Examining Effects of Gamification Elements in an Intelligent Tutoring System for 7th Grade English Learners on Their Motivation – A Randomized Controlled Field Trial (Pre-registration)

Parrisius, C.[1,2], Pieronczyk, I.[2], Wendebourg, K.[2], Holz, H.[3,4], Rieger, S.[2], Schmidt, T.[5], Meurers, D.[4], Trautwein, U.[2], & Nagengast, B.[2,6]

[1] Karlsruhe University of Education, Germany

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany

[3] Ludwigsburg University of Education, Germany

[4] Department of Computational Linguistics, University of Tübingen, Germany

[5] Institute of English Studies, Leuphana University Lüneburg, Germany

[6] Department of Education and the Brain & Motivation Research Institute (*b*MRI), Korea University, South Korea

Date of pre-registration: May 5th, 2023

Introduction

Digital learning environments are growing in popularity in supporting learning or practice phases in institutionalized settings. Multiple studies suggest a link between the use of digital environments and achievement motivation, often explained by individualized feedback as achieved in intelligent tutoring systems (ITS; e.g., Wilson & Czik, 2016). Gamification elements added to digital environments are considered beneficial to maintaining the learners' motivation (Jackson & McNamara, 2013; Sailer & Homner, 2020).

The gamification of digital environments received increasing interest also in second language learning research (Dehghanzadeh et al., 2021). In particular, gamified elements in vocabulary tools were examined and found to make learning more enjoyable than non-gamified tools (Liu et al., 2011; Wu & Huang, 2017). Even though Sailer et al. (2017) showed that gamification is not efficient per se, but different game elements are linked to specific motivational outcomes, the impact of *specific* gamified elements (rather than gamification in general) on students' motivation has rarely been examined in the second language learning context (Dehghanzadeh et al., 2021). To fill this gap, in the current study, we will examine different theory-driven gamification elements based on underlying mechanisms of achievement motivation implemented in the ITS FeedBook (Meurers et al., 2019; Parrisius, Wendebourg, et al., 2023), a web-based English workbook targeting seventh-grade learners of English as a second language in German academic-track schools. In an institutionalized learning setting, we will investigate the motivational impact of a learner dashboard, a pedagogical agent, and individualized feedback with a large randomized controlled field trial.

Research Questions and Hypotheses

We aim to answer the following research questions:

- RQ1: Do students report higher levels of motivation (i.e., expectancies, intrinsic value, utility value, attainment value, and cost; cf. Eccles & Wigfield, 2020) when provided with (1) a learner dashboard, (2) a learner dashboard and a pedagogical agent, or (3) none of these elements? We assume that providing learners with such gamification elements will positively affect their motivation for English. In particular, the learner dashboard highlighting the individual learner's progress might impact their expectancies, intrinsic value, utility value, and subjective cost of engaging in English learning (Sailer et al., 2017). Further, we assume that students who have not only access to the learner dashboard but also to a pedagogical agent will benefit more concerning their intrinsic value and subjective cost.

- RQ2: Does individualized feedback contribute to students' motivation for English? Feedback is a relevant driver of learning gain (Hattie & Timperley, 2007). Receiving individualized feedback in the FeedBook was found to have positive effects on students' learning gain when compared with no feedback (Meurers et al., 2019), but no evidence for effects were found when compared to true/false feedback (Parrisius, Wendebourg, et al., 2023). Irrespective of downstream impacts such as effects on students' learning gains, individualized feedback might initially lead to higher general levels of motivation in students because they might experience higher levels of expectancies and intrinsic value (because they are immediately aware of their performance and are given the means to improve). At the same time, learners might either not see the relevance of the feedback and, as a consequence, might not use it, or they might get even frustrated or annoyed by the intelligent feedback in cases where they struggle understanding it, or in cases where the hints generated by the intelligent algorithms do not fit ideally, which could be revealed in higher levels of subjective cost. Thus, we do not formulate specific expectations concerning this research question.

Additionally, we seek to answer the following exploratory research questions:

- RQ3 (interaction effect): Is the effect of access to a learner dashboard or a learner dashboard *and* a pedagogical agent larger if it is accompanied by individualized feedback? That is, is there a multiplicative effect of individualized feedback and either of the gamification elements?

- RQ4 (individual differences): To what extent do students differ in their effects of the learner dashboard and the pedagogical agent on their motivation for English as a function of their baseline motivation for English?

