Analog and Digital Mentoring in Mathematics Education: Effects on Participation, Exam Success, and Cohort Variability

Miriam Weigel[1*], Reinhold Hübl[1], David Obermayr[1] and Marc Peterfi[1]

[1] Duale Hochschule Baden-Wuerttemberg Mannheim, ZeMath, Coblitzallee 1-9, 68163 Mannheim, Germany, miriam.weigel@dhbw.de (Corresponding Author)

Abstract. This study investigates the long-term effects of analog and digital mentoring formats on undergraduate mathematics exam performance. Based on a dataset of 489 students across six semesters, the results confirm that prior domain knowledge is the strongest predictor of exam success in first-year mathematics courses, while structured learning formats and mentoring exert a moderate, indirect influence. The study's novelty lies in its systematic analysis of how digital mentoring practices are sustained beyond the project phase, with particular attention to their support intensity. The findings underline the need for sustainable, adaptive support concepts in mathematics education.

Keywords: Mathematics Education, Formative Assessment, Digital Learning Environments, Student Engagement, Blended Instructional Design.

1 Introduction

The digital transformation of higher education has accelerated markedly over the past decade, driven by the proliferation of online tools and the growing demand for flexible, scalable instructional formats. The COVID-19 pandemic further intensified this shift, compelling institutions to rapidly adopt digital teaching methods and revealing both the potential and the limitations of e-learning within traditional academic structures. In mathematics education, where abstract reasoning and procedural fluency are central, digital environments offer unique opportunities for individualized practice, automated feedback, and scalable assessment formats [1, 2].

However, the transition to digital formats has also introduced new challenges. Voluntary participation in face-to-face tutorials has declined, and student engagement with optional support formats remains uneven [3]. Research suggests that academic success is more closely linked to active engagement than to mere attendance, underscoring the need for instructional models that foster sustained participation [4]. Blended learning - combining digital and analog elements - has emerged as a promising approach to address these challenges, particularly in STEM disciplines [5].

S. Moebs et al. (eds.), *Proceedings of the FIREtalk Conference - Research on FIRE! (research-on-fire 2025)*,
Advances in Social Science, Education and Humanities Research 1018,
https://doi.org/10.2991/978-94-6239-705-7_17

Against this backdrop, the EduFIT project at DHBW Mannheim, funded by the Stiftung Innovation in der Hochschullehre (September 2022 - February 2024), implemented a blended learning framework in first-year mathematics courses within the computer science program. The initiative aimed to standardize instruction and enhance student support through increased practice opportunities, individualized feedback mechanisms, and reduced tutorial group sizes [6, 7]. Key components included submission tasks that were corrected by teaching staff, STACK-based online quizzes with adaptive feedback, and differentiated tutorial formats.

While the initial implementation offered intensive support, this level of mentoring proved unsustainable beyond the project phase. As a result, tutorial structures were scaled back, and only core elements such as submission tasks and online quizzes were retained. The present study investigates the pedagogical impact of this transition, focusing on how changes in mentoring intensity and task engagement influenced student performance. Specifically, we examine the following hypotheses:

(H1) Reduced support activities lead to lower exam performance.
(H2) Participation in training sessions and submission tasks improves exam performance.

2 Evolving Tutorial Structures in Higher Mathematics Education

The EduFIT program was launched in the winter semester of 2022/23 and developed through a structured, multi-phase pilot aimed at enhancing mathematics instruction through digital support and individualized mentoring. Central to the concept was the integration of STACK-based formative assessments and the reduction of tutorial group sizes to half their usual capacity to enable more targeted support and increased student engagement. STACK (System for Teaching and Assessment using a Computer algebra Kernel), developed by Chris Sangwin [8], allows automated evaluation of mathematical responses based on mathematical properties and provides adaptive feedback tailored to common misconceptions. Its ability to randomize parameters enables scalable, reusable online exercises that support flexible and repeated practice.

