ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

Scenario-Driven Assessment of Engineering Group Work Integrating AI-Assisted Behavioral Coding and the Individual Weighting Factor Algorithm

Junxiao Zhu[1,a], Shiyi Wang[1,b], Chenliang Chu[1,c], Yu Liang[1,d], Yongxian Li[2,e], Ping Gao[3,f], Yiyang Liang[1,g], Yingjun Liu[1,h], Liang Qin[1,i,*], Deyun Ma[1,j,*]

[1]School of Food and Pharmaceutical Engineering, Zhaoqing University, Zhaoqing, 526061, China
[2]Guangzhou Juli Pharmaceutical Development Co., Ltd., Guangzhou, 510220, China
[3]Zhanjiang Institute for Food and Drug Control, Zhanjiang, 524000, China
[a]zhujunxiao@zqu.edu.cn, [b]shiyi_wang2877@163.com, [c]ccl15918621025@163.com, [d]liangyu.127@foxmail.com, [e]571718020@qq.com, [f]gaoyang2008694@126.com, [g]2062753100@qq.com, [h]1930372474@qq.com, [i]liangq@zqu.edu.cn, [j]mady@zqu.edu.cn
**Corresponding authors*

Abstract: *To address two long-standing problems in engineering group assessment (the difficulty of quantifying individual contributions and the difficulty of identifying social loafing), this study proposes a "scenario-driven" assessment model integrating the Leaderless Group Discussion (LGD) scenario, Bales' Interaction Process Analysis (IPA) coding, and the Individual Weighting Factor (IWF) algorithm, with AI introduced to improve data collection efficiency. An empirical study was conducted with 84 pharmaceutical engineering students (12 groups of seven) using a high-fidelity Good Manufacturing Practice (GMP) case. Results showed IPA coding reliability of $\kappa = 0.87$ and peer rating reliability of ICC = 0.79; the IWF algorithm effectively identified social loafing and significantly improved grade differentiation for low contributors; constructive-conflict groups achieved significantly higher risk-identification coverage than weak-conflict groups (87.3% vs. 52.1%, $p < 0.001$); speaking frequency correlated only weakly with contribution weight ($r = 0.23$), whereas being cited by peers correlated strongly ($r = 0.76$, $p < 0.001$). The model provides a workable approach to process-based assessment for engineering education accreditation.*

Keywords: *Educational Evaluation Reform; Engineering Education; Artificial Intelligence; Leaderless Group Discussion; Process-Based Assessment*

1. Introduction

China's *Overall Plan for Deepening Educational Evaluation Reform in the New Era* explicitly calls for "improving outcome-based evaluation, strengthening process evaluation, exploring value-added evaluation, and refining comprehensive evaluation." The Washington Accord and the *Standards for Engineering Education Accreditation* (2024 revision) require graduates to demonstrate the core competencies of solving complex engineering problems and collaborating effectively in teams [1-3]. Group-based cooperative learning has become the mainstream instructional model in engineering courses, yet its assessment continues to face three persistent pain points.

First, individual contributions are difficult to quantify. Learning occurs collectively, whereas assessment must ultimately be assigned to individuals, creating an "individual attribution paradox" [4]. In eight-person groups, individual output falls to about 49% of the level achieved when working alone [5], and a seven-person group entails 21 interaction paths along which loafing is easily concealed. Karau and Williams' meta-analysis identified three driving mechanisms: diffusion of responsibility, reduced evaluation apprehension, and free-rider motivation [6]. Traditional intra-group peer review suffers from leniency and convergent scoring, lacking a scientific mechanism for converting qualitative judgments into quantitative adjustments. Second, process data collection is inefficient: conventional assessment relies on manual video observation and subjective coding, which is costly and difficult to scale. Third, assessment scenarios lack fidelity: paper-and-pencil tests reach only lower-order cognitive levels such as "remembering" and "understanding" [7], not higher-order abilities such as "analyzing" and "evaluating."

