


Exercise parameters influencing exercise difficulty

Tanja Heck^a and Detmar Meurers^b

^aDepartment of Linguistics, University of Tübingen, , tanja.heck@uni-tuebingen.de and ^bDepartment of Linguistics, University of Tübingen, , detmar.meurers@uni-tuebingen.de

How to cite: Heck, T.; Meurers, D. (2023). Exercise parameters influencing exercise difficulty. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik.
<https://doi.org/10.4995/EuroCALL2023.2023.16921>

Abstract

Macro-adaptive systems aim to assign practice exercises to language learners that match their proficiency levels. While learner-dependent parameters of exercise difficulty need to be considered online, learner-independent parameters can inform an exercise's difficulty level in a resource-efficient offline procedure. We present an evaluation of learners' responses to form-based grammar exercises that aims to identify learner-independent exercise parameters affecting exercise complexity. The results indicate that the exercise type can yield coarse-grained complexity estimates, whereas exercise type specific features can inform more fine-grained estimates. For fine-grained estimates, we show that syntactic variants significantly impact exercise difficulty. Since there is strong variation between learning targets and learners with respect to the impact of different exercise parameters on a learner's performance, exercise difficulty can only be reliably determined if the exercises are created in a systematic way and by also considering characteristics of the learner.

Keywords: ICALL, ILTS, exercise difficulty.

1. Introduction

Learners' performance on language exercises is closely linked to exercise difficulty (Buckledee, 2008), which depends on learner-specific parameters on the one hand, and on exercise-specific parameters on the other hand (Pelánek et al., 2021). Macro-adaptive systems that assign exercises to learners in a personalized manner for best possible learning outcomes, therefore need to consider both types of parameters when selecting an exercise (Liu et al., 2021). Learner-specific parameters, such as cognitive abilities or personal experience, are dynamic features and therefore need to be factored in online individually for each learner at the time of selecting an exercise (Kunichika et al., 2002). Exercise complexity, on the other hand, comprises learner-independent, static parameters of exercise difficulty, such as linguistic complexity of the textual material and characteristics of the exercise types, thus constituting a property of the exercise. In order to provide this meta-information to the exercise selection algorithm of a macro-adaptive system, exercise complexity can therefore be determined once offline before adding the exercise to the system's resources (Pandiarova et al., 2019). Considering the stress that online calculations put on a system's performance, these offline calculations should take into account all learner-independent parameters of exercise difficulty in order to speed up the system's exercise selection process at runtime. It is thus necessary to not only identify parameters impacting exercise difficulty, but also to determine which of them are learner-independent.

Little is known about the impact of different exercise parameters on learner-independent exercise complexity and learner-dependent exercise difficulty. With our analysis of real-world learner data from two field studies, we shed light onto the relevance of a selection of exercise features of form-based, English grammar exercises in order to provide macro-adaptive systems with the means to more effectively and efficiently select exercises tailored to the individual learner.

2. Data

The evaluations are based on data collected from German 7th grade learners of English in the Interact4School (I4S) (Parrisius, Pieronczyk, et al., 2022; Parrisius, Wendebourg, et al., 2022) and the Digbindiff (Didi)¹ studies, which are based on the Intelligent Language Tutoring System FeedBook. The system offers exercises for practice of English as a second language, incorporating intelligent feedback provided as a learner works on the exercises. While both studies were conducted over the course of a school year, I4S focused on motivational aspects in a task based setting whereas Didi investigated the effects of user-adaptive exercise sequencing. For form-based grammar exercises, the FeedBook covers the seven exercise types Fill-in-the-Blanks (FiB), Single Choice (SC), Jumbled Sentences (JS), Categorization, Memory, Short Answers (SA), and Mark-the-Words (MtW). The exercises provide a total of 3,143 actionable elements, ($N_{I4S}=1,140$; $N_{Didi}=2,003$) such as blanks of FiB or SC exercises, chunks of JS exercises, elements to sort into a category, Memory pairs, answers to SA questions, or clickable words in MtW exercises. They are distributed across 11 distinct learning targets ($N_{I4S}=9$; $N_{Didi}=4$). While revising and re-submitting an exercise was possible in the studies, the evaluations only consider the first submission ($N_{I4S}=153,596$; $N_{Didi}=120,431$).