Across two cohorts, each comprising six classes with 20 - 30 students, the instructional model was continuously refined in response to student feedback and performance data. Initially, weekly digital lecture-based assessments informed tutorial group allocation, while online quizzes and handwritten submission tasks offered structured opportunities for practice (see Fig. 1). Over time, the frequency and format of these components were adjusted to balance workload, autonomy, and instructional sustainability.

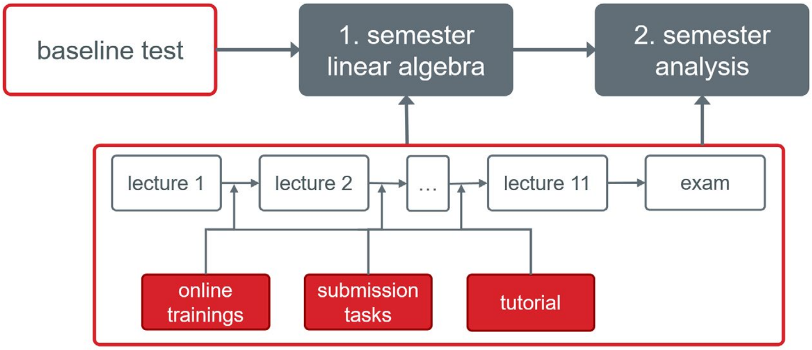

Fig. 1. An overview of the semester schedule.

Following the pilot phase, the EduFIT framework was integrated into regular teaching operations beginning in the summer semester of 2024. Upon the termination of project funding and the departure of key staff members, tutorial offerings were consolidated into a single group per class, reverting to the organizational structure that was in place prior to the project. Nevertheless, core elements such as online trainings and submission tasks were retained, albeit with reduced grading weight and increased flexibility. Participation in digital training became voluntary, and the complexity of proof-based assignments was moderated to reflect the shift toward greater student autonomy and the evolving use of AI tools in student workflows (see Fig. 2).

This streamlined version of the EduFIT model has remained in use since its institutional adoption and continues to inform mathematics instruction at DHBW Mannheim as of the summer semester 2025.

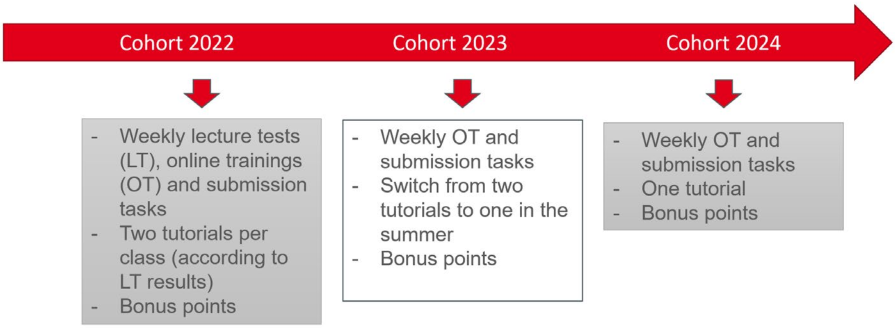

Fig. 2. An overview of the project phases.

3 Data Collection and Analytical Scope

During the project period and subsequent semesters until summer of 2025, comprehensive data was collected from three cohorts comprising a total of 489 computer science students. While exam results and results for submission tasks were available for all cohorts, participation in online trainings was only recorded for cohorts 2022 and 2023.

To ensure analytical consistency and comparability across participants, only results from the initial written examinations were included in the dataset. Students who completed the exam in a single semester were included solely in the analysis for that respective semester and excluded from analyses pertaining to the other semester. 18 students who did not participate in any form of examination were omitted to preserve the integrity of the comparative framework.

Supplementing this data, the analysis also incorporated results from a mathematics baseline test administered during the first week of the students' first semester. This instrument forms part of DHBW Mannheim's longstanding digital mathematics preparation program, which has been in use since 2011 to assess incoming students' prior knowledge in mathematics through a pre-post test design [9, 10, 11].