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

The assessment dilemma is particularly acute in the Good Manufacturing Practice (GMP) course within pharmaceutical engineering programs: GMP decisions carry severe consequences (the 2008 heparin contamination incident caused serious patient harm); practical scenarios involve trade-offs among technical compliance, production efficiency, and cost; and ethics education remains confined to theory, leaving students who "know" but cannot "act" [8]. High-fidelity, contextualized assessment instruments are needed to close this gap between knowing and doing.

This study focuses on one core question: how can a group assessment model be constructed that both elicits authentic engineering decision-making behavior and fairly quantifies individual contributions? The innovative contributions of this study include: (1) constructing a high-fidelity GMP scenario-based assessment framework; (2) integrating Bales' Interaction Process Analysis (IPA) coding with the IWF algorithm to form a complete assessment evidence chain; and (3) exploring a classroom application pathway for AI-assisted behavioral coding.

2. Construction of the Assessment Model

2.1 Design Philosophy: Scenario–Mechanism Coupling

This study proposes "scenario–mechanism coupling" as the core design philosophy: effective group assessment requires the simultaneous alignment of scenario fidelity and the quantification mechanism. If a high-fidelity scenario lacks a fair quantification mechanism, the rich behavioral evidence it generates will be wasted; conversely, if a sophisticated quantification algorithm is applied to a low-fidelity scenario, it cannot produce meaningful differentiation, because the task itself fails to elicit genuine differences in competence. This philosophy is supported by Guzmán's Monte Carlo simulations, which demonstrated that "simple methods perform best" and that increasing algorithmic complexity does not significantly improve accuracy [9]. Rather than pursuing algorithmic sophistication, then, assessment design effort is better invested in improving the quality of the assessment scenario.

Building on this philosophy, this study integrates Leaderless Group Discussion (LGD) scenario-based assessment theory [10, 11], social loafing intervention mechanisms [6], and the IWF algorithm [12] into a closed-loop evidence chain of "scenario elicitation → behavioral coding → contribution weighting → individual assessment." Three key features of LGD (ill-structured scenarios with no standard answers, naturally emergent roles without a pre-assigned leader, and processes that can be quantified through video coding) fit the assessment of "complex engineering problems" well. The model further integrates three loafing intervention strategies: making contributions visible through IPA coding and video recording (identifiability); driving IWF-based grade adjustment through anonymous peer evaluation (evaluation pressure); and stimulating professional identity through a high-fidelity disciplinary scenario (intrinsic motivation).

2.2 Three-Dimensional Assessment Indicator System

In accordance with the *Standards for Engineering Education Accreditation*, a three-dimensional indicator system with balanced weights was constructed (Table 1). Task performance and process performance each account for 40%, giving equal weight to process and outcome; peer evaluation accounts for 20%, giving peer review enough moderating power, with instructor scoring as a backstop.

Table 1. Assessment Indicator System.

Dimension	Weight	Assessment Content	Corresponding Graduate Attribute	Data Source
Task Performance	40%	Accuracy of regulatory citation, proper use of analytical tools, feasibility of the proposed solution	Graduate Attribute 3	Instructor scoring
Process Performance	40%	Frequency and quality of behavioral events across the 12 IPA categories	Graduate Attributes 6 and 7	Video coding
Peer Evaluation	20%	Task contribution, idea quality, collaborative attitude (input to the IWF algorithm)	—	Intra-group peer review

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

2.3 Core Mechanism of the IWF Algorithm

Consider a learning group of n students. Let r_{ji} denote the contribution rating given by member j to member i ($i \neq j$). The mean peer rating received by member i is:

$$R_i = \frac{1}{n-1}\sum_{j=1,\, j\neq i}^{n} r_{ji} \tag{1}$$

The group mean peer rating is:

$$\bar{R} = \frac{\sum_{i=1}^{n} R_i}{n} \tag{2}$$

On this basis, the Individual Weighting Factor is defined as:

$$IWF_i = \frac{R_i}{\bar{R}} \tag{3}$$

The final individual score is computed through a secondary distribution model:

$$S_i = T_i \times 0.4 + P_i \times 0.4 + \bar{R} \times IWF_i \times 0.2 \tag{4}$$

where T_i is the instructor's task-performance score, P_i is the process-performance score derived from IPA coding, and $\bar{R} \times IWF_i$ equals member i's mean peer rating. Equation (4) ensures that the IWF exerts its redistributive effect only within the peer evaluation dimension (20%), and because the mean IWF within a group is constantly 1.0, group grades are redistributed rather than inflated. The validity of the algorithm is strengthened by three design features: exclusion of self-ratings (to prevent strategic manipulation), multi-dimensional rating (to reduce halo effects), and instructor scoring as a backstop (to guard against peer evaluation failure) [9, 12].