3. Evaluation

3.1. Exercise parameters impacting exercise difficulty

In order to determine those exercise parameters that are most predictive of exercise difficulty, we trained statistical models from the Python *scikit-learn* library to predict exercise difficulty based on a selection of exercise parameters. These parameters cover: a) generally applicable parameters including the number of actionable elements in the exercise or the length of an actionable element; b) exercise type specific parameters such as the number of distractors of SC exercises, the number of chunks of a JS exercise, the number of categories of a Categorization exercise, the number of pairs of a Memory exercise, or whether a FiB exercise requires the learner to determine the correct lemma in addition to transforming it into the correct form; and c) the exercise type itself. All predictors were encoded as numerical features. Difficulty of the actionable elements was operationalized as item difficulty scores obtained from an Item Response Theory model². While the continuous scores served directly as outcome variables for the regression models, they were transformed into categorical values for the classification models. Since the FeedBook distinguishes three proficiency levels of learners, we applied the same amount of exercise difficulty levels. The thresholds between the three levels were determined through K-means clustering. All employed statistical models support feature ranking, which allows to easily determine their most predictive features. In order to identify the overall most predictive features, we added up the predictor ranks of all regression and those of all classification models, thus obtaining overall rankings for regression and classification, respectively.

¹ <http://digbindiff.de>

² The implementation is based on the Rasch model of the TAM package for R.

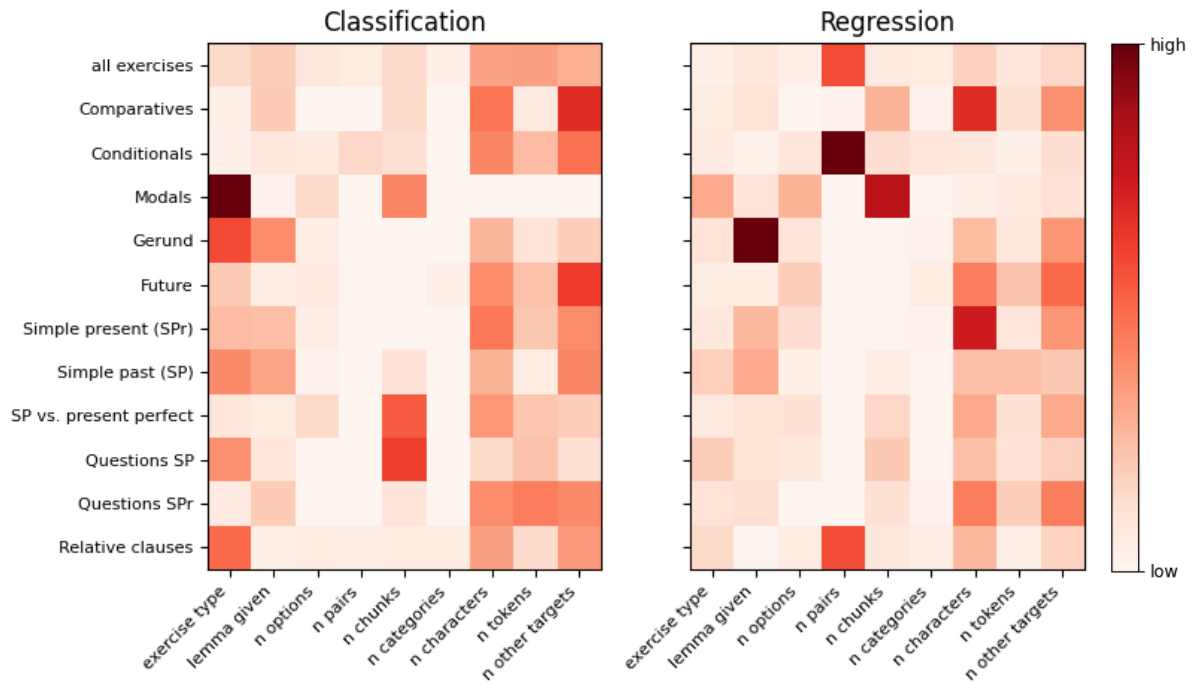


Figure 1. Feature importances in statistical models predicting exercise difficulty.

The heatmaps in Figure 1 assign colours of increasing darkness to parameters on the x-axis for learning targets on the y-axis the more important the parameter is for that learning target. They show that for the generally applicable parameters, the rankings are rather similar across learning targets for both the regression and the classification models. They all occupy ranks in the middle ranges, indicating that while they do not constitute the most informative of the evaluated parameters, their predictive power is rather constant and reliable across exercises. The exercise type specific features show considerably more variance across learning targets especially with regression, appearing at both extremes of the rankings. A general trend sees the exercise type as rather important for classification, whereas it ranks among the least predictive features for regression. Type-specific parameters hold more predictive power with those models. The regressions per exercise type, illustrated in the heatmap in Figure 2, highlight that the parameters applicable to only a particular exercise type indeed are more important for the respective exercises. This might indicate that the exercise type can inform coarse-grained difficulty estimations, while fine-grained distinctions require more detailed, type-specific parameters.