Method[1]

Data were collected in academic track schools in three German federal states (namely, Baden-Württemberg, North Rhine-Westphalia, and Hamburg). Approval for this study was received from the appropriate state agencies of the three German federal states where the study was conducted as well as from the institutional review board of the lead university.

Context of the Study

The FeedBook System

The FeedBook system is an ITS for English as a second language (Rudzewitz et al., 2018). It is a web-based workbook containing exercises appropriate to the seventh-grade curriculum and is designed to be used as part of the regular teaching and learning routine (e.g., for individual learning phases or homework). The FeedBook version in this study contains 301 exercises with different question formats. Exercises in the FeedBook are aligned with, and prepare for, four complex target tasks that require the integration of several skills and competences, such as the use of certain lexical material and grammatical structures. The target tasks are carried out in class.

The pedagogical sequence leading up to a target task is called a task cycle and was planned to last approximately 3 weeks. The FeedBook contains digital practice material for each of the four task cycles and provides teachers with detailed lesson plans and teaching

[1] This preregistration contains sections adapted from Parrisius, Wendebourg, et al. (2023), sometimes with only minor changes. The authors have agreed on this. Both studies describe the same randomized controlled field trial and sample.

material (e.g., handouts and other additional digital and paper-based activities). Teachers were asked to use the teaching material while working on the task cycles and to use the FeedBook for individual practice of the targeted grammatical constructs.

Participants

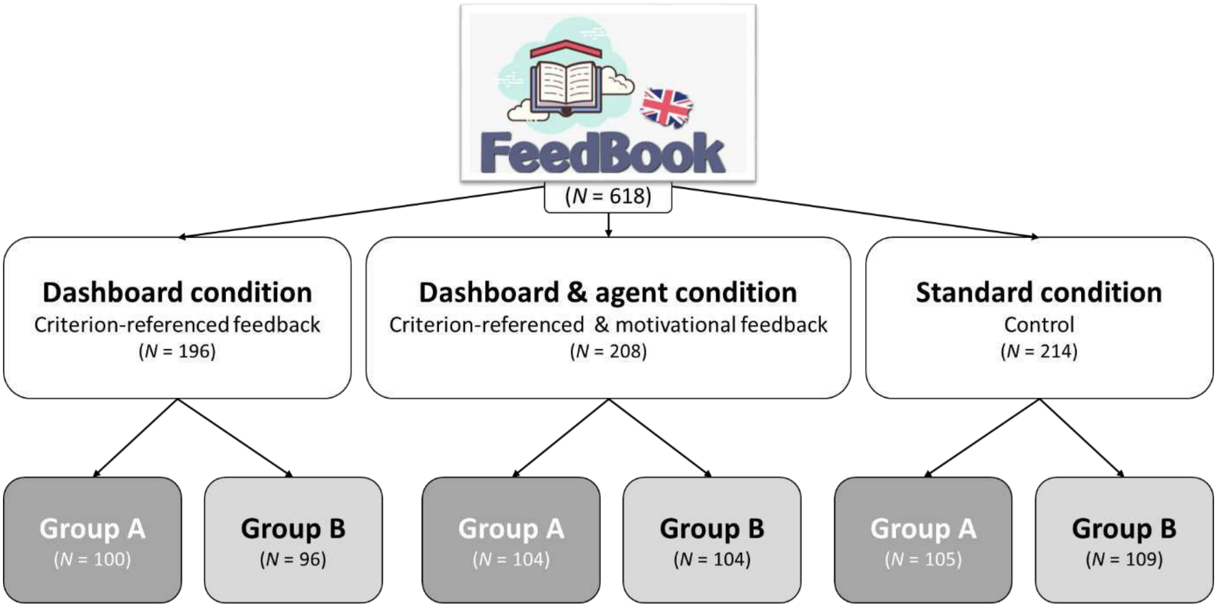
Data for this investigation was collected throughout the entire school year 2021/2022 in German academic track schools ("Gymnasium"). For the recruitment process and for information on the power analyses based on which we determined our aspired sample size, we refer the reader to the pre-registration of the full study design (Parrisius et al., 2022) or Parrisius, Wendebourg, et al. (2023). A total of 13 schools with 36 classes participated in the study. Out of the 13 participating schools, seven schools were randomly assigned to use the FeedBook (*FeedBook condition*), whereas six schools were assigned to a *waiting control condition* in which "business as usual" in English class took place. In this study, we will exclusively consider the subsample of schools in the FeedBook condition, consisting of N_{S} = 7 schools with N_{T} = 21 teachers and their N_{C} = 24 classes, comprising a total of N_{St} = 618 students with written consent (corresponding to a participation rate in the FeedBook condition of 96.7%). All students were seventh-grade learners of English as a second language. In our subsample of schools, 53% of the students were female adolescents and students' mean age was 12.51 (*SD* = 0.41).