2.4 AI-Assisted IPA Behavioral Coding

Discussion behaviors were coded using the IPA framework [13, 14], which classifies speaking turns into 12 categories across four domains: positive socio-emotional reactions, task answers, task questions, and negative socio-emotional reactions. Two coders completed 20 hours of training. An "AI-assisted pre-coding plus human calibration" model was introduced: Whisper automatic speech recognition performs automatic turn segmentation and speaker attribution, replacing manual frame-by-frame counting; a pre-trained language model performs preliminary IPA category classification, providing reference annotations for the coders; and semantic similarity matching assists in identifying "idea citation" behaviors (i.e., instances in which a member's contribution is subsequently invoked or developed by other members), which are then confirmed manually by the coders. At the pre-coding stage, the agreement between AI and human coding reached $\kappa = 0.79$; after calibration, the formal coding reliability reached $\kappa = 0.87$, corresponding to "almost perfect agreement" [15]. Compared with fully manual coding, the AI-assisted model reduced the coding workload by approximately 40%.

3. High-Fidelity Scenario Case Design

Grounded in situated cognition theory [16], the cases were developed following four principles: technical authenticity (embedding domain knowledge barriers so that professional competence, not generic expression skills, determines performance), ethical tension (conflicts between legitimate values such as efficiency versus safety), stakeholder conflict (multi-role bargaining that prevents premature consensus), and solution openness (no single correct answer). These principles are transferable across disciplines: civil engineering scenarios can be built around safety margin versus cost conflicts, and software engineering scenarios around delivery deadline versus security trade-offs.

The primary assessment case is an engineering change control dispute: in an aseptic preparation workshop, the water-for-injection circulation pump failed during the night shift, and the maintenance engineer, finding the 316L stainless-steel pump out of stock, replaced it with a 304 pump without authorization as a claimed "like-for-like replacement," bypassing change control; two weeks later, rouge and microbial abnormalities were detected in the water system. The case embeds three layers of tension: technical (304 steel's lower molybdenum content gives weaker corrosion resistance than 316L, and the substitution altered the system's validated state [8]); ethical (the engineer faced a genuine dilemma, since keeping the pump offline risked shutting down production); and institutional (whether students can move beyond "punishing the individual" to propose innovations such as an emergency change fast-track procedure). Students were assigned roles including maintenance engineer, QA

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

manager, production supervisor, and engineering director, naturally forming multi-party bargaining. Two training cases, an out-of-specification (OOS) investigation and an aseptic process failure, were used for practice sessions.

4. Empirical Results and Analysis

Eighty-four third-year students majoring in pharmaceutical engineering at a Chinese university were randomly assigned to 12 groups of seven. The sample size, verified through G*Power analysis, met the requirement for detecting medium effect sizes ($\alpha = 0.05$, $1 - \beta = 0.80$, $d \geq 0.62$). The experimental procedure was: training discussions (two sessions) → a formal 40-minute video-recorded LGD → anonymous peer evaluation → AI-assisted coding → IWF calculation. A total of 2,847 speaking turns were collected. Informed consent was obtained throughout; video recordings were stored in encrypted form, and peer evaluations remained anonymous and confidential.

4.1 Reliability and Validity of the Assessment Instruments

Table 2. Summary of Reliability and Validity Tests.

Test Content	Indicator	Result	Criterion	Conclusion
IPA coding reliability	Cohen's κ	0.87 (0.81–0.93)	≥ 0.81: almost perfect	Almost perfect
Peer rating reliability	ICC (2,1)	0.79 [0.72, 0.85]	≥ 0.75: good	Good
Criterion validity	Pearson r	0.71***	≥ 0.50: strong	Strong correlation
Systematic bias	Paired t-test	t = 1.24, p = 0.22	p > 0.05: no bias	No systematic bias

Note. ***p < 0.001.

