engineering projects. The specificity of road geometric design—with its regulatory constraints and standardized procedures—may limit generalizability to other engineering domains where creativity and innovation are more central (Dym et al., 2005).

Measurement and modeling choices further delimit interpretation. Outcomes relied on post-project self-report surveys, preventing longitudinal assessment of team dynamics and limiting insight into the observed dissociation between structural composition and perceived experience; future work should integrate objective performance indicators, behavioral data, and developmental measures. In addition, our optimization framework explored only a subset of the possible parameter space, and the TREO operationalization employed here may not fully capture multidimensional complementarity, as suggested by inconsistent CI patterns despite successful overlap manipulation. Future research should therefore examine alternative allocation algorithms, richer role-space representations, and potential moderators (e.g., prior teamwork experience, project type), as well as qualitative inquiry to clarify why structural complementarity does not consistently translate into perceptual advantages. Research on team effectiveness in educational settings has identified several critical moderators including task interdependence, duration, and instructor support that warrant systematic investigation (Hoegl & Gemuenden, 2001; Loughry et al., 2007).

4.7. Theoretical Contributions and Broader Significance

Beyond practical implications for engineering education, this study contributes to broader theoretical understanding of team composition effects. The paradoxical finding that higher baseline diversity complicates rather than facilitates team formation challenges implicit assumptions in diversity research and highlights the importance of configurational thinking—considering not just marginal distributions of characteristics but how those characteristics combine subject to allocation constraints.

The dissociation between structural metrics and lived experience resonates with ongoing debates in organizational psychology about the relationship between team composition and team processes. Our findings support interactionist models positing that compositional characteristics interact with emergent dynamics rather than deterministically producing outcomes (Mathieu et al., 2015), challenging input-output models that dominated early team research. Engineering education researchers should adopt more nuanced models acknowledging that team formation establishes initial conditions that are then transformed through complex developmental processes.

Finally, this study demonstrates the value of pre-registered, theory-driven parameter optimization as an alternative to post hoc rationalization of arbitrary choices. By systematically exploring 80 configurations per section and transparently reporting all results

(Appendix A), we provide a methodological model for rigorous intervention research in educational settings. The dual-optimization approach—independently optimizing for each section while using identical procedures—balances standardization with contextualization, a tension central to educational research more broadly.

5. Conclusions

This quasi-experimental study proposes and validates a context-sensitive framework for TREO-based team formation in civil engineering education, demonstrating that baseline section diversity fundamentally constrains both the feasibility and impact of algorithmic team assignment. Contrary to common assumptions, higher baseline role diversity (Section B: $R_{ef} = 5.57$) hindered the generation of structurally distinct team configurations, whereas lower-diversity sections (Section A: $R_{ef} = 2.94$) enabled clear differentiation between balanced and unbalanced conditions. Importantly, even when structural manipulation was successful, it did not translate into improved student perceptions of complementarity, team dynamics, or performance, highlighting a disconnect between objective role composition and subjective experience. These findings challenge one-size-fits-all approaches to team formation and underscore the need for diagnostic, section-specific strategies. Methodologically, the study contributes a dual-optimization approach that systematically identifies empirically optimal configurations and provides a transparent, replicable template for adaptation across contexts. For civil engineering education, the results suggest that role-based team formation is best suited to specific instructional scenarios rather than universal application and that both low- and high-diversity sections require empirical validation of intended compositional effects. Overall, the study emphasizes that effective team-based learning depends on aligning pedagogical goals, section characteristics, and structural interventions within an evidence-based, context-aware framework.

Author Contributions: Conceptualization, Y.G.-R.; methodology, Y.G.-R.; software, P.C.-G. and Y.G.-R.; validation, Y.G.-R. and P.C.-G.; formal analysis, P.C.-G. and Y.G.-R.; investigation, Y.G.-R.; resources, Y.G.-R.; data curation, Y.G.-R.; writing—original draft preparation, P.C.-G. and Y.G.-R.; writing—review and editing, Y.G.-R. and P.C.-G.; visualization, Y.G.-R.; supervision, Y.G.-R.; project administration, Y.G.-R.; funding acquisition, Y.G.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad Técnica Particular de Loja, grant number POA VIN-56.