3.1 Standardization of Examination Data

The comparability of exam results is not always straightforward. On the one hand, the presence of actual differences in the participants' abilities gives rise to varying average scores. On the other hand, task difficulty may differ, particularly because both the subject matter and the tasks themselves may be perceived differently by students and instructors. As illustrated in Table 1, average exam scores varied considerably across semesters and cohorts, despite deliberate efforts to ensure consistency in competency expectations and assessment rigor.

To mitigate these discrepancies in the joint analysis, examination scores were standardized within each cohort using a z-score transformation prior to pooling. Accordingly, the interpretation of the results is based on standard deviations rather than absolute scores, with a z-score of 0.3 indicating that the corresponding result was 0.3 standard deviations above the exam mean.

Table 1. Descriptive statistics of exam results by cohort.

Cohort	Sem.	*N*	Mean	*SD*	Min	Max
2022	1.	157	59.67	18.49	12.86	98.57
	2.	121	60.90	19.80	10.00	95.71
2023	1.	162	57.26	19.90	8.33	100.00
	2.	142	44.77	19.95	2.50	98.33
2024	1.	148	43.68	18.50	3.33	85.83
	2.	124	44.81	19.27	4.17	88.33

3.2 Statistical Analyses

All statistical analyses were conducted using IBM SPSS Statistics (version 29.0). Group differences in prior knowledge and exam results were examined using analysis of variance (ANOVA) on the raw data. To assess the influence of assessment activities and support intensity on exam performance, multiple regression analyses were conducted. To remove exam- and topic-related biases, exam data were z-standardized (see section above) and subsequently pooled for joint analyses.

The regression analyses composed of two models. The baseline model included pre-test scores and the assessment variables (submission task scores and training participation, both expressed as percentages) as predictors. The extended model added support intensity as a predictor. Support intensity was dummy-coded (0 = low support, 1 = high support, i.e., semesters benefitting from the additional EduFIT project capacities). To test for potential moderating effects, interaction terms between support intensity and both assessment variables were computed. All predictors were mean-centered before constructing the interaction terms to reduce nonessential multicollinearity.

Due to missing data on training participation in the 2024 cohort, separate regression models were estimated to maximize the available sample size. Models M.1 and M.2 included the training variable and excluded the 2024 cohort, whereas Models M.3 and M.4 omitted the training variable but incorporated the full dataset.

Since participation numbers in the pre-test and the exams varied across cohorts and semesters, the sample sizes (N) also differed depending on the specific variables under investigation. The exact N for each model is reported in the corresponding results tables. Effect sizes are reported as standardized regression coefficients (β) for predictor variables, adjusted R^2 for overall model fit, and η^2 for ANOVAs. Statistical significance was evaluated at $\alpha = .05$.

All regression analyses were conducted in long format, treating each exam as an independent data point. This approach is appropriate since the influence of students' prior knowledge on first- and second-semester exam performance has been shown to be comparable [12], and it increases the effective sample size for the analyses.

4 Results

4.1 Assessing Prior Knowledge

Internal records of the annual mathematics pre-test indicate a gradual but continuous decline in prior knowledge over the past decade. To ensure comparability of prerequisites, an ANOVA was first conducted with pre-test scores as the dependent variable and cohort as the factor. Running the Shapiro-Wilk test ($\alpha = .05$) showed that exam results were normally distributed only for the cohort 2024, but not for cohorts 2022 and 2023. Given that previous simulation studies have demonstrated the robustness of ANOVA against violations of normality [13, 14] the analysis was carried out as planned. However, Levene's test indicated significant differences in variance across cohorts ($p = .013$), suggesting heterogeneity of variances. Taking this into account, the ANOVA indicated that there were no significant differences in prior knowledge between the cohorts (Welch's $F(2, 288.82) = 0.24$, $p = .79$). Hence it could be assumed that prior knowledge of students of the three cohorts was comparable.

