Effective Features of Feedback in an Intelligent Language Tutoring System – A Randomized Controlled Field Trial

(Pre-registration)

Parrisius, C.*[1], Wendebourg, K.*[1], Rieger, S.[1], Loll, I.[1,2], Pili-Moss, D.[3], Colling, L.[4], Blume, C.[5], Pieronczyk, I.[1], Holz, H.[4,6], Bodnar, S.[4], Schmidt, T.[3], Trautwein, U.[1], Meurers, D.[4], & Nagengast, B.[1,7]

* both authors contributed equally

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

[2] University of Trier, Germany

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

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

[5] Competence Centre for Teacher Education and Research (DoKoLL), Technical University Dortmund, Germany

[6] Novatec Consulting GmbH

[7] The Brain & Motivation Research Institute (*b*MRI), Korea University, South Korea

Date of pre-registration: August 31, 2022

Introduction

Individual practice is a central component of learning at school, both in the classroom and at home (Trautwein et al., 2006). In this context, students who complete their homework with great commitment (i.e., fast and with high effort) show better learning performance (Flunger et al., 2015). Furthermore, motivation is a decisive factor contributing to learning success (Wentzel & Miele, 2016). Individualized, elaborate feedback on the completed tasks has been shown to be very effective in this regard (Elawar & Corno, 1985). However, in school practice—and especially in language teaching—the situation often arises that students have very different learning requirements. For this reason, teachers often argue that they are unable to provide immediate, individualized, concrete feedback due to time constraints alone. Part of the solution can be provided by intelligent tutoring systems (ITS) that aim at personalizing instructions to users by means of artificial intelligence technology. Such systems are often developed to give individual, adaptive, and scaffolded feedback without a human instructor that guides learners step by step to the correct solution.

Whereas positive effects have been found with various types of ITS, there have been few scientifically-evaluated ITS approaches to supporting foreign language learning; they are correspondingly absent altogether from Kulik and Fletcher's (2016) meta-analysis of the effectiveness of such systems. One of the few exceptions is the ITS FeedBook (Rudzewitz et al., 2017, 2018) for the subject English as a second language. The FeedBook provides scaffolded, individualized grammatical feedback regarding certain language means (e.g., simple past, regular verbs) for seventh-grade English learners. In a first efficacy study, Meurers et al. (2019) found larger learning gains regarding these language means as compared with students who only received default feedback merely including information on the correctness of the given answer (i.e., knowledge of correct response feedback; KCR).

The efficacy study by Meurers et al. (2019) was a first step to bring the FeedBook into the field and to test for its efficacy under real-life though still relatively controlled conditions.

As proposed in the field of intervention and implementation research (e.g., Gottfredson et al., 2015; Herbein et al., 2020), a next step towards scaling up such an intervention is to test for the effectiveness under ever more realistic conditions and ideally with increasing sample size. Thus, the first aim of the current investigation is to replicate the findings of this first study concerning the FeedBook by means of a follow-up trial testing for the effectiveness of said scaffolded feedback in a bigger sample and under controlled conditions. As a second goal, we seek to examine whether there are additional positive effects on students' English proficiency when—besides the scaffolded, individualized grammatical feedback—criterion-referenced feedback (i.e., a student dashboard providing information about students' learning progress and performance level in relation to a set learning goal) and motivational elements (i.e., a pedagogical agent presenting praise sentences after each exercise) are added to the FeedBook.

Past research provided evidence for the positive effects of both such game-based elements on students' achievement. Providing feedback about learning progress to learners can improve their learning regulation (Sedrakyan et al., 2020) and learning performance (Attali & van der Kleij, 2017; Hattie & Timperley, 2007; Sailer et al., 2017; Wilbert et al., 2010; Wollenschläger et al., 2011), whereas visualized information seems to be beneficial (Park & Jo, 2015). It is important that learners' reflection is stimulated in the process (Duijnhouwer et al., 2012). Motivational feedback has also been shown to enhance learning. For example, pedagogical agents (Lane, 2016), gamification elements (Chu & Fowler, 2020), and other attributional motivational feedback (Schrader & Grassinger, 2021) have been shown to improve learning performance. Acceptance of game-based learning approaches (e.g., perceived usefulness of the approach; Ninaus et al., 2017) and emotional engagement (Ninaus et al., 2019) are helpful in this regard.

