On the Relevance and Learner Dependence of Co-text Complexity for Exercise Difficulty

Tanja Heck
Universität Tübingen / Germany
tanja.heck@uni-tuebingen.de

Detmar Meurers
Universität Tübingen / Germany
detmar.meurers@uni-tuebingen.de

Abstract

Adaptive exercise sequencing in Intelligent Language Tutoring Systems (ILTS) aims to select exercises for individual learners that match their abilities. For exercises practicing forms in isolation, it may be sufficient for sequencing to consider the form being practiced. But when exercises embed the forms in a sentence or bigger language context, little is known about how the nature of this co-text influences learners in completing the exercises.

To fill the gap, based on data from two large field studies conducted with an English ILTS in German secondary schools, we analyze the impact of co-text complexity on learner performance for different exercise types and learners at different proficiency levels. The results show that co-text complexity is an important predictor for a learner's performance on practice exercises, especially for gap filling and Jumbled Sentences exercises, and particularly for learners at higher proficiency levels.

1 Introduction

Exercise difficulty, which constitutes the probability of a learner answering the exercise correctly, plays an important role in intelligent tutoring systems. Macro-adaptive systems in particular rely on it to select exercises at the learner's proficiency level. Assigning a global difficulty score to an exercise, however, fails to consider the many facets of factors contributing to exercise difficulty and the varied learner profiles instantiating them (Beinborn, 2016). Approaches like Multidimensional Item Response Theory (Park et al., 2019) and Knowledge Tracing (Liu et al., 2021b) address this issue by tracking individual skills instead of a single, accumulated one. Yet they usually focus on the skills the learner is supposed to acquire through the exercises. More stable skills such as a learner's language affinity or their general language proficiency are therefore often neglected in these approaches. Such skills might not be relevant in mechanical drill exercises that practice the linguistic forms of the learning target in isolation (Wong and Van Patten, 2003). However, contextualized exercises, which practice linguistic constructions in the broader context of a coherent text, require learners to understand the clues provided by this co-text in order to give the correct answer (Walz, 1989). Yet understanding of how form-specific clues relate to general linguistic properties is still lacking. Approaches aligning a text's linguistic complexity with a learner's general language proficiency have so far been limited to the domain of readability assessment (Chen and Meurers, 2019). In order to apply it to adaptive exercise selection, the relationship between an exercise's co-text complexity and the learner's language proficiency level must have an impact on the learner's performance on an exercise. If the relevance of a relationship between these two factors can be established, it constitutes a valuable indicator to determine initial parameter settings while the system lacks learner data for more individualized adaptation.

Approaches trying to determine difficulty based on exercise parameters, thus allowing to calibrate exercise difficulty without available learner performance data in order to solve the cold start problem, have indeed found that general language parameters influence exercise difficulty (Pandarova et al., 2019). However, these approaches focus on a specific exercise type each. Since different exercise types elicit different processing of the linguistic co-material and target different skills (Grellet, 1981, p. 5), the relevance of individual linguistic parameters can be expected to vary from one exercise type to the other.

The cold-start problem is not only an issue with new exercises, but also with learners interacting with the system for the first time or starting to practice a new learning target. If the learner has already completed other lessons, overall performance data might be used to determine initial exercise difficulty. Performance metrics for one particular learning target might, however, not be indicative of performance on another learning target. If the learner is new to the system, determining the appropriate exercise difficulty level becomes a matter of randomness. Many systems rely on user questionnaires asking about the proficiency level and in addition offer placement tests (Vesselinov and Grego, 2016). While specifically testing a learner's proficiency in the learning targets of the particular learning unit would provide the most representative picture of a learner's knowledge state, this could turn the first contact with the system into a frustrating experience for low-proficient learners. In addition, linguistic co-text material of exercises always contains linguistic constructions other than the learning targets. In order to process the semantic context of the exercises, learners need to have passive knowledge of of these constructions. Since text readability is traditionally linked to general language proficiency (Chen and Meurers, 2019), a measure reflecting this learner characteristic in relation to the complexity of the exercises' linguistic co-material might be more suitable to determine the optimal initial exercise difficulty. C-tests constitute a popular method of providing such a measure (Drackert and Timukova, 2020).