Institutional Review Board Statement: Ethical review and approval were waived for this study according to the regulations of the Universidad Técnica Particular de Loja (UTPL). Under the standardized procedures of the local Ethics Committee (CEISH-UTPL), this research is classified as “Research Without Risk” as it involves the anonymous evaluation of educational practices and pedagogical scenarios. The study was conducted in accordance with the Declaration of Helsinki and the Ecuadorian Organic Law on Personal Data Protection, ensuring participant anonymity and the use of aggregated data only.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Prior to the study, all participants were presented with a digital consent form outlining the study purpose, the voluntary nature of participation, the right to withdraw at any time, and a guarantee of anonymity. Participants were required to check a mandatory acknowledgment box to document their consent before proceeding. No identifiable private information or images are included in this publication; therefore, additional written informed consent for publication was not required.

Data Availability Statement: The original data presented in the study are openly available on Zenodo at <https://doi.org/10.5281/zenodo.18305293>. However, academic grades and final survey responses

cannot be shared due to privacy constraints. Access to these specific datasets is restricted to protect participant confidentiality.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-4, 2024 version) for minor editing and improvements in the English language of specific sections. Additionally, Claude (Anthropic, Claude 3, 2024 version) were utilized to review and correct certain segments of the R code script included in the Data Availability Statement. The authors have carefully reviewed and edited the output from these tools and take full responsibility for the content and accuracy of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TREO	Team Role Experience and Orientation survey
CI	Complementarity Index
OI	Overlap Index
RD	Role Diversity
IDR	Role Diversity Index
H	Shannon’s Diversity Index
Re _f	Effective Number of Roles (Shannon’s effective diversity)
T	Role Strength Threshold
N	Sample size
n _g	Group Size
S	Separation Score

Appendix A

Course Learning Objectives and Assessment Alignment

The Road Construction I course was designed around the following learning objectives, operationalized through the assessment components described in Section 2.1:

- 1. Apply traffic analysis procedures to characterize vehicular demand on road corridors. (assessed via in-class activities, weekly problem sets)
- 2. Design horizontal and vertical road alignments in compliance with Ecuadorian and AASHTO geometric design standards. (assessed via engineering report, weekly problem sets)
- 3. Evaluate road design consistency and safety using quantitative indicators. (assessed via theoretical defense, midterm exam)
- 4. Communicate technical design decisions clearly and persuasively in written and oral form. (assessed via engineering report, oral presentation)
- 5. Synthesize and consolidate weekly content through individual and team-based outputs. (assessed via weekly infographics, reading checks)
- 6. Demonstrate effective collaboration, role differentiation, and shared accountability within a project team. (assessed via post-project survey, team-based deliverables)

Objectives 1–5 were evaluated through both individual and team-based assessments. Objective 6 was the primary focus of the experimental manipulation in this study.

Appendix B

Full Optimization Results for Threshold (T) and Group Size (n_g) Combinations

This appendix reports the complete grid of optimization results obtained from the systematic evaluation of all threshold (T) and group size (n_g) combinations. For each configuration, the Separation Score (S) is presented independently for Section A and

Section B, along with the combined score (S_{ScoreSum}) and the minimum score across sections (S_{ScoreMin}). Table A1 is sorted in descending order of S_{ScoreSum} to facilitate identification of high-performing parameter configurations.

Table A1. Separation Scores for All Threshold (T) and Group Size (n_g) Combinations by Section.