The Present Study

The present study is embedded in a bigger research project ("Interact4School") and has two objectives that correspond to Projects 2 and 4 of the pre-registration of

Interact4School (Parrisius et al., 2022). First, we aim to replicate the findings of a first efficacy study on the scaffolded feedback as provided by the FeedBook (Meurers et al., 2019). Second, we aim at extending previous findings by investigating the additional and potentially interactive effects of criterion-referenced and motivational feedback on students' English proficiency after using the FeedBook for multiple weeks over the course of one school year. More precisely, we aim to answer the following research questions:

- RQ1: Do seventh-grade students show higher English proficiency for those language means for which they received scaffolded, individualized grammatical feedback by the FeedBook compared with students who received only default feedback (i.e., KCR feedback)? We hypothesize that, in line with the efficacy trial (Meurers et al., 2019), we will find positive effects for the group of students receiving scaffolded, individualized grammatical feedback compared with the group of students receiving default feedback for identical language means (e.g., simple past, regular verbs) in an achievement test focusing specifically on these language means.
- RQ2: Does criterion-referenced feedback and motivational feedback lead to bigger learning gains in seventh-grade students? We assume that the groups of students using the FeedBook with additional features (i.e., either *criterion-referenced feedback* or *criterion-referenced feedback and motivational elements*) as compared with the group of students using the FeedBook without such features (*original FeedBook*) show higher English proficiency. However, the question of whether additional motivational elements also lead to an even bigger effect as compared with "only" the add-on of criterion-referenced feedback is purely explorative and we do not formulate precise expectations in this regard.

Additionally, we seek to answer the following explorative questions for which we have no clear expectations:

- RQ3: Is there an interactive effect of receiving scaffolded, individualized grammatical feedback as well as *criterion-referenced feedback* or *criterion-referenced feedback and motivational elements*?
- RQ4: Do the intervention effects cumulate over the course of one school year?

Method

Data stem from a large-scale test of the FeedBook and were collected in 36 seventh-grade classrooms in 13 academic track schools in three German federal states (namely, Baden-Württemberg, North Rhine-Westphalia, and Hamburg) from September 2021 to July 2022. For Baden-Württemberg, the Ministry of Education and Cultural Affairs in Baden-Württemberg approved the study and the collection of the data (date of approval: July 19, 2021, file number: 31-6499.21/629/1). For Hamburg, the Authority for School and Vocational Training approved the study and the collection of the data (date of approval: July 13, 2021, file number e514.101.5000-002/221,025). In line with school law in North Rhine-Westphalia, according to which no superior ministry is responsible for approving school studies, the study and the collection of the data in this federal state were approved by the individual headmasters and headmistresses from the respective participating schools. The Ethics Committee for Psychological Research at the University of Tübingen confirmed that the study procedures were in line with ethical standards of research with human subjects (date of approval: August 4, 2021, file number: A2.5.4-184_ns).

Sample

The overall Interact4School study sample comprises a total of 13 schools with 33 teachers and their 36 classes from three different German federal states (eight from Baden-Württemberg, four from North Rhine-Westphalia, and one from Hamburg). For information on the recruitment process, we refer the reader to Parrisius et al. (2022). Within these schools, 844 students and their parents provided written consent to participate in the study, which corresponds to an overall participation rate of 91.1%. As part of the randomized controlled

field design, seven schools were asked to use the FeedBook in the school year 2021/22 (*FeedBook condition*), whereas six schools were allocated to a *waiting control condition* in which the FeedBook has not been used but business as usual has been taking place in the school year 2021/22 (for more detail, see Parrisius et al., 2022). As the focus of the current investigation lies on the group of students using the FeedBook, we will exclusively consider the subsample of schools in the FeedBook condition, consisting of $N_S = 7$ schools (three from Baden-Württemberg, three from North Rhine-Westphalia, and one from Hamburg) with $N_T = 21$ teachers and their $N_C = 24$ classes, comprising a total of $N_{St} = 616$ students with written consent (corresponding to a participation rate in the FeedBook condition of 96.7%).

The FeedBook and the Interact4School Intervention Conditions

The FeedBook contains exercises that are aligned with, and prepare for, a total of four complex Target Tasks that require the integration of several skills and competences, such as the use of certain grammatical structures, the knowledge of certain vocabulary fields, and listening, speaking, or writing skills. As these Target Tasks are generally communicative, they are carried out in class. The process of preparation for a Target Task is called a Task Cycle. The FeedBook contains digital practice material for each of the four Task Cycles. Each Task Cycle is planned to last approximately 3 weeks, respectively. The FeedBook also provides teachers with detailed lesson plans, which give them ideas of how to optimally integrate the digital exercises.