All indicators reached the level of "good" or above (Table 2). Criterion validity was supported by a strong correlation between the IWF and the instructor's individual process ratings ($r = 0.71$, $p < 0.001$). Among the 12 IPA behavioral categories, coding agreement was highest in the task-answers domain (giving suggestions, opinions, and information; $\kappa = 0.89$–0.93) and relatively lower in the negative socio-emotional domain ($\kappa = 0.81$–0.84), reflecting the inherent coding challenges posed by the ambiguity of negative affective behaviors in tone and intensity. The IWF-adjusted scores ($M = 79.6$, $SD = 15.8$) did not differ significantly from instructor scores ($M = 78.2$, $SD = 14.3$), indicating no systematic bias.

4.2 Precise Identification of Social Loafing

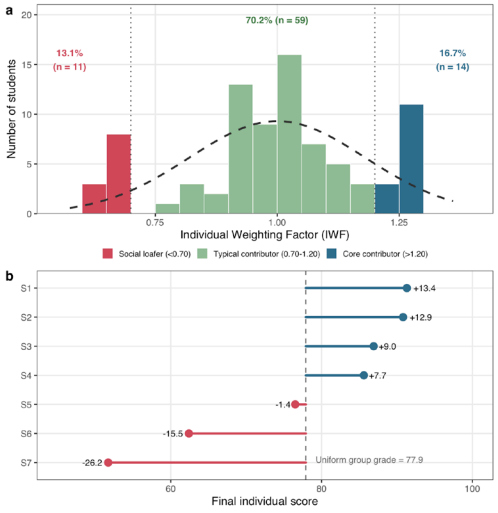

Figure 1. (a) Distribution of IWF scores for the 84 students (M = 1.00, SD = 0.18; tiers: IWF < 0.70 social loafers, 0.70–1.20 typical contributors, IWF > 1.20 core contributors; dashed curve = normal reference); (b) IWF-based grade redistribution in Group G-07 relative to uniform group grading (77.9).

As shown in Figure 1a, the IWF scores of the 84 students formed an approximately symmetric, unimodal distribution centered on 1.00 ($M = 1.00$, $SD = 0.18$), with heavier tails than a normal curve at

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

both ends. Of all students, 13.1% were significant loafers (IWF < 0.70), 70.2% fell within the typical contribution range (0.70 ≤ IWF ≤ 1.20), and 16.7% were core contributors (IWF > 1.20), a clear three-tier distribution. Group G-07 is a typical case (Table 3).

Table 3. Behavioral Characteristics and Multi-Dimensional Scores of Group G-07 Members.

Member	Emergent Role	Task Score	Process Score	Peer Rating Mean	IWF	Final Score	Speaking Turns	Times Cited
S1	Core leader	92	88	96.5	1.23	91.3	28	12
S2	Technical expert	95	85	94.0	1.20	90.8	18	15
S3	Process facilitator	80	92	90.5	1.15	86.9	24	8
S4	Critical thinker	88	85	82.0	1.04	85.6	21	6
S5	Follower	75	78	76.5	0.97	76.5	15	3
S6	Ineffective contributor	60	65	62.0	0.79	62.4	32	1
S7	Social loafer	55	50	48.5	0.62	51.7	6	0
Mean	—	77.9	77.6	78.6	1.00	77.9	20.6	6.4

S7 (the social loafer) remained disengaged throughout, contributing only 6 effective speaking turns and being cited 0 times; the peer rating mean of 48.5 was far below the group mean of 78.6. Under uniform grading, S7 would have shared the group score of 77.9; after IWF adjustment, the score dropped to 51.7, a failing grade 26.2 points (33.6%) below the group mean. This breaks the shelter that egalitarian grading provides for free-riding and sends students a clear signal that reward matches contribution. Psychologically, the IWF algorithm suppresses loafing motivation through two pathways, reinstating evaluation apprehension and raising the cost of free-riding, which provides educational intervention evidence for the collective effort model [6].