Procedure and Study Design

Data collection took place before and after every task cycle. To address the research questions, students had access to different versions of the FeedBook based on the conditions to which they were randomly assigned to. Entire classes were assigned to one of three FeedBook versions (see Figure 1): (1) the FeedBook with a learner dashboard and progress bars (*dashboard condition*); (2) the FeedBook with a learner dashboard and progress bars, and additionally with a pedagogical agent stating motivational feedback messages (*dashboard & agent condition*), and (3) the standard FeedBook in which no such features occurred (*standard* *condition*). Finally, individual students within classes were randomly assigned to *Group A* or *Group B*. Students in *Group A* received individualized feedback throughout their FeedBook use for all grammatical structures targeted in task cycles 1 and 3, whereas students in *Group B* received individualized feedback for all grammatical structures targeted in task cycles 2 and 4. That is, for each of the grammatical structures targeted in the FeedBook, only one-half of the students received individualized feedback. For the other students, feedback for these grammatical structures was not provided throughout the entire FeedBook use, and they received true/false feedback instead (i.e., a default feedback message reading “This is not what I am expecting” or, where applicable, feedback that targeted potential misconceptions other than those related to the grammatical structures of interest).

Figure 1

Intervention Conditions for the Subgroup of Schools Focused on in this Investigation



Note. Students in Group A received individualized feedback for the grammatical structures targeted task cycles 1 and 3, and true/false feedback for the grammatical structures targeted in task cycles 2 and 4. Students in Group B received individualized feedback for the grammatical structures targeted task cycles 2 and 4, and true/false feedback for the grammatical structures targeted in task cycles 1 and 3.

Instruments

Survey data were collected at the beginning of the school year (i.e., before the FeedBook was introduced; T1) and after every task cycle (i.e., after task cycle 1: T2; after task cycle 2: T3, after task cycle 3: T4; after task cycle 4: T5). Students' self-reported motivation for English (expectancies and values) forms the primary outcomes of this investigation (for a full variable list, see Parrisius et al., 2022).

Students' Motivation for English

We asked students to report their motivation for English at each of the time points after the introductory sentence "To what degree do these statements apply to you?" In line with SEVT (Eccles & Wigfield, 2002, 2020), we asked students about their subjective expectancies and task values. Their competence-related beliefs concerning English were measured with a scale adapted from Baumert et al. (1997), Ramm et al. (2006), and Gaspard et al. (2017; e.g., "I am good at English"), as well as from McAuley et al. (1987; e.g., "I am satisfied with my English performance"). Measures for students' English-related intrinsic value (e.g., "English is fun for me"), attainment value (e.g., "English as a subject is important to me"), utility value (e.g., "Knowledge in English comes in handy during everyday life and leisure time") and cost (e.g., "Dealing with English drains a lot of my energy") were adapted from Gaspard et al. (2017). We shortened the utility-value scale because—other than in the original publication—students' utility value was not the main focus of the present investigation. Additionally, we expanded the cost scale, thus fitting the individual learning context by emphasizing emotional cost (items adapted from Pekrun et al., 2002; e.g., "Dealing with English usually frustrates me"). All items were answered on a 4-point response format from 1 = *not true at all* to 4 = *completely true*.

Covariates

We will consider a set of 10 variables as covariates (if not already in the focus of the respective analysis): students' gender, self-concept, intrinsic value, utility value, cost, ideal L2 self, English effort, homework effort, conscientiousness, and computer proficiency. For the selection process, we refer the reader to Parrisius, Wendebourg, et al. (2023). Additionally, we will consider adding students' English proficiency scores at the respective time points as covariates.