In order to investigate the aforementioned research questions, students were randomly assigned to different conditions. More precisely, classes within schools using the FeedBook were assigned to a set of three conditions: the *original FeedBook condition*, the *criterion-referenced feedback condition* (in which a criterion-referenced dashboard was additionally provided to the students), and the *criterion-referenced feedback and motivational elements condition* (in which next to the criterion-referenced dashboard also a pedagogical agent and affective praise feedback messages were provided to the students). Finally, individual students

within classes were randomly assigned to *Group A* or *Group B*, which differed concerning the extent of scaffolded, individualized grammatical feedback that has been provided to the students. Students in *Group A* received scaffolded, individualized grammatical feedback for all language constructs targeted in Task Cycles 1 and 3, while students in *Group B* received scaffolded, individualized grammatical feedback for all language constructs targeted in Task Cycles 2 and 4. That is, while working on a Task Cycle, always one half of the students received scaffolded feedback regarding specific language constructs or grammatical structures that are primarily focused on by the respective Task Cycle (e.g., Conditional Clauses Type 2). The other students did not receive this feedback for the same specific grammatical forms and constitute the respective control group. The allocation of feedback/no feedback alternates with each Task Cycle, so that by the end of the school year, all students in a class have had an equal opportunity to benefit from the various forms of grammatical feedback in the FeedBook. For a full overview of all intervention conditions realized in the Interact4School study, please consult Parrisius et al. (2022).

To address RQ1, students in *Group A* will be compared with students in *Group B*. To address RQ2, each set of two conditions at the class level (i.e., *original FeedBook condition*, *criterion-referenced feedback condition*, *criterion-referenced feedback and motivational elements condition*) will be compared with each other. Ultimately, we realized a 2×3 factorial design (i.e., (*Group A* vs *Group B*) × (*original FeedBook* vs *criterion-referenced feedback* vs *criterion-referenced feedback and motivation elements*)), thus also allowing to check for potential interaction effects between the implemented conditions at the different levels (RQ3). Considering the four different Task Cycles simultaneously allows for a longitudinal inspection of potential intervention effects (RQ4). For more detail, see the Statistical Analysis Plan section.

The teachers working with the FeedBook had access to a wide range of teaching material, including lesson plans, handouts and other additional paper-based exercise sheets,

solutions, PowerPoint-Slides, and additional video material for viewing and listening comprehension. Teachers were asked to use the teaching material while working on the Task Cycles and to conduct the Target Tasks during class. Whereas teachers in the *waitlist control condition* did not receive any material, material provided to the teachers in the *FeedBook condition* did not differ between the different intervention conditions.

The intervention materials (i.e., the English exercises used in the FeedBook as well as the additional material provided to the teachers to prepare their students for the Target Tasks in class, including paper-based handouts, slides for presentations, and videos) were developed by a group of experts, consisting of teachers in Hamburg and Baden-Württemberg as well as researchers of English didactics. The entire teaching material is based on the contents of the textbook Camden Town 3 (version 2012) by the Westermann publishing house and is thus aligned with the year seven English curricula of the various German federal states in academic-track schools.

Instruments

During the Interact4School study, the FeedBook was used throughout an entire school year including four Task Cycles, each lasting approximately 3 weeks. Achievement test data and 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). For the present study, students' achievement test results at the respective time points are of particular interest and constitute the primary outcome of the current investigation. However, the full list of variables that were assessed during the Interact4School study can be retrieved from the pre-registration of the full Interact4School study design (Parrisius et al., 2022).

English Proficiency. Students' English proficiency concerning the language means of interest were assessed by an achievement test before and after each Task Cycle (i.e., at all five time points with a total of eight tests because of four pretests and four posttests spread across

T1 to T5). The achievement tests focus on those grammatical structures that are the target constructs of the respective Task Cycles. Part of the achievement tests were developed for the first FeedBook study (Meurers et al., 2019) and were adapted and extended for the current study. They consist of 30 items each and constitute either a pre- or posttest for the respective Task Cycles. An overview of the number of items per achievement test and the language constructs they cover is provided in the Appendix. After collecting the data, we realized some ambiguity concerning a few items and decided to exclude them from the analyses. This involves two items from pretest Task Cycle 1 (i.e., "I (not go) to see the movie. I am not a fan of horror films." and "This summer I (buy) a hat for my cousin in Italy") and one item from posttest Task Cycle 1 (i.e., "This winter we (spend) a lot of time at home."). Ultimately, we will end up using 28/29 items for pretest/posttest Task Cycle 1 and 30 items each for pretest/posttest Task Cycle 2, pretest/posttest Task Cycle 3, and pretest/posttest Task Cycle 4.