4.2 Impact of Support on Exam Performance

Given that support intensity varied across semesters but not cohorts, an ANOVA was conducted with exam performance as the dependent variable and semester count as the factor. The Shapiro-Wilk test (α = .05) indicated that exam results were normally distributed for all exams, except for those of the second semesters of cohorts 2022 and 2023. Due to the same rationale as in the previous section the ANOVA was nevertheless conducted.

By Levene's test homogeneity of variances could be assessed (p = .872). The ANOVA revealed a significant difference in exam performance across semesters, $F(5, 848) = 24.91$, $p < .001$, $\eta^2 = .128$. Tukey post-hoc analysis identified two homogeneous groups. The first group, with higher mean exam scores, included both semesters of cohort 2022 and the first semester of cohort 2023, while the second group, with lower mean scores, comprised the second semester of cohort 2023 and both semesters of 2024. This grouping is also apparent in Fig. 3 and aligns with the changes in support intensity, with semesters receiving higher support corresponding to the group with higher mean exam scores, and semesters with lower support corresponding to the group with lower mean scores.

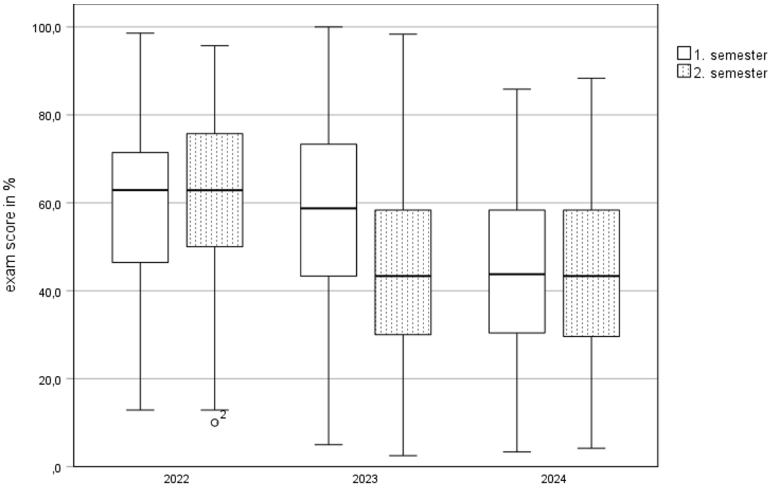

Fig. 3. Boxplot of exam results by cohort and semester

4.3 Regression Analyses

To examine the impact of support intensity and student engagement - specifically, participation in training sessions and performance on submission tasks - on academic performance, multiple linear regression analyses were conducted with standardized exam scores as the dependent variable. Consistent with prior research identifying subject-specific prior knowledge as a strong predictor of early academic success [10], all models included prior knowledge as a covariate.

Model 1 further incorporated the exercise-specific variables, which have previously been shown to affect exam outcomes [12], while Model 2 extended the analysis by adding support intensity as well as its interaction with the exercise-specific variables to test for a potential moderation effect.

Due to missing data on training participation for the 2024 cohort, separate regression models were estimated to maximize the available sample size. Models M.1 and M.2 included training participation as a predictor and therefore excluded the 2024 cohort. In contrast, Models M.3 and M.4 omitted the training variable but incorporated the full dataset, including the 2024 cohort. The corresponding regression results and coefficients are presented in Table 2 and Table 3, respectively.

Table 2. Regression results for M.1 and M.2

	M.1		M.2	
	B (SE)	*β*	*B* (SE)	*β*
Prior Knowledge	.026 (.002)	.527***	.026 (.002)	.528***
Submission	.011 (.002)	.210***	.013 (.002)	.251***
Training	.005 (.002)	.116*	.013 (.004)	.292***
Support Intensity			.155 (.372)	.067
Support × Submission			-.006 (.005)	-.193
Support × Training			-.010 (.004)	-.396*
Adj. R^2	.39		.40	
N	558		558	

Note. M.1 and M.2 exclude the 2024 cohort due to missing training participation data. Submission = performance on submission tasks; Training = participation in training sessions. Standardized exam scores served as the dependent variable.