Statistical Analysis Plan

The significance level for all analyses will be set at 5% (two-tailed). To answer our research questions, we will specify one multilevel two-way ANCOVA per task cycle and per outcome (expectancies, intrinsic value, attainment value, utility value, and cost) in Mplus considering our 3×2 factorial design. We aim to investigate mean differences in students' motivation for English after task cycles 1 to 3. We will not consider task cycle 4 because there was a continuous dropout throughout the school year, resulting in too few classes after task cycle 4 (classes dropped out by T2: none; T3: three; T4: five; T5: 13 in total). Notably, because of schedule complications due to residual COVID-19 effects, 13 of the 24 classes did not manage to work on task cycle 4 and, consequently, did not participate in T5.

For this purpose, we will implement separate two-level models with the students at Level 1 (individual level) and the classes at Level 2 (class level) for each of the three task cycles, resulting in 15 models in total. The clustering of classrooms within schools will be addressed by including dummy variables for those schools that participated with more than one class and with different conditions at the class level in the regression model.

To estimate the effects of receiving (1) the standard FeedBook (standard condition) versus the FeedBook with dashboard (dashboard condition) versus the FeedBook with dashboard and pedagogical agent (dashboard & agent condition; RQ1), and of receiving (2) individualized versus true/false feedback, and of (3) all combinations of these (RQ3), the following independent variables will be included in the models: A dummy variable indicating the group of students receiving individualized feedback for the grammatical structures focused on in the respective task cycle will be used as predictor at the individual level; two dummy variables indicating the dashboard condition and the dashboard & agent condition as compared with the standard condition, respectively, will be used as predictors at the class level; two interaction terms between the individual-level dummy variable and the class-level dummy variables will be included at the individual level. Additionally, we will include the respective pretest score of the outcome variable of interest as a covariate in the analyses. Finally, we will additionally include the variables for which we found statistically significant differences between the intervention groups before introducing the FeedBook (i.e., at T1) to yield unbiased estimates of the intervention effects (if not already in the focus of the analyses). All continuous variables will be standardized before running the analyses, so that the regression coefficients of the dummy variables indicating the respective treatment groups can directly be interpreted as Cohen's d (see Marsh et al., 2009; Tymms, 2004).

To answer RQ1 and RQ2 (intervention effects), we will report the *main* effects of the individualized feedback (i.e., Group A vs Group B), of the dashboard availability (i.e., a blend of the dashboard condition and the dashboard & agent condition vs the standard condition), and of the additional agent availability (i.e., dashboard & agent condition vs dashboard condition). These effects will be calculated via the MODEL CONSTRAINT option available in Mplus. To answer RQ3 (interaction effect), we will report the respective interaction effects. To answer RQ4 (individual differences), we will investigate whether students' baseline motivation for English moderates the effects of the learner dashboard and the pedagogical agent. To do so, we will add an interaction term between the respective motivation variable at baseline and the dummy variables indicating the dashboard condition and the dashboard & agent condition, respectively, at the class level as an additional predictor in our multilevel regression analyses.

Missing Values

As is common in longitudinal studies (Enders, 2010), we had missing data at all measurement occasions because of absence of individual students (see Table 1) and because of nonresponses to single items. All analyses will be conducted using full information maximum likelihood estimation implemented in Mplus (Graham, 2009). All covariates will be used as auxiliary variables by including correlations between these variables and the predictor variables as well as the residuals of the outcome variables at both levels (see Collins et al., 2001; Enders, 2010).

Table 1

Overview of Dropout and Individual Attendance in % per Time Point

	T1	T2	T3	T4	T5
Cumulated drop out of classes	0	0	12.5	20.8	54.2
Absence of individual students within participating classes	3.6	7.8	6.8	11.0	19.4

Note. T = time point. The high missing rate at T5 was a consequence of COVID-19 side effects. We did not consider T5 in our analyses.

Knowledge of Data

Work in Progress and Previous Publications

Multiple authors of this pre-registration have previously used this data for different research questions. Even though the data have been used before, including at least some of the variables and measures in this study, all analyses focused on other (outcome) variables (e.g., students' English proficiency). Previous work, including the measures used, is listed below, and in each case, it is indicated who among the authors was involved.

- Blume et al. (2022): In this conference contribution presented at EUROCALL 2022, Blume, Meurers, Middelanis, Pili-Moss, and Schmidt compared students' learning gains across the three FeedBook conditions *standard FeedBook*, *dashboard* and *dashboard & agent* for task cycle 1.