#	T	n_g	S_{SectionA}	S_{SectionB}	S_{ScoreSum}	S_{ScoreMin}
1	2.75	4	6.56	−1.5	5.06	−1.5
2	2.75	3	6.14	−2.23	3.91	−2.23
3	3	3	3.97	−1.64	2.33	−1.64
4	3.75	4	4.42	−2.1	2.32	−2.1
5	2	4	6.42	−4.28	2.14	−4.28
6	0.25	4	6.13	−4.28	1.84	−4.28
7	0.5	4	6.13	−4.28	1.84	−4.28
8	0.75	4	6.13	−4.28	1.84	−4.28
9	1	4	6.13	−4.28	1.84	−4.28
10	2.5	3	5.28	−3.52	1.76	−3.52
11	3	4	5.16	−3.5	1.66	−3.5
12	2.5	6	4.81	−3.25	1.56	−3.25
13	1.5	4	5.8	−4.28	1.52	−4.28
14	2.25	3	5.22	−3.75	1.47	−3.75
15	2.75	6	5.32	−3.86	1.46	−3.86
16	2.25	4	5.84	−4.61	1.23	−4.61
17	2	3	4.59	−3.54	1.05	−3.54
18	2	6	5.31	−4.31	1	−4.31
19	0.25	3	4.51	−3.54	0.97	−3.54
20	0.5	3	4.51	−3.54	0.97	−3.54
21	0.75	3	4.51	−3.54	0.97	−3.54
22	1	3	4.51	−3.54	0.97	−3.54
23	0.25	6	5.11	−4.31	0.79	−4.31
24	0.5	6	5.11	−4.31	0.79	−4.31
25	0.75	6	5.11	−4.31	0.79	−4.31
26	1	6	5.11	−4.31	0.79	−4.31
27	3.75	3	3.3	−2.59	0.71	−2.59
28	1.5	3	4.24	−3.54	0.7	−3.54
29	1.75	4	4.86	−4.28	0.58	−4.28
30	2.25	6	4.95	−4.41	0.54	−4.41
31	2.5	5	4.61	−4.1	0.51	−4.1
32	1.5	6	4.8	−4.31	0.49	−4.31
33	3.5	4	3.95	−3.54	0.41	−3.54
34	2.75	5	5.05	−4.64	0.41	−4.64
35	2.25	5	4.57	−4.2	0.37	−4.2
36	1.25	4	4.59	−4.28	0.31	−4.28
37	4	6	4.21	−3.91	0.3	−3.91
38	2.5	4	4.95	−4.66	0.29	−4.66
39	3	6	4.2	−3.92	0.29	−3.92
40	3.25	6	3.51	−3.34	0.17	−3.34
41	1.75	3	3.61	−3.54	0.07	−3.54
42	3.75	6	3.81	−3.95	−0.14	−3.95
43	1.25	3	3.26	−3.54	−0.29	−3.54
44	1.75	6	3.99	−4.31	−0.33	−4.31
45	3.25	3	4.43	−4.83	−0.4	−4.83
46	4	4	3.16	−3.57	−0.41	−3.57
47	1.25	6	3.58	−4.31	−0.73	−4.31
48	3	5	4.88	−5.73	−0.86	−5.73
49	3.25	4	3.45	−4.33	−0.88	−4.33
50	4	3	3.08	−3.98	−0.89	−3.98
51	4	5	4.53	−5.43	−0.9	−5.43
52	3.5	6	3.18	−4.08	−0.91	−4.08
53	3.75	5	4.12	−5.22	−1.11	−5.22
54	0.25	5	5.35	−6.62	−1.28	−6.62
55	0.5	5	5.35	−6.62	−1.28	−6.62

Table A1. Cont.

#	T	ng	S _{SectionA}	S _{SectionB}	S _{ScoreSum}	S _{ScoreMin}
56	0.75	5	5.35	−6.62	−1.28	−6.62
57	1	5	5.35	−6.62	−1.28	−6.62
58	3.25	5	4.22	−5.56	−1.35	−5.56
59	2.75	7	3.57	−5.24	−1.68	−5.24
60	1.5	5	4.72	−6.62	−1.9	−6.62
61	2	5	4.66	−6.62	−1.96	−6.62
62	3.5	3	2.51	−4.49	−1.98	−4.49
63	3	7	4.08	−6.16	−2.08	−6.16
64	1.75	5	4.51	−6.62	−2.12	−6.62
65	0.25	7	3.61	−6.17	−2.56	−6.17
66	0.5	7	3.61	−6.17	−2.56	−6.17
67	0.75	7	3.61	−6.17	−2.56	−6.17
68	1	7	3.61	−6.17	−2.56	−6.17
69	2	7	3.57	−6.17	−2.59	−6.17
70	3.5	5	3.07	−5.75	−2.68	−5.75
71	1.5	7	3.4	−6.17	−2.77	−6.17
72	1.25	5	3.82	−6.62	−2.8	−6.62
73	2.5	7	2.76	−5.58	−2.82	−5.58
74	3.25	7	3.04	−5.88	−2.84	−5.88
75	1.75	7	3.15	−6.17	−3.01	−6.17
76	4	7	2	−5.46	−3.46	−5.46
77	2.25	7	3.16	−6.64	−3.48	−6.64
78	1.25	7	1.63	−6.17	−4.53	−6.17
79	3.75	7	1.71	−6.26	−4.55	−6.26
80	3.5	7	0.59	−6.09	−5.5	−6.09