Covariates. In addition to students' English proficiency, we will consider a number of covariates. Whether a variable will be included as a covariate will be decided based on recommendations by the What Works Clearinghouse (Institute of Education Sciences, 2022). That is, variables will be included as covariates if their intervention group-specific means differ by $d > .05$. We will test for group differences concerning the following variables reported by the schools, stemming from another achievement test, and as self-reported by the students at T1 (for a full overview of items, see Parrisius et al., 2022): students' gender, age, English grade at the end of grade level 6 (reported by the schools); general English achievement (C-test); English motivation (self-concept, intrinsic value, attainment value, utility value, cost, ideal L2 self), English effort, homework effort, conscientiousness, computer proficiency, migration background, parents' educational background, language at home, use of English in everyday life (self-reported by the students).

Statistical Analysis Plan

The significance level for all analyses will be set to 5% (two-tailed).

Achievement Test Evaluation. To evaluate the quality of the achievement tests (i.e., for each of the eight tests separately), a range of analyses based on item response theory (IRT) will be conducted. The overall goal is to derive person parameters that subsequently can be used in the analyses to answer our four research questions. To do so while considering the constitution of the data (i.e., item discriminations), we will scale the responses of the students by multiple approaches (e.g., with the Rasch model; Rasch, 1960; or the 2PL Birnbaum model; Birnbaum, 1968). We will decide for the final model based on the fit of the models to the test data.

Furthermore, all items will be examined regarding item fit and differential item functioning (for the following variables: intervention group, sex, English as language at home). Finally, based on the scaling results, we will derive person parameters (i.e., weighted maximum likelihood estimates; WLE; Warm, 1989; and expected a posteriori estimates; EAP; Uebersax, 1993) from the final model.

Research Questions 1 to 3. To answer RQ1, RQ2, and RQ3, we will specify one multilevel two-way ANCOVA per Task Cycle in Mplus (Muthén & Muthén, 1998-2017), thus, considering our 2×3 factorial design. More precisely, we will investigate mean differences in students' English proficiency after each Task Cycle between the different treatment groups. We will specify one model per Task Cycle or outcome, respectively (i.e., four models in total). For this purpose, we will separately conduct two-level analyses with the students at Level 1 and the classes at Level 2. The clustering of classrooms within schools will be accounted for by including dummy variables for the schools (i.e., fixed effects for Level 3).

To estimate the effects of receiving scaffolded versus default feedback (RQ1), of receiving the original FeedBook versus the criterion-referenced feedback versus the criterion-referenced feedback plus motivational elements (RQ2), or any combination of these (RQ3), the following independent variables will be included in the model: a dummy variable

indicating *Group B* with *Group A* as the comparison group will be used as predictor at the individual level; two dummy variables indicating the intervention conditions *criterion-referenced feedback* as well as *criterion-referenced feedback and motivational elements* as compared with the *original FeedBook* condition will be used as predictors as the class level; finally, two interaction terms between the individual-level dummy variable and the class-level dummy variables will be included at the individual level. To answer RQ1 and RQ2, we will consult the main effects of the independent variables. To answer RQ3, we will report the interaction effects between the two independent variables of students' English proficiency.

Furthermore, we will include the pretest score of the respective outcome variable as a covariate in the analyses. To yield more accurate estimates of the intervention effects, we will additionally include variables as covariates for which we will find substantial differences between the intervention groups before introducing the Task Cycles, that is, at T1 ($d > .05$; see Institute of Education Sciences, 2022). 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).

Research Question 4. Ultimately, we aim at testing for cumulative intervention effects. To do so, we will model a multivariate multilevel ANCOVA, including all four English proficiency outcomes simultaneously. For this purpose, we will specify all four aforementioned two-way ANCOVAS simultaneously in Mplus and conduct Wald tests as omnibus tests for the overall intervention effect at the class level. We cannot test for cumulative intervention effects at the individual level (i.e., effects of receiving scaffolded vs default feedback) because of the within-person design (i.e., students received scaffolded feedback during two of the four Task Cycles and default feedback during the remaining two Task Cycles).