B: regression coefficient; SE: standard error; *β*: standardized regression coefficient. *** $p < .001$; ** $p < .01$; * $p < .05$

Model M.1 yielded a significant overall result, $F(3, 554) = 120.46$, $p < .001$, explaining 39% of the variance in exam scores. All three predictors were statistically significant. Prior knowledge demonstrated a strong effect on exam performance ($\beta = .527$), whereas performance on submission tasks ($\beta = .210$) and participation in online training sessions ($\beta = .116$) showed effects of medium to small magnitude.

The introduction of support and interaction variables in Model M.2 resulted in only a minor improvement in model fit, reflected in an increase in adjusted R^2 of .01. The model nevertheless remained statistically significant, $F(6, 551) = 62.107$, $p < .001$. As expected in moderation models, the interaction terms and the support variable showed high multicollinearity (VIF > 20). Although all predictors were mean-centered prior to creating interactions to reduce nonessential multicollinearity, the coefficients of this model should be interpreted cautiously.

Table 3. Regression results for M.3 and M.4

	M.3		M.4	
	B (SE)	*β*	*B* (SE)	*β*
Prior Knowledge	.025 (.001)	.505***	.025 (.001)	.505***
Submission	.016 (.002)	.298***	.016 (.002)	.292***
Support Intensity			.039 (.263)	.019
Support × Submission			.001 (.003)	.055
Adj. R^2	.37		.37	
N	792		792	

Note. M.3 and M.4 include the 2024 cohort but omit training data. Submission = performance on submission tasks. Standardized exam scores served as the dependent variable. *B*: regression coefficient; SE: standard error; *β*: standardized regression coefficient
*** p <.,001; ** $p < .01$; * $p < .05$

The effects observed in the models including training data (M.1 and M.2) were largely mirrored by the models that incorporated the 2024 cohort but omitted the training variables (M.3 and M.4). Model M.3 was statistically significant, $F(2, 789) = 236.89$, $p < .001$, and explained 37% of the total variance in exam performance. This slight reduction in explained variance was expected, as training participation had shown a significant contribution in model M.1. Consistent with the previous models, prior knowledge emerged as the strongest predictor ($\beta = .505$), while submission tasks demonstrated a medium effect ($\beta = .298$). The addition of support and interaction variables in Model M.4 again introduced high multicollinearity (VIF > 20). However, this issue was less consequential here, as the corresponding coefficients were not statistically significant.

Taken together, these findings suggest that exam performance is primarily driven by prior knowledge and sustained engagement with submission tasks, whereas the influence of support intensity appears less consistent.

5 Discussion

This study investigated the influence of mentoring intensity and task engagement on students' exam performance across three pilot phases of the EduFIT project and three subsequent consolidation semesters conducted under standard university conditions. All examined cohorts entered their studies with comparable levels of prior mathematical knowledge, as confirmed by a well-established baseline test.

Our findings provide empirical support for hypothesis (H1): Reduced mentoring during the consolidation phase was associated with a measurable decline in examination performance. While confounding factors - such as variations in teaching staff, perceived exam difficulty, and scheduling - cannot be ruled out, it is noteworthy that performance patterns consistently reflected the level of mentoring provided across otherwise comparable cohorts. Although regression analyses in Tables 2 and 3 did not reveal a clear direct or moderating effect of support level on standardized examination scores, likely due to multicollinearity, we argue that mentoring intensity indirectly influences learning outcomes by shaping students' engagement with formative assessments, tutorials and lectures.