In contrast to S7, the core contributors earned high ratings through different pathways. S1, the naturally emergent core leader, structured the discussion, synthesized viewpoints, and managed time (28 speaking turns, 12 citations, IWF = 1.23). S2, the technical expert, spoke 18 times but was cited 15 times, the highest in the group (IWF = 1.20). One led through organization, the other through professional input; the algorithm accommodates both.

Figure 1b visualizes this redistribution: under uniform grading every member would have received 77.9, whereas IWF adjustment produced a spread from 51.7 to 91.3. The gains of the four above-average contributors (+7.7 to +13.4) are funded entirely by the deductions of S5, S6, and S7 (−1.4, −15.5, and −26.2), keeping the group total unchanged and confirming that the algorithm redistributes rather than inflates grades.

4.3 Constructive Conflict Enhances Decision Quality

The analysis of S4 (the critical thinker) shows another mechanism at work in LGD interaction. The group initially inclined toward attributing the contamination to the maintenance engineer's individual violation. S4 questioned this hasty conclusion and proposed the alternative hypothesis that "the standard operating procedure (SOP) itself may contain design flaws." This provoked a brief dispute, and S4's peer rating ended up only slightly above the group mean (IWF = 1.04), which did not fully reflect the cognitive value of this contribution. Yet the remark broke the group's mental set: the discussion shifted from pursuing individual responsibility to examining systemic loopholes in the change control system. Based on IPA coding, the 12 groups were divided into strong-conflict and weak-conflict groups (six each), and their decision quality was compared (Table 4).

Table 4. Effect of Constructive Conflict on Solution Quality.

Group Type	Number of Groups	Risk Identification Coverage (%)	Regulatory Provisions Cited	Innovative Suggestions
Strong-conflict	6	87.3 ± 6.2	8.5 ± 1.8	4.2 ± 1.1
Weak-conflict	6	52.1 ± 8.9	3.8 ± 1.2	1.5 ± 0.8
t value	—	7.95***	5.32***	4.86***
Cohen's d	—	4.59	3.07	2.81

Note. ***p < 0.001. Risk identification coverage = the proportion of expert-predetermined risk points covered by the group's solution.

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

The strong-conflict groups outperformed the weak-conflict groups in risk identification coverage by 35.2 percentage points (t = 7.95, d = 4.59), consistent with Jehn's theory of task conflict [17] (Figure 2a).

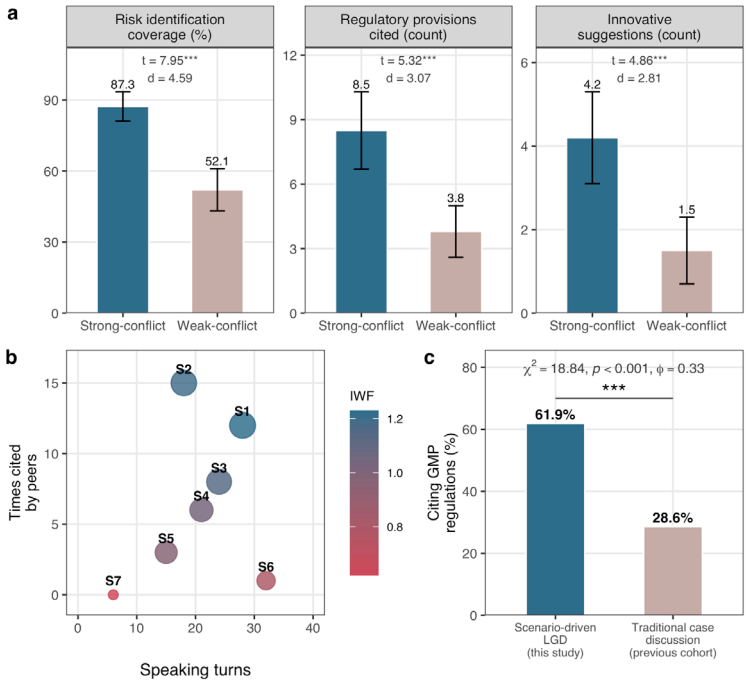

*Figure 2. (a) Decision-quality outcomes of strong-conflict versus weak-conflict groups (means ± SD; ***p < 0.001); (b) speaking turns versus times cited by peers in Group G-07 (bubble size and color = IWF); (c) percentage of students citing GMP regulations to defend compliance, by discussion format.*

The role-differentiation design caused conflict to be perceived as a "role requirement" rather than a "personal attack": when an external role framework provides a legitimate attribution for disagreement, participants are more likely to experience conflict as task-oriented cognitive stimulation, and its cognitive benefits are preserved. This finding complements task conflict theory: the distinction between task conflict and relationship conflict depends not only on the content of the conflict but also on the structural context in which it occurs.