Missing Values. As is common in longitudinal designs, we expect missing data at all measurement occasions because of absence of individual students or because of nonresponses to single items. Additionally, eight of the 24 classes did not manage to work on Task Cycle 4 (e.g., because of missing time due to COVID side effects) and, consequently, did not participate in T5. That is, there is no posttest Task Cycle 4 data available for these classes. All analyses will thus be conducted using full information maximum likelihood estimation implemented in Mplus (Graham, 2009). To make the necessary assumption of missing-at-random more plausible, 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).

Knowledge of Data

Work in Progress and Previous Publications

Multiple authors of this pre-registration have previously used (preliminary versions of) the Interact4School data for different research questions already. Even though earlier versions of the dataset have been used before, including at least some of the variables and measures in this study, all analyses were on a mutually exclusive subset of the dataset (i.e., a subset of participants and/or a selection of time points). Most importantly, the co-authors who will conduct the analyses for the current investigation (Parrisius and Rieger) have not yet worked with the data. Previous work, including the measures used, is listed below, and in each case, it is indicated who among the authors was involved.

- Deininger et al. (in prep): This project focuses on the prediction of English proficiency by behavioral trace data. Students' English proficiency assessed after each Task Cycle will be used as the outcomes, students' interactions with the system will be used as basis to calculate the input variables (e.g., number of exercises worked on, time on task). At this point in time, no analyses concerning students' English proficiency as measured at the respective time

points T1 to T5 have been conducted. However, to check for a general relationship between the raw interaction data and students' English competences, a first machine learning model trying to predict the correctness of individual task field entries based on information given about this specific task field (e.g., task type) and the specific student (e.g., age) was trained successfully. Only Deininger was involved in this initial analysis. Parrisius, Pieronczyk, Colling, Trautwein, Meurers, Kasneci, and Nagengast are involved as co-authors.

- Loll (in prep): In her Master's thesis, Loll will examine the intervention effects of scaffolded versus default feedback on students' posttest value for Task Cycle 1 at T2. For this purpose, she will consider students' English proficiency concerning the language means targeted during Task Cycle 1 as predictor and outcome (i.e., pretest and posttest Task Cycle 1). Additionally, she will include students' gender as a covariate. Parrisius and Rieger were involved in planning the statistical analyses. However, no analyses have been conducted at this point in time.
- Pieronczyk et al. (in prep): In this study, Pieronczyk 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, as well as their self-reported English self-concept, intrinsic value, attainment value, utility value, cost, conscientiousness, effort regarding English, and effort regarding English

homework at T1 were used. That is, students' achievement test results are not in the focus of this investigation but might be included in the future as an additional variable to validate our behavioral measure of student engagement. Further analyses are currently in preparation, and we plan to finalize a manuscript based on the full set of results. Overlapping co-authors include Parrisius, Wendebourg, Trautwein, Meurers, and Nagengast at this point in time.

- 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 *original FeedBook*, *criterion-referenced feedback* and *criterion-referenced feedback and motivational elements* for Task Cycle 1. For this purpose, they first checked for statistical significance regarding the learning gains from T1 to T2 (i.e., from pretest Task Cycle 1 to posttest Task Cycle 1) within each of the three groups separately. Afterwards, they descriptively compared the learning gains across groups. The nested data structure, the intervention conditions at the individual level (*Group A* vs *Group B*), or other covariates have not been considered. Prior to publishing this pre-registration, they have not shared any results with the co-authors who wrote the main part of the current preregistration (Parrisius & Wendebourg) nor with the co-authors who will conduct the analyses for the current investigation (Parrisius & Rieger).
- Parrisius et al. (2022): This publication is the pre-registration of the Interact4School study design. All listed co-authors except for Rieger and Loll have also been involved in this pre-registration. No data has been worked with for this purpose.

- Pili-Moss et al. (2022): In this conference contribution presented at EUROCALL 2022, Pili-Moss, Schmidt, Blume, Middelanis, and Meurers investigated the efficacy of FeedBook in a sub-sample of 77 students (three intact classes) who used the platform within the *criterion-referenced feedback and motivational elements 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 Cycles 2 and 3, as well as data from the written Target Task performed by the students at the end of 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).
- 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.