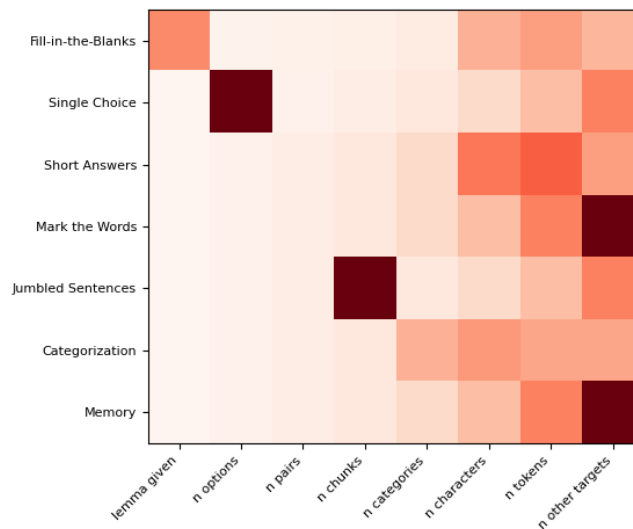


Figure 2. Feature importances for regression predicting exercise difficulty per exercise type.

3.2. Impact of syntactic variations on exercise difficulty

Since both classification and regression highlight the relevance of the generally applicable parameters, including linguistic complexity features such as token and character counts, we took a closer look at the impact of a range of linguistic complexity features on exercise difficulty. Didi's exercises of the learning targets *conditionals* and *relative clauses* were generated with the approach to systematic variability presented by Heck et al. (2022), so that they contain learner data for exercises with identical textual material and varying only in a selection of controlled, syntactic features. These variations target the *clause order*, *targeted clause* and *negation of clauses* for conditionals, and *clause order* for relative clauses. In order to determine their effect on exercise difficulty, we compared the distributions of difficulty scores across the different realizations of the variations. We tested for statistical significance with a two-tailed T-test, applying the commonly used threshold of $p < .05$ for statistical significance.

For the overall dataset, all effects were statistically significant, indicating that syntactic variations indeed impact exercise difficulty. In order to verify whether this is the case for all exercise types, we performed evaluations for the individual types. The violin plots given in Figure 3 illustrate that the results vary considerably across different exercise types, yet almost all effects are statistically significant. For the clause order of conditionals, exercises of almost all types are more difficult when putting the if-clause before the main clause. The effect, although not significant ($t=1.7796$, $p=.0760$), is inverted for Categorization exercises. With respect to the targeted clause, exercise items are slightly more difficult if the actionable element is in the if-clause rather than in the main clause with significant effects for all exercise types. Although we hypothesize that items simultaneously targeting both clauses are more difficult, the dataset does not contain according exercises. The question whether this variation of the exercise parameter makes a difference thus remains an open research question. Concerning negation, there is a statistically significant effect indicating that exercises are easiest if only the if-clause is negated, and most difficult when both clauses are negated. The only contradictory – and non-significant – effect appears with JS exercises between negation of the if-clause and of both clauses ($t=-.1993$, $p=.8421$). For relative clauses, exercise items appear more difficult if the prompt gives the clause corresponding to the relative clause before that corresponding to the main clause. The effect is significant for all exercise types but Memory ($t=-.4771$, $p=.6333$).

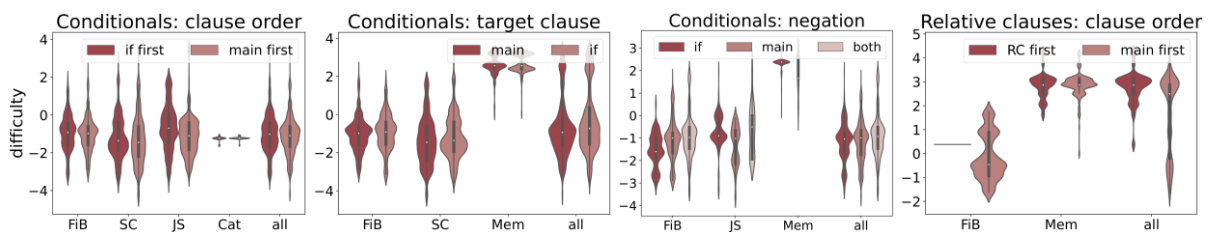


Figure 3. Difficulty distributions for exercises of syntactic variations.