4.4 Effective Separation of Verbal Quantity from Contribution Quality

Table 5. Correlation Analysis between IWF and Behavioral Indicators.

Behavioral Indicator	Correlation with IWF	p value	Interpretation
Speaking frequency	r = 0.23	0.035	Weak, significant
Times cited by peers	r = 0.76	< 0.001	Strong, highly significant

Note. Difference between the two correlations: Steiger's Z test, p < 0.001.

The contrast between S6 (the ineffective contributor) and S2 (the technical expert) in Group G-07 illustrates this point (Figure 2b). S6 spoke 32 times, but most turns consisted of interruptions and entanglement in details; S6 was cited only once and obtained an IWF of merely 0.79. S2 spoke 18 times but was cited 15 times, achieving an IWF of 1.20. Within Group G-07 the same dissociation holds (speaking turns vs. IWF: r = 0.43; times cited vs. IWF: r = 0.91). Correlation analysis across the full sample (Table 5) shows that speaking frequency correlated only weakly with IWF, whereas the number of citations correlated strongly, and the difference between the two correlations was significant. Peer evaluation, then, is not the "popularity score" that critics worry about: in an appropriate assessment scenario, peers can tell a "chatterbox" from an "expert." They enjoy a genuine observational advantage, since they sit through the entire discussion and directly perceive whose remarks advanced it and whose ideas entered the final solution; external observers can hardly capture this in full.

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

4.5 Engineering Ethics: From "Knowing" to "Doing"

During the discussions, 61.9% of the students cited GMP regulations at least once to defend the compliance baseline, significantly higher than the 28.6% observed in a comparison cohort of the same size (n = 84) from the previous course offering, which was taught with traditional case discussions ($\chi^2(1) = 18.84$, $p < 0.001$, $\varphi = 0.33$; Figure 2c). Debriefing interviews suggest that role assignment created a space of psychological permission: several students reported that when assigned the QA role they felt defending compliance was their responsibility, whereas in ordinary discussions they might have stayed silent to avoid conflict. The peer evaluation mechanism further reinforced the seriousness of role engagement; some students said they had looked up the regulatory provisions carefully because they knew teammates would evaluate their contribution. Role assignment and peer accountability worked together, and students moved from ethical cognition to ethical behavior, a central goal of engineering education for professional engineers.

5. Conclusion and Outlook

Taking the GMP course in pharmaceutical engineering as the entry point, this study constructed a closed-loop evidence chain of "scenario elicitation → behavioral coding → contribution weighting → individual assessment," responding to the reform requirement of the *Overall Plan* to strengthen process evaluation. The IWF algorithm produced substantial grade differentiation for social loafers (in the typical group, the loafer scored 26.2 points below the group mean); constructive conflict raised group risk-identification coverage by approximately 35 percentage points; peer evaluation took idea influence rather than speaking quantity as its core criterion; and AI-assisted coding reduced the workload by approximately 40%. The evidence chain is documented and auditable, and can directly support the evaluation of graduate attribute attainment in engineering education accreditation; the four scenario design principles are likewise transferable to engineering disciplines such as civil engineering, software engineering, and environmental engineering.

The study also has limitations: the single-institution sample limits external validity; classroom simulation differs from authentic workplace settings; and AI is currently used only for pre-coding. Future work will pursue large-scale cross-institutional studies, AI coding models dedicated to engineering education, and virtual reality for higher scenario fidelity.