- Colling et al. (2023): In this conceptual work, Colling, Pieronczyk, Parrisius, Holz, Bodnar, Nuxoll, and Meurers describe the development process behind the learner dashboard. No data have been worked with for this purpose.
- Deininger, Lavelle-Hill, et al. (2023) and Deininger, Parrisius, et al. (2023): These two projects focus on the prediction of English proficiency (on an exercise level, Deininger, Lavelle-Hill, et al., 2023; or after a full task cycle, Deininger, Parrisius, et al., 2023) by behavioral trace data. Students' interactions with the system were used as basis to calculate the input variables (e.g., number of exercises worked on, time on task). Results regarding academic performance prediction on an exercise level have been submitted; analyses for the prediction of academic performance after a full task cycle are currently performed and prepared for publication. Among the co-authors of the current pre-registration, Parrisius, Pieronczyk, Colling, Trautwein, Meurers, Kasneci, and Nagengast were involved as co-authors.
- Deininger, Pieronczyk, et al. (2023): In this study, Deininger et al. aim at investigating students' effort and engagement in using the FeedBook during task cycle 1 while considering the full set of available survey data from T1 as potential predictors. For this purpose, two outcomes are of interest: students' self-reported engagement and students' engagement derived from behavioral indicators while using the FeedBook during task cycle 1 (e.g., number of exercises worked on, time on task). For preliminary analyses, students' gender, prior achievement at the end of grade level 6, their self-reported conscientiousness, effort regarding English and regarding homework, as well as their self-reported English self-concept, intrinsic value, attainment value, utility value, and cost value were used. Further analyses are currently in preparation, and we plan to finalize a manuscript based on the full set of

results. Overlapping co-authors include Pieronczyk, Parrisius, Wendebourg, Trautwein, Meurers, and Nagengast at this point in time.

- Parrisius, Wendebourg, et al. (2023): In this project, Parrisius, Wendebourg, et al. investigated intervention effects on students' English proficiency after task cycles 1 to 3. They compared the intervention conditions at the class level (i.e., standard FeedBook, learner dashboard, learner dashboard & pedagogical agent) as well as on the individual level (i.e., Group A and Group B). Students' motivation (i.e., expectancies, intrinsic value, attainment value, utility value, and cost) at T1 were used as covariates in this investigation. The analyses are pre-registered at *PsychArchives* (https://doi.org/10.23668/psycharchives.8152) and have been submitted. Our results show positive effects of the learner dashboard on students' English proficiency scores, mixed effects when additionally introducing the pedagogical agent, and no evidence for effects of the individualized feedback on the test scores. The set of co-authors fully overlaps with the co-authors of the current investigation. The planned analyses for the current investigation were chosen to be in parallel with the Parrisius, Wendebourg, et al. analyses and merely differ in the outcomes of interest (motivation instead of achievement).
- Parrisius et al. (2022): This publication is the pre-registration of the full study design. All listed co-authors have also been involved in this pre-registration. No data have been worked with for this purpose.
- Pieronczyk et al. (2022): In this conference contribution presented at the annual conference of the German Society for Empirical Educational Research (Gesellschaft für Empirische Bildungsforschung; GEBF), Pieronczyk et al. investigated the number of exercises students and classes (on average) worked on during task cycle 1. For this purpose, they investigated the behavioral trace

data from using the FeedBook. The analyses were purely descriptive. Parrisius, Wendebourg, Bodnar, Colling, Holz, Trautwein, Meurers, and Nagengast have been involved as co-authors.

- Pili-Moss et al. (2022): In this conference contribution presented at EUROCALL 2022, Pili-Moss, Schmidt, Blume, Middelanis, and Meurers investigated the efficacy of the FeedBook in a sub-sample of 77 students (three intact classes) who used the platform within the *dashboard & agent condition*. Specifically, Pili-Moss et al. investigated: (1) pre-posttest gains comparing constructions for which both digital and classroom instruction were provided to gains relative to linguistic targets for which only digital instruction was given, and (2) the relationship between the efficacy of hybrid instruction (digital + classroom-based) and the learners' ability to accurately employ practiced linguistic targets in classroom-based communicative target tasks. For this purpose, they used pre-posttest English proficiency data from task cycles 2 and 3, as well as data from the written target task performed by the students at the end of task cycle 3. However, their knowledge of data is restricted to this subgroup of students who only had access to one version of the FeedBook (i.e., no comparisons possible).