References

- Bell, S. T., Villado, A. J., Lukasik, M. A., Belau, L., & Briggs, A. L. (2011). Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *Journal of Management*, 37(3), 709–743. [\[CrossRef\]](#)
- Borrego, M., Karlin, J., McNair, L. D., & Beddoes, K. (2013). Team effectiveness theory from industrial and organizational psychology applied to engineering student project teams: A research review. *Journal of Engineering Education*, 102(4), 472–512. [\[CrossRef\]](#)
- Boyce-Jacino, C. M., & Harrington, N. T. (2022). *Optimizing team staffing: A review of computational approaches to team formation*. Available online: <https://apps.dtic.mil/sti/citations/trecms/AD1179137> (accessed on 10 January 2026).
- Burris-Melville, T. S., Burris, S. T., & Bledsoe, K. (2024). *Empowering teams in higher education*. IGI Global. [\[CrossRef\]](#)
- Castro, M. (2012). Highway design software as support of a project-based learning course. *Computer Applications in Engineering Education*, 20(3), 468–473. [\[CrossRef\]](#)
- Chapman, K. J., Meuter, M., Toy, D., & Wright, L. (2006). Can't we pick our own groups? the influence of group selection method on group dynamics and outcomes. *Journal of Management Education*, 30(4), 557–569. [\[CrossRef\]](#)
- Curşeu, P. L., & Pluut, H. (2013). Student groups as learning entities: The effect of group diversity and teamwork quality on groups' cognitive complexity. *Studies in Higher Education*, 38(1), 87–103. [\[CrossRef\]](#)
- Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., & Leifer, L. J. (2005). Engineering design thinking, teaching, and learning. *Journal of Engineering Education*, 94(1), 103–120. [\[CrossRef\]](#)
- Dzvonyar, D., Alperowitz, L., Henze, D., & Bruegge, B. (2018). Team composition in software engineering project courses. In *SEEM '18: Proceedings of the 2nd international workshop on software engineering education for millennials* (pp. 16–23). ACM. [\[CrossRef\]](#)
- Fausett, C. M., Simonson, R. J., Griggs, A. C., & Keebler, J. R. (2021). Team role experience and orientation dimensions in predicting team performance in a spaceship bridge simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1), 812–816. [\[CrossRef\]](#)
- Fiegel, G. L., & Denatale, J. S. (2011). Civil engineering capstone design: Team formation, preparation, and performance. *The International Journal of Engineering Education*, 27(6), 1295–1307.
- García-Ramírez, Y. (2024). Implementing multiple different active learning techniques in online road geometric design courses. *Advances in Engineering Education*, 12(1), 25–44. [\[CrossRef\]](#)
- García-Ramírez, Y. (2025). Impact of team role distribution on performance in civil engineering education: A TREO survey analysis. *Edelweiss Applied Science and Technology*, 9(3), 1403–1414. [\[CrossRef\]](#)