This relationship can be better understood through Self-Determination Theory [15], which posits that increased mentoring enhances perceived competence and relatedness, thereby supporting sustained engagement. Additionally, the findings align with Cognitive Load Theory [16]: Personalized support and smaller group settings reduce extraneous cognitive load and promote deeper learning. More direct contact with instructors increases accountability and resource awareness, fostering commitment and enabling more individualized feedback - factors known to enhance academic outcomes. Our results thus extend existing models by demonstrating that the long-term effectiveness of mentoring depends on its integration into adaptive and formative support structures.

Regarding hypothesis (H2), regression analyses further showed that participation in training sessions and performance on submission tasks were positively associated with examination outcomes. However, these effects were modest compared to the predictive strength of prior domain knowledge - a pattern consistent with previous studies [10, 17, 18].

Participation in optional support formats may be subject to selection effects and influenced by motivational and affective factors not directly captured in this study [19, 20]. It is plausible that students with higher intrinsic motivation were more likely to engage in training activities, while others were primarily incentivized by the bonus point system. This incentive, designed to encourage consistent participation, may have helped mitigate selection effects and broaden engagement.

Nevertheless, the opportunity to earn bonus points may have increased the temptation to copy solutions, either from peers or AI tools. Lecturers observed that similar or identical solutions frequently appeared in coursework submissions. In mathematics, a subject often outside students' primary focus, some may resort to copying to secure bonus points and pass exams more easily. In the 2023 cohort, tasks emphasizing argumentation and proof were introduced to limit AI-assisted copying; however, this increased cognitive demand may have inadvertently led to more superficial copying. These results highlight the need for assessment formats that balance authenticity, difficulty and resistance to automation - an increasingly important concern with the integration of generative AI in education.

To further examine these issues, the 2024 teaching evaluation included questions on AI use and task completion. With a response rate of almost 40%, results showed that while approximately 18% of respondents admitted to sometimes or frequently copying from their peers, only 9% reported directly copying from AI tools. However, over 80% reported using AI at least sometimes for support or clarification. This suggests that most students engage constructively with AI, though misuse remains prevalent.

Despite these limitations, our study underscores the value of structured analog and digital support for mathematical understanding. Importantly, we demonstrate that digital mentoring practices can be sustained and adapted beyond the project phase - a novel contribution compared to earlier work, which rarely examines long-term sustainability. Future research should focus on identifying scalable mentoring strategies, refining task design to maintain engagement within resource constraints, and incorporating validated measures of motivation and affect.

Practical Implications: For educators, our findings indicate that mentoring - especially when combining analog and digital components - should be institutionalized as an ongoing feature of mathematics education, not treated as a temporary project. Instructors are advised to establish regular online feedback, foster peer interactions in digital spaces, and tailor support intensity to individual student needs. Careful task design is essential to ensure authenticity and limit the potential for unreflective copying, especially given the increasing prevalence of AI tools.

In summary, this study contributes to the understanding of how sustainable, adaptive mentoring can be integrated into mathematics education and points to the need for continued research on effective, scalable support mechanisms.

Acknowledgments. The EduFIT project was funded by Stiftung Innovation in der Hochschullehre (Foundation for Innovation in Higher Education) (grant number FRFMM-718/2022).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brame, C. J. & Biel, R.: Test-Enhanced Learning: The Potential for Testing to Promote Greater Learning in Undergraduate Science Courses. CBE—Life Sciences Education, **14**(2), es4 (2015)
2. Förster, M., Weiser, C. & Maur, A.: How feedback provided by voluntary electronic quizzes affects learning outcomes of university students in large classes. Computers & Education, **121**, 100-114 (2018)
3. Büchele, S.: Evaluating the link between attendance and performance in higher education: the role of classroom engagement dimensions. Assessment & Evaluation in Higher Education, **46**(1), 132-150 (2021)
4. Massingham, P. & Herrington, T.: Does attendance matter? An examination of student attitudes, participation, performance and attendance. Journal of university teaching & learning practice, **3**(2), 20-42 (2006)
5. Yu, Z., Xu, W. & Sukjairungwattana, P.: Meta-analyses of differences in blended and traditional learning outcomes and students' attitudes. Frontiers in psychology, **13**:926947 (2022)
6. Hübl, R., Kneher, K., Obermayr, D. & Peterfi, M.: Einsatz von Online-Tests in der Mathematiklehre am Beispiel des Projekts EduFIT. In: Hufnagel, J., Ternes, D. & Schnekenburger, C. (eds.), Digitale Lehre an der DHBW - The New Normal? #DUAL: ZHL-Schriftenreihe für die DHBW. vol. 7, pp. 77-89. DHBW CAS ZHL (2024)