Prior Knowledge About the Dataset

The survey data assessed at T1 to T5 have been fully cleaned. This work has primarily been executed by Ines Pieronczyk and supervised by Cora Parrisius. Due to the analyses in Parrisius, Wendebourg, et al. (2023) and the preliminary analyses in Deininger, Pieronczyk, et al. (2023), we have knowledge about students' motivation at T1: means, standard deviations, as well as Cronbach's alpha were calculated for students' English self-concept, intrinsic value, attainment value, utility value, and cost. Additionally, we have already estimated correlations among the variables of interest at T1 and have checked for baseline differences in students' motivation between the different intervention groups. The results are presented in Tables 2, 3, 4a, and 4b.

Table 2
Descriptive Statistics at T1

	ICC	*N*	Missing values (%)	*M*	*SD*	Min	Max	Reliability
Self-concept	0.07	559	9.5	2.85	0.65	1.00	4.00	0.87
Intrinsic value	0.10	559	9.5	2.92	0.79	1.00	4.00	0.90
Attainment value	0.05	547	11.5	3.14	0.57	1.00	4.00	0.78
Utility value	0.04	550	11.0	3.20	0.55	1.00	4.00	0.68
Cost	0.04	538	12.9	2.17	0.60	1.00	4.00	0.81

Note. ICC = intraclass correlation, relative = missing values in % with respect to students in classes that still participated at that time point, TC = task cycle, T = time point.
In Germany, grades range from 1 to 6 with lower values indicating better achievement. To facilitate interpretation, however, we recoded students' grades so that higher values indicated better achievement.

Table 3
Intercorrelations at T1

	(1)	(2)	(3)	(4)	(5)
(1) Self-concept	-	0.76	0.58	0.52	-0.90
(2) Intrinsic value	0.59	-	0.78	0.65	-0.69
(3) Attainment value	0.34	0.54	-	0.63	-0.51
(4) Utility value	0.33	0.45	0.58	-	-0.50
(5) Cost	-0.73	-0.59	-0.23	-0.27	-

Note. TC = task cycle, T = time point. The class level is presented above the diagonal, the individual level below the diagonal.

Table 4a
Baseline Differences

	Self-concept			Intrinsic value			Attainment value		
	Est	95% CI	*p*	Est	95% CI	*p*	Est	95% CI	*p*
Group A – Group B	-0.07	[-0.24, 0.09]	.368	-0.12	[-0.27, 0.04]	.368	-0.05	[-0.21, 0.10]	.512
Cond 1 – Cond 3	**-0.27**	**[-0.48, -0.06]**	**.013**	**-0.29**	**[-0.57, -0.01]**	**.013**	-0.06	[-0.26, 0.14]	.577
Cond 2 – Cond 3	-0.14	[-0.40, 0.11]	.278	-0.12	[-0.41, 0.18]	.278	-0.06	[-0.27, 0.15]	.576
Cond 1 – Cond 2	-0.13	[-0.37, 0.11]	.297	-0.17	[-0.50, 0.15]	.297	0.00	[-0.25, 0.26]	.990

Note. 95% CI = 95% confidence interval, Cond 1 = dashboard condition, Cond 2 = dashboard & agent condition, Cond 3 = standard condition. Bold values are statistically significantly different from 0 at $p < .05$ (two-tailed).

Table 4b
Baseline Differences

	Utility value			Cost		
	Est	95% CI	*p*	Est	95% CI	*p*
Group A – Group B	-0.10	[-0.22, 0.02]	.092	0.05	[-0.13, 0.23]	.583
Cond 1 – Cond 3	**-0.19**	**[-0.35, -0.03]**	**.020**	**0.19**	**[0.05, 0.33]**	**.007**
Cond 2 – Cond 3	0.02	[-0.15, 0.19]	.839	-0.03	[-0.25, 0.20]	.827
Cond 1 – Cond 2	**-0.21**	**[-0.39, -0.03]**	**.022**	**0.22**	**[0.03, 0.41]**	**.024**

Note. 95% CI = 95% confidence interval, Cond 1 = dashboard condition, Cond 2 = dashboard & agent condition, Cond 3 = standard condition. Bold values are statistically significantly different from 0 at $p < .05$ (two-tailed).

References

Baumert, J., Gruehn, S., Heyn, S., Köller, O., & Schnabel, K. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU). Dokumentation–Band 1: Skalen Längsschnitt Welle 1–4 [Learning processes, educational careers, and psychosocial development in adolescence and young adulthood (BIJU). Documentation–Volum.* Max-Planck-Institute for Human Development.