- Garrido de la Torre, M. E., Pérez-Zuriaga, A. M., Martínez-Ibáñez, V., López-Maldonado, G., & Cuadrado-Tarodo, Á. (2020). PIME aprendizaje basado en proyectos: Ingeniería de carreteras e ingeniería geotécnica. In *IN-RED 2020: VI congreso de innovación educativa y docencia en red*. Editorial Universitat Politècnica de València. [CrossRef]
- Hadgraft, R. G., & Smith, D. W. (2008). Civil engineering education of the future. In *Proceedings of the 2008 AaeE Conference* (pp. 1–12). Australasian Association for Engineering Education. Available online: https://acquire.cqu.edu.au/articles/conference_contribution/Civil_engineering_education_of_the_future/13425377 (accessed on 17 January 2026).
- Han, J. (2020). Development of a teamwork skill scale for engineering students. *The International Journal of Engineering Education*, 36(1), 483–490.
- Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, 32(4), 1199–1228. [CrossRef]
- Hoegl, M., & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, 12(4), 435–449. [CrossRef]
- Humphrey, S. E., Morgeson, F. P., & Mannor, M. J. (2009). Developing a theory of the strategic core of teams: A role composition model of team performance. *Journal of Applied Psychology*, 94(1), 48–61. [CrossRef]
- Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, 56, 517–543. [CrossRef]
- Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, 38(5), 365–379. [CrossRef]
- Johri, A. (2023). Introduction to the international handbook of engineering education research (IHEER). In *International Handbook of Engineering Education Research*. Routledge. [CrossRef]
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63–85. [CrossRef]
- Karimi, A., & Manteufel, R. D. (2020, June 22–26). *Performance balanced team formation for group study and design projects*. 2020 ASEE Virtual Annual Conference Content Access, Virtual Online. [CrossRef]
- Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, Supplement*, 7(3), 77–124. [CrossRef]
- Kurtis Younkin, S. (2017). *Mixed methods assessment of project-based learning in a senior-level highway design course* [Master's thesis of Science on Civil Engineering (Transportation Engineering) Major, Iowa State University].
- LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2), 273–307. [CrossRef]
- Loughry, M. L., Ohland, M. W., & Dewayne Moore, D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement*, 67(3), 505–524. [CrossRef]
- Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review*, 26(3), 356. [CrossRef]
- Mathieu, J. E., Tannenbaum, S. I., Kukenberger, M. R., Donsbach, J. S., & Alliger, G. M. (2015). Team role experience and orientation: A measure and tests of construct validity. *Group & Organization Management*, 40(1), 6–34. [CrossRef]
- Meslec, N., & Curşeu, P. L. (2015). Are balanced groups better? Belbin roles in collaborative learning groups. *Learning and Individual Differences*, 39, 81–88. [CrossRef]
- Michalaka, D., & Golub, M. (2016, June 26–29). *Effective building and development of student teamwork using personality types in engineering courses*. 2016 ASEE Annual Conference & Exposition (pp. 1–14), New Orleans, LA, USA. [CrossRef]
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2600–2606. [CrossRef] [PubMed]
- Oakley, B., Felder, R. M., Brent, R., & Elhajj, I. (2004). Turning student groups into effective teams. *Journal of Student Centered Learning*, 2(1), 9–34.
- Opdecam, E., & Everaert, P. (2018). Seven disagreements about cooperative learning. *Accounting Education*, 27(3), 223–233. [CrossRef]
- Paoletti, J., Bisbey, T. M., Reyes, D. L., Wettergreen, M. A., & Salas, E. (2020). A checklist to diagnose teamwork in engineering education. *The International Journal of Engineering Education*, 36(1), 365–377.
- Pérez-Acebo, H. (2017). Teaching geometric design of roads by means of project-based learning. In *INTED2017 proceedings* (Vol. 1, pp. 229–238). IATED. [CrossRef]
- Pieterse, V., Kourie, D. G., & Sonnekus, I. P. (2006). Software engineering team diversity and performance. In *ACM international conference proceeding series* (Vol. 204, pp. 180–186). ACM. [CrossRef]
- R Core Team. (2023). *R: A language and environment for statistical computing*. R foundation for statistical computing. Available online: <https://www.r-project.org/> (accessed on 19 December 2025).
- Rico, R., Sánchez-Manzanares, M., Gil, F., & Gibson, C. (2008). Team implicit coordination processes: A team knowledge-Based approach. *Academy of Management Review*, 33(1), 163–184. [CrossRef]

- Salas, E., Diaz-Granados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, 50(6), 903–933. [CrossRef]
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. [CrossRef]
- Stewart, G. L., Fulmer, I. S., & Barrick, M. R. (2005). An exploration of member roles as a multilevel linking mechanism for individual traits and team outcomes. *Personnel Psychology*, 58(2), 343–365. [CrossRef]
- Swanson, E., McCulley, L. V., Osman, D. J., Scammacca Lewis, N., & Solis, M. (2019). The effect of team-based learning on content knowledge: A meta-analysis. *Active Learning in Higher Education*, 20(1), 39–50. [CrossRef]
- The Group for Organizational Effectiveness. (2012). *Team role experience and orientation (TREO) exercise*. Available online: <https://rdl.train.army.mil/catalog-ws/view/ARI-MSCP/pdfs/TREO-exercise-working-document-Mar12.pdf> (accessed on 29 November 2025).
- Van den Beemt, A., MacLeod, M., Van der Veen, J., Van de Ven, A., van Baalen, S., Klaassen, R., & Boon, M. (2020). Interdisciplinary engineering education: A review of vision, teaching, and support. *Journal of Engineering Education*, 109(3), 508–555. [CrossRef]
- Wilde, D. J. (2011). *Jung's personality theory quantified*. Springer London. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.