Prior Knowledge About the Dataset by the Co-Authors Who Will Conduct the Analyses (Parrisius & Rieger)

The achievement test and survey data assessed at T1 to T5 have been cleaned in great parts. This work has been executed by student assistants not involved in the current investigation as well as by Ines Pieronczyk. The cleaning process was supervised by Cora Parrisius.

There are some variables that overlap between this current question and previous work (Blume et al., 2022; Pili-Moss et al., 2022). Specifically, students' English proficiency regarding Task Cycle 1 (i.e., pretest Task Cycle 1 at T1 and posttest Task Cycle 1 at T2) and their English proficiency regarding Task Cycles 2 and 3 (i.e., pretest and posttest each; however, only in a sub-sample of 77 students) has been looked at before. However, as the two co-authors who will conduct the analyses for the current investigation, we were not involved in these projects and are not aware of their results.

We have not seen any descriptive or other statistics of the outcome variables planned to be used in the current investigation.

References

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education*, *110*, 154–169. https://doi.org/10.1016/j.compedu.2017.03.012

Birnbaum, A. (1968). Some latent trait models and their use in inferring and examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

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.

Chu, M.-W., & Fowler, T. A. (2020). Gamification of formative feedback in language arts and mathematics classrooms: Application of the learning error and formative feedback (LEAFF) model. *International Journal of Game-Based Learning*, *10*(1), 1–18. https://doi.org/10.4018/IJGBL.2020010101

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989x.6.4.330

Duijnhouwer, H., Prins, F. J., & Stokking, K. M. (2012). Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process, and performance. *Learning and Instruction*, *22*(3), 171–184. https://doi.org/10.1016/j.learninstruc.2011.10.003

Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology*, *77*(2), 162–173. https://doi.org/10.1037/0022-0663.77.2.162

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press. https://doi.org/10.4135/9781412983907.n1075

Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., & Schnyder, I. (2015). The Janus-faced nature of time spent on homework: Using latent profile analyses to predict academic achievement over a school year. *Learning and Instruction*, *39*, 97–106. https://doi.org/10.1016/j.learninstruc.2015.05.008

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893–926. https://doi.org/10.1007/s11121-015-0555-x

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

Herbein, E., Golle, J., Nagengast, B., & Trautwein, U. (2020). Förderung von Präsentationskompetenz: Schrittweise Implementation und Effektivitätsüberprüfung eines Präsentationstrainings für Grundschulkinder. *Zeitschrift für Erziehungswissenschaft*, *23*, 83–120. https://doi.org/10.1007/s11618-019-00923-y

Institute of Education Sciences. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. 234.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Lane, H. C. (2016). Pedagogical agents and affect: Molding positive learning interaction. In *Emotions, technology, design, and learning* (pp. 47–62). Academic Press. https://doi.org/10.1016/B978-0-12-801856-9.00003-7

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764–802. https://doi.org/10.1080/00273170903333665

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. (1998-2017). *Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén*. https://doi.org/10.1111/j.1600-0447.2011.01711.x

Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., Karnath, H.-O., & Moeller, K. (2019). Increased emotional engagement in game-based learning—A machine learning approach on facial emotion detection data. *Computers and Education*, *142*, Article 103641. https://doi.org/10.1016/j.compedu.2019.103641

Ninaus, M., Moeller, K., McMullen, J., & Kiili, K. (2017). Acceptance of game-based learning and intrinsic motivation as predictors for learning success and flow experience. *International Journal of Serious Games*, *4*(3), 15–30. https://doi.org/10.17083/ijsg.v4i3.176

Park, Y., & Jo, I.-H. (2015). Development of the learning analytics dashboard to support students' learning performance. *Journal of Universal Computer Science*, *21*(1), 110–133.

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

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Mesa Press.

Rudzewitz, B., Ziai, R., De Kuthy, K., & Meurers, D. (2017). Developing a web-based workbook for English supporting the interaction of students and teachers. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017.*, *134*, 36–46.