3.3. Learner dependence of exercise parameter predictiveness

Since the exercise type holds predictive power for coarse-grained difficulty estimates, we investigated whether this is a global parameter for general exercise complexity or has to be determined on a per-learner basis. To this purpose, we determined the rankings of exercise types with respect to their difficulty for each learner and compared the distributions of ranking positions across learners. Following Pelánek et al.'s (2021) approach, we operationalized exercise difficulty as learners' performance on the exercises. More precisely, the rankings were created based on the ratio of incorrect to all submissions for an exercise item. If two exercise types obtained similar accuracies, they were both assigned the same rank.

The results are visualized in the heatmaps in Figure 4, where darker colours of a matrix cell indicate higher numbers of learners for which the exercise type is placed at the corresponding rank relative to the other exercise types. A single dark cell and white colour for all remaining cells of the row would indicate perfect agreement in

ranking for that exercise type among all learners; uniform colouring of a row would indicate highest possible diversity among learners. The heatmaps show that while there are differences between learners, there are definite tendencies as to what exercise types are most often solved incorrectly. The exact rankings are not identical for all learners, yet for most exercise types, the most frequent rank positions correspond to adjoining cells of the matrix, indicating that rough difficulty placements are similar for most learners. The results are clearest for Fib and SC exercises, where the majority of learners make the least errors amongst all exercise types. MtW exercises constitute an interesting case as they feature two peaks at opposite ends of the ranking, indicating that they are among the least critical types for some learners, and among the most critical ones for others. JS exercises exhibit a similar trend, although they paint an overall more diverse picture with a number of learners also placing them in the middle ranking positions. SA exercises generally constitute the most difficult exercise type. In addition, the heatmaps differ considerably from one learning target to another, sometimes even reversing ranking positions. Assuming that learners perform better on exercises if they are more proficient in the skill that the exercise practices, this seems to indicate that the exercises of the dataset do not encode the same skill for an exercise type across all learning targets. It is therefore imperative to systematically create exercises so that they target the same skill for a particular exercise type, or else to also consider the skill when estimating an exercise's complexity.