Blume, C., Meurers., D., Middelanis, L., Pili-Moss, D., & Schmidt, T. (2022). *Strengthening form-focused practice in task-based language teaching through intelligent CALL.* EUROCALL Conference 2022, virtual.

Colling, L., Pieronczyk, I., Parrisius, C., Holz, H., Bodnar, S., Nuxoll, F., & Meurers, D. (2023). *Towards task-orinted ICALL: A criterion-referenced learner dashboard organising digital practice* [Unpublished manuscript].

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2021). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning*, *34*(7), 934–957. https://doi.org/10.1080/09588221.2019.1648298

Deininger, H., Lavelle-Hill, R., Parrisius, C., Pieronczyk, I., Colling, L., Meurers, D., Trautwein, U., Nagengast, B., & Kasneci, G. (2023). *Can you solve this on the first try? - Understanding exercise field performance in an intelligent tutoring system* [Manuscript accepted for publication].

Deininger, H., Parrisius, C., Colling, L., Meurers, D., Trautwein, U., Nagengast, B., & Kasneci, G. (2023). *Analyzing behavioral trace data with machine learning and explainable AI to predict learning success* [Manuscript in preparation].

Deininger, H., Pieronczyk, I., Parrisius, C., Plumley, R., Colling, L., Meurers, D., Kasneci, G., Trautwein, U., Nagengast, B., Greene, J., & Bernacki, M. (2023). *Self-perception = self-deception? - The relationship between homework effort self-reports and* *observable homework behavior and their impact on achievement* [Manuscript in preparation].

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A development, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, Article 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, *48*, 67–84. https://doi.org/10.1016/J.CEDPSYCH.2016.09.003

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*, *105*(4), 1036–1049. https://doi.org/10.1037/a0032580

Liu, M., Horton, L., Olmanson, J., & Toprac, P. (2011). A study of learning and motivation in a new media enriched environment for middle school science. *Educational Technology Research and Development*, *59*(2), 249–265. https://doi.org/10.1007/s11423-011-9192-7

McAuley, E., Duncan, T., & Tammen, V. V. (1987). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*, 48–58.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, *36*, 161–188. https://doi.org/10.1017/S0267190519000126

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.*

Parrisius, C., Pieronczyk, I., Blume, C., Wendebourg, K., Pili-Moss, D., Assmann, M., Beilharz, S., Bodnar, S., Colling, L., Holz, H., Middelanis, L., Nuxoll, F., Schmidt-Peterson, J., Meurers, D., Nagengast, B., Schmidt, T., & Trautwein, U. (2022). *Using an intelligent tutoring system within a task-based learning approach in English as a foreign language classes to foster motivation and learning outcome (Interact4School): Pre-registration of the study design*. PsychArchives. https://doi.org/10.23668/psycharchives.5366

Parrisius, C., Wendebourg, K., Rieger, S., Blume, C., Pili-Moss, D., Colling, L., Pieronczyk, I., Holz, H., Bodnar, S., Loll, I., Schmidt, T., Trautwein, U., Meuers, D., & Nagengast, B. (2023). *Effective features of feedback in an intelligent language tutoring system* [Manuscript submitted for publication].

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist*, *37*, 91–106.

Pieronczyk, I., Parrisius, C., Bodnar, S., Colling, L., Deininger, H., Holz, H., Nuxoll, F., Wendebourg, K., Trautwein, U., Meurers, D., & Nagengast, B. (2022). *Motivation und Übungsverhalten von Schüler:innen bei der längerfristigen Verwendung eines* *intelligenten Tutorsystems im Fremdsprachenunterricht*. Jahreskonferenz der Gesellschaft für Empirische Bildungsforschung (GEBF), virtual.

Pili-Moss, D., Schmidt, T., Blume, C., Middelanis, L., & Meurers, D. (2022). *Enhancing EFL classroom instruction via an ICALL platform: Effects on language development and transfer to tasks*. EUROCALL Conference 2022, virtual.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, R., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [Pisa 2003: Scale documentation]*. Waxmann.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers and Education*, *100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Wu, T.-T., & Huang, Y.-M. (2017). A mobile game-based English vocabulary practice system based on portfolio analysis. *Journal of Educational Technology and Society*, *20*(2), 265–277.