Acknowledgments

This work was supported by the 2026 Guangdong Provincial Higher Education Teaching Reform Project (No. 20260105); the 2023 Guangdong Provincial Quality Engineering Pharmaceutical Analysis Course Teaching and Research Office Project; the 2026 General Project of Educational Evaluation System and Mechanism Reform of Zhaoqing University (No. JYPJ202601); the 2025 University-Level First-Class Undergraduate Course Construction Project of Zhaoqing University (No. ylkc202539); the 2025 "Artificial Intelligence +" Course Construction Project of Zhaoqing University (Pharmaceutical Technology); the 2025 Teaching Reform Project of Zhaoqing University "Exploration of AI-Empowered Teaching Reform in Pharmacology" (No. zlgc2025018); the 2021 Quality Engineering and Teaching Reform Project of Zhaoqing University (No. zlgc202113); and the Innovation and Entrepreneurship Training Program for College Students of Zhaoqing University (Nos. 202610580019X, 202610580020, and 202610580021), Zhaoqing University-Zhangjiang Institute for Food and Drug Control Joint Laboratory (No. 51), Zhaoqing University-Institute of Pharmaceutical Engineering.

References

[1] The Central Committee of the Communist Party of China, The State Council. Overall plan for deepening educational evaluation reform in the new era[Z]. Beijing, 2020. (in Chinese)

[2] China Engineering Education Accreditation Association. Standards for engineering education accreditation: 2024 edition[S]. Beijing: China Engineering Education Accreditation Association, 2024. (in Chinese)

[3] Zhu L, Tang H X, Hu D X, et al. An evaluation model of complex engineering problem-solving competence for engineering undergraduates[J]. Research in Higher Education of Engineering, 2023(4): 86-99. (in Chinese)

ISSN 2522-6398 Vol. 9, Issue 9: 39-46, DOI: 10.25236/FER.2026.090906

[4] ALQASSAB M, STRIJBOS J W, PANADERO E, et al. A systematic review of peer assessment design elements[J]. Educational Psychology Review, 2023, 35(1): 18.
[5] LATANÉ B, WILLIAMS K, HARKINS S. Many hands make light the work: The causes and consequences of social loafing[J]. Journal of Personality and Social Psychology, 1979, 37(6): 822-832.
[6] KARAU S J, WILLIAMS K D. Social loafing: A meta-analytic review and theoretical integration[J]. Journal of Personality and Social Psychology, 1993, 65(4): 681-706.
[7] ANDERSON L W, KRATHWOHL D R, AIRASIAN P W, et al. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives[M]. New York: Longman, 2001.
[8] Yuan Q J. Pharmaceutical good manufacturing practice engineering[M]. Beijing: Chemical Industry Press, 2024. (in Chinese)
[9] GUZMÁN S G. Monte Carlo evaluations of methods of grade distribution in group projects: simpler is better[J]. Assessment & Evaluation in Higher Education, 2018, 43(6): 893-907.
[10] HEIMANN A L, INGOLD P V, LIEVENS F, et al. Actions define a character: Assessment centers as behavior-focused personality measures[J]. Personnel Psychology, 2022, 75(3): 675-705.
[11] BREIL S M, LIEVENS F, FORTHMANN B, et al. Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness[J]. Personnel Psychology, 2023, 76(3): 759-795.
[12] BENNING T M. Reducing free-riding in group projects in line with students' preferences: Does it matter if there is more at stake?[J]. Active Learning in Higher Education, 2024, 25(2): 242-257.
[13] BALES R F. Interaction process analysis: A method for the study of small groups[M]. Cambridge, MA: Addison-Wesley, 1950.
[14] FRANCIS N, PRITCHARD C, PRYTHERCH Z, et al. Making teamwork work: Enhancing teamwork and assessment in higher education[J]. FEBS Open Bio, 2025, 15(1): 35-47.
[15] LANDIS J R, KOCH G G. The measurement of observer agreement for categorical data[J]. Biometrics, 1977, 33(1): 159-174.
[16] LAVE J, WENGER E. Situated learning: Legitimate peripheral participation[M]. Cambridge: Cambridge University Press, 1991.
[17] JEHN K A. A multimethod examination of the benefits and detriments of intragroup conflict[J]. Administrative Science Quarterly, 1995, 40(2): 256-282.