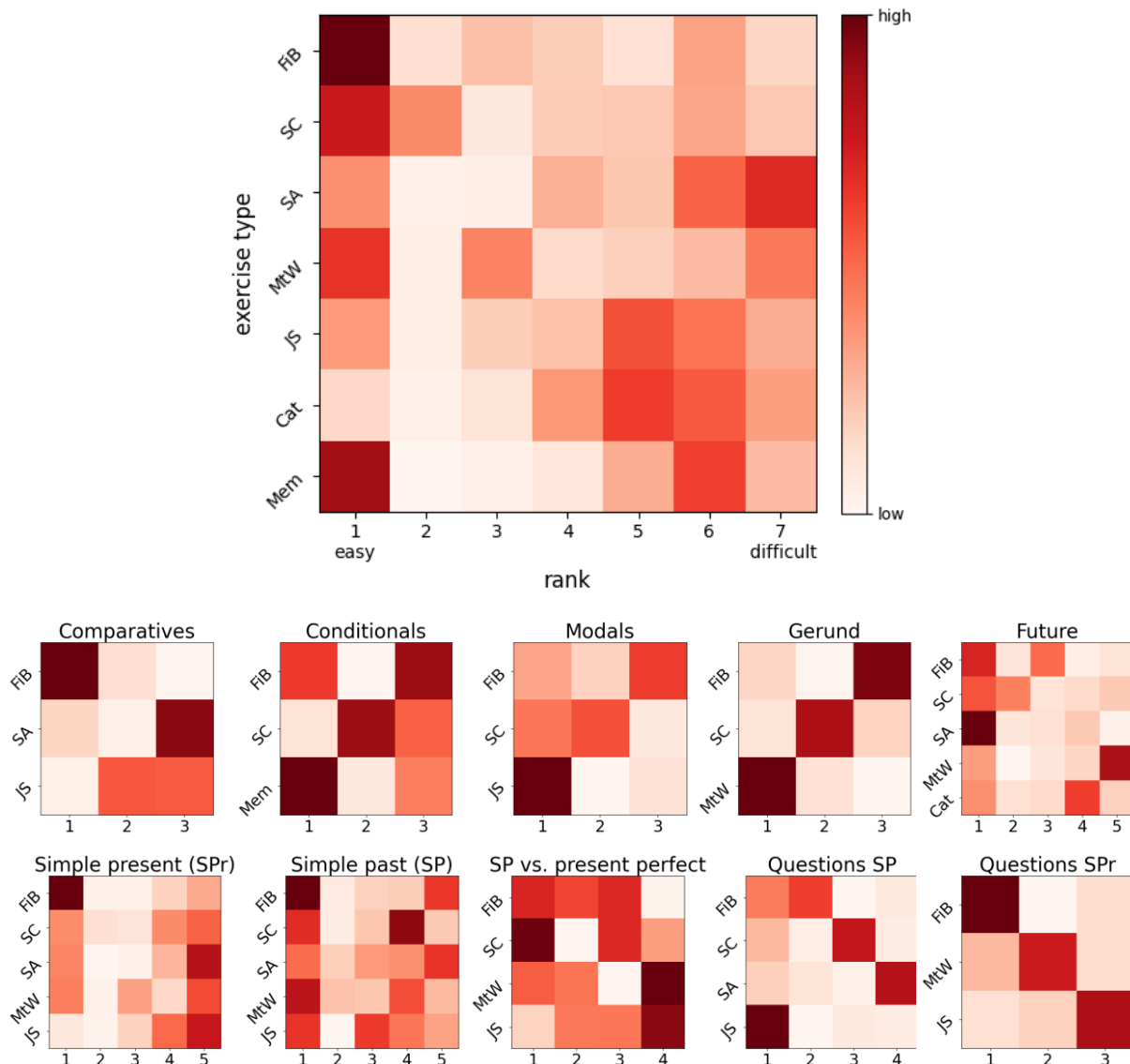


Figure 4. Distributions of learner-based exercise difficulty rankings

4. Conclusions

We presented an evaluation of exercise parameters with respect to their influence and learner-dependence on exercise difficulty. We found that the exercise type is indicative of coarse-grained difficulty estimates, while exercise-type specific parameters can yield more fine-grained predictions. Although the parameters hold some general predictive power, even coarse estimates are best based on exercise features in conjunction with learner characteristics. Syntactic variants do have an impact on exercise difficulty, so that macro-adaptive systems should take these linguistic features into account when calibrating exercise difficulty. While the approach presented by Pandarova et al. (2019) could be extended to consider parameters of syntactic variations, also taking learner characteristics into account requires maintaining and consulting a learner model at runtime.

References

- Buckledee, S. (2008). Motivation and Second Language Acquisition. In Z. Dörnyei & R. W. Schmidt (Eds.), *ELOPE: English Language Overseas Perspectives and Enquiries* (Vol. 5, Issues 1–2, pp. 159–170). Second Language Teaching & Curriculum Center, University of Hawai'i at Mānoa. <https://doi.org/10.4312/elope.5.1-2.159-170>
- Heck, T. & Meurers, D. & Nuxoll, F. (2022). Automatic exercise generation to support macro-adaptivity in intelligent language tutoring systems. *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, Research-publishing.net, pp. 162-167. <https://doi.org/10.14705/rpnet.2022.61.1452>
- Kunichika, H., Urushima, M., Hirashima, T., & Takeuchi, A. (2002). *A Computational Method of Complexity of Questions on Contents of English Sentences and its Evaluation*. 97–101. <https://doi.org/10.1109/CIE.2002.1185873>
- Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A survey of knowledge tracing. *ArXiv Preprint ArXiv:2105.15106*.
- Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the Difficulty of Exercise Items for Dynamic Difficulty Adaptation in Adaptive Language Tutoring. *International Journal of Artificial Intelligence in Education*, 29(3), 342–367. <https://doi.org/10.1007/s40593-019-00180-4>
- Parrisius, C., Pieronczyk, I., Blume, C., Wendebourg, K., Pili-Moss, D., Assmann, M., Beilharz, S., Bodnar, S., Colling, L., Holz, H., & others. (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., Loll, I., Pili-Moss, D., Colling, L., Blume, C., Pieronczyk, I., Holz, H., Bodnar, S., & others. (2022). *Effective Features of Feedback in an Intelligent Tutoring System-A Randomized Controlled Field Trial (Pre-Registration)*. PsychArchives. <https://doi.org/10.23668/psycharchives.8152>
- Pelánek, R., Effenberger, T., & Čechák, J. (2021). Complexity and Difficulty of Items in Learning Systems. *International Journal of Artificial Intelligence in Education*, 32, 1–37. <https://doi.org/10.1007/s40593-021-00252-4>