7. Kneher, K. G., Hübl, R., & Peterfi, M.: Using STACK to individualise tutorials in mathematics. In: International Meeting of the STACK Community 2023, Tallinn, Estonia. (2023)
8. Sangwin, C.: Computer Aided Assessment of Mathematics, Oxford University Press. ISBN: 9780199660353 (2013)
9. Derr, K.: Identifying Consistent Variables in a Heterogeneous Data Set: Evaluation of a Web Based Pre Course in Mathematics. Electronic Journal of e-Learning, **15**(1), 82-93 (2017)
10. Derr, K., Hübl, R. & Ahmed, M. Z.: Prior knowledge in mathematics and study success in engineering: informational value of learner data collected from a web-based pre-course. European Journal of Engineering Education, **43**(6), 911-926 (2018)
11. Derr, K., Hübl, R., Huckenbeck, U., Mechelke-Schwede, E., Obermayr, D. & Weigel, M.: Studien-Mathematikvorbereitung. In: Küstermann, R., Kunkel, M., Mersch, A. & Schreiber, A. (eds.), Selbststudium im digitalen Wandel, pp. 21-42. Springer Spektrum (2021)
12. Hübl, R., Obermayr, D., Peterfi, M. & Weigel, M.: EduFIT: Verstetigung gelungen? Eine quantitative Analys. In: Merker, J., Hübl, R., Liebscher, E. & Eulitz, M. (eds.), Tagungsband 3. Workshop Digitale Lehre im Rahmen der Grundlagenausbildung in MINT-Fächern an Hochschulen, pp. 71-82. Hochschule für Technik, Wirtschaft und Kultur Leipzig (2025)
13. Blanca Mena, M. J., Alarcón Postigo, R., Arnau Gras, J., Bono Cabré, R., & Bendayan, R.: Non-normal data: Is ANOVA still a valid option?. Psicothema, **29**(4), 552-557 (2017)
14. Schmider, E., Ziegler, M., Danay, E., Beyer, L. & Bühner, M. (2010). Is it really robust? Methodology, **6**(4), 147–151.
15. Deci, E. L., & Ryan, R. M.: The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, **11**(4), 227-268 (2000)
16. Sweller, J.: Cognitive load theory. In: Mestre, J. P. & Ross, B. H. (eds.), The psychology of learning and motivation, vol. 55, pp. 37-76. Academic Press (2011)
17. Hall, C. W., Kauffmann, P. J., Wuensch, K. L., Swart, W. E., DeUrquidi, K. A., Griffin, O. H. & Duncan, C. S.: Aptitude and personality traits in retention of engineering students. Journal of Engineering Education, **104**(2), 167-188 (2015)
18. Kokkelenberg, E. C. & Sinha, E.: Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students. Economics of Education Review, **29**(6), 935-946 (2010)
19. Büchele, S. & Feiste, L.: Exploring determinants of engagement with voluntary mathematics assignments in economics education: a correlational study. Teaching Mathematics and its Applications: An International Journal of the IMA, **43**(4), 339-355 (2024)
20. Kirsten, K. & Greefrath, G.: On-campus vs distance tutorials in preparatory courses for mathematics student teachers–performance gains and influencing factors. International Journal of Research in Undergraduate Mathematics Education, **10**(2), 547-576 (2024)