Rudzewitz, B., Ziai, R., De Kuthy, K., Möller, V., Nuxoll, F., & Meurers, D. (2018). Generating feedback for English foreign language exercises. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 127–136. https://doi.org/10.18653/v1/w18-0513

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

Schrader, C., & Grassinger, R. (2021). Tell me that I can do it better. The effect of attributional feedback from a learning technology on achievement emotions and performance and the moderating role of individual adaptive reactions to errors. *Computers and Education*, *161*, Article 104028. https://doi.org/10.1016/j.compedu.2020.104028

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, *107*, Article 105512. https://doi.org/10.1016/j.chb.2018.05.004

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, *90*(2), 334–349. https://doi.org/10.1037/0022-3514.90.2.334

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). National Foundation for Educational Research.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment apropriateness. *Journal of the American Statistical Association*, *88*, 421–427.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wentzel, K. R., & Miele, D. B. (Eds.). (2016). *Handbook of motivation at school* (2nd ed.). Routledge.

Wilbert, J., Grosche, M., & Gerdes, H. (2010). Effects of evaluative feedback on rate of learning and task motivation: An analogue experiment. *Learning Disabilities: A Contemporary Journal*, *8*(2), 43–52.

Wollenschläger, M., Möller, J., & Harms, U. (2011). Effekte kompetenzieller Rückmeldung beim wissenschaftlichen Denken. *Zeitschrift für Pädagogische Psychologie*, *25*(3), 197–202.

Appendix

Table A1

Overview of Tests, Time Points, Language Constructs, Sample Items and Target Answers, and the Full Number of Items Included in the Achievement Tests for the Respective Language Constructs

Achievement Test	Time point	Language construct	Sample item	Target answer(s) for sample item	Number of items
Pretest Task Cycle 1	T1	Simple past	Two weeks ago Martin (walk) 10 miles.	Walked	15
		Modals	I think you ___ help your brother more.	Should	5
		Gerund	I would try (talk) to her.	Talking	10
					Total: 30
Posttest Task Cycle 1	T2	Simple Past	An hour ago Susan (receive) an important phone call.	Received	15
		Modals	You ___ do something about the situation if you can.	Should	5
		Gerund	The jar doesn't open. Let us try (use) a piece of cloth.	Using	10
					Total: 30
Pretest Task Cycle 2	T2	Conditional sentence type 2 (simple past)	They could apply at university, if they (go) to school for at least 12 years.	Went	13
		Conditional sentence type 2 (would)	If you sang at night, if (disturb) your neighbours.	Would disturb | could disturb	7
		Comparative	Paul wasn't (smart) than other students, he just studies really hard.	Smarter	5
		Superlative	Is January the (cold) month in Europe?	Coldest	5
					Total: 30

Posttest Task Cycle 2	T3	Conditional sentence type 2 (simple past)	If Paul (try) hard enough, I think he could make it.	Tried	13
		Conditional sentence type 2 (would)	If he really cared about you, he (not miss) your party.	Wouldn't miss | couldn't miss	7
		Comparative	It is much (safe) to wear a helmet when you go cycling.	Safer	5
		Superlative	How old is the (old) tree in the world?	Oldest	5
					Total: 30
Pretest Task Cycle 3	T3	Questions (simple present)	A: Mark and Amy love the Canary Islands in winter. B: Really? Where (fly) to when they visit?	Do they fly	15
		Questions (simple past)	A: My parents spent their honeymoon in Capri. B: ____ decide to go there? A: Because it was cheaper than the Maldives.	Why did they	15
					Total: 30
Posttest Task Cycle 3	T4	Questions (simple present)	A: Patrick and Nick adore trekking in the North of England. B: Really? Where (go) when they are there?	Do they go	15
		Questions (simple past)	A: Jill and I spent a night in Dover last August. B: ____ decide to stay there? A: Because we missed the ferry.	Why did you	15
					Total: 30
Pretest Task Cycle 4	T4	Going to-future	A: Rob say he is still stuck in traffic. B: Oh no, he (miss) his flight.	Is going to miss	10
		Will-future	I promise I (study) harder next year.	Will study	10
		Relative clause (defining)	The man ___ is wearing a hat is my uncle.	Who	10
					Total: 30

Posttest Task Cycle 4	T5	Going to-future	A: It was minus 20 again this morning. B: It looks like it (be) a cold winter.	Is going to be	10
		Will-future	You have my word I (be) back before dinner.	Will be	10
		Relative clause (defining)	The dog ___ is barking at your dad is not very dangerous.	Which	5
		Relative clause (non-defining)	Nuclear missiles, ___ are a serious threat to human kind, should be banned.	Which	5
					Total: 30