In this paper, we establish the groundwork to overcome the shortcomings of previous work on exercise difficulty calibration in terms of narrow exercise type coverage and learner-dependence of global exercise parameters. We determine for a range of different exercise types whether the global parameter of co-text complexity impacts learners' performance on the exercise. This will inform macro-adaptive algorithms as to which exercises warrant adaptive assignment with respect to co-text complexity. In addition, we analyze the relevance of the learner's proficiency to this parameter in order to determine whether co-text complexity has a similar impact on exercise difficulty for all learners.

The rest of the paper is structured as follows: Section 2 presents work on exercise difficulty calibration in the domain of language learning. Section 3 describes the dataset used for the evaluations. Section 4 presents the analyses and their results before discussing their implications for adaptive exercise selection. Section 5 concludes with a summary, including a discussion of some limitations of the approach and directions for future research.

2 Related Work

Macro-adaptive systems aim to provide personalized learning experiences by selecting exercises matching a learner's abilities (Slavuj et al., 2017). This has been tackled by a variety of approaches including the proportion of correct answers, Item Response Theory (IRT), Elo rating, and learner and expert ratings (Wauters et al., 2012). Human rating based approaches are subjective in nature and require human effort. Data based approaches are more objective, yet they rely on large amounts of learner answers in order to provide reliable difficulty estimates. Aiming to overcome this shortcoming, multiple strategies have been explored to determine exercise difficulty based on a range of exercise parameters instead. Hartig et al. (2012) point out that the relevant parameters vary depending on the skill targeted by the exercise so that the set of parameters needs to be determined individually for any domain. For language exercises, most work so far has focused on Cloze exercises with a particular emphasis on C-tests. In an early approach, Wilson (1994) used co-text readability as a single determining feature of exercise difficulty, acknowledging the need to yet establish its correlation with exercise difficulty. Others have identified a range of linguistic features on the word, sentence, and text levels that impact exercise difficulty (e.g. Galasso, 2018; Beinborn et al., 2014; McCarthy et al., 2021; Settles et al., 2020; Brown, 1989). The effect of exercise format parameters such as gap size, deletion pattern and deletion frequency on exercise difficulty varied across studies (Sigott, 1995; Lee et al., 2019; Kamimoto, 1993). Abraham and Chapelle (1992) explored different input types and found dropdown selection to be easier than text input. A number of Single Choice (SC) reading comprehension exercises applied machine learning and statistical approaches generating predictors of exercise difficulty from the text, the question, and answer options (Liu et al., 2021a; Huang et al., 2017; Loukina et al., 2016). While Holzknecht et al. (2021)

found that such exercises were more difficult when the correct option was in the last position, studies on SC exercises in other domains found exercises with the correct option in the first or last position (Attali and Bar-Hillel, 2003), or next to the most attractive distractor (Shin et al., 2020) to be harder. Also not focusing on language exercises, Swanson et al. (2006) explored the number of distractors, and Kubinger and Gottschall (2007) the number of correct options as indicators of exercise difficulty. Since language exercises are often automatically generated, their complexity is sometimes already determined and controlled for at generation time (Kurdi et al., 2020). In this line of work, Pilán et al. (2017) only considered the co-text complexity of their SC exercises for vocabulary practice. Generating the same type of exercises, Susanti et al. (2017) in addition used semantic similarity between the correct option and the distractors, as well as the word-level complexities of the distractors. In their comparisons of syntactically, paradigmatically and not related distractors, Hoshino (2013) found that syntactically related ones were the most difficult distractors, yet only in exercises that require semantic parsing of the co-text. Very little research has focused on grammar exercises. A noticeable exception constitutes the approach by Pandarova et al. (2019), which examines the effect on exercise difficulty of various linguistic properties on the gap, item, and text levels of Fill-in-the-Blanks (FiB) exercises to practice tenses.

Almost all of these analyses targeting difficulty parameters of language exercises use co-text complexity as one of the influencing features. However, they all consider only a single exercise type. In order to fill this gap and establish whether the results of such narrowly focused studies can be generalized to other exercise types, we present an evaluation of the impact of co-text complexity on exercise difficulty for seven exercise types.

Using a feature to predict static exercise difficulty only makes sense if the impact of the feature is similar for all learners. To the best of our knowledge, none of the approaches to exercise difficulty calibration have looked into learner dependence of the features impacting exercise difficulty. We therefore evaluate whether co-text complexity can be used as a static exercise complexity feature or whether it needs to be considered dynamically based on learner characteristics.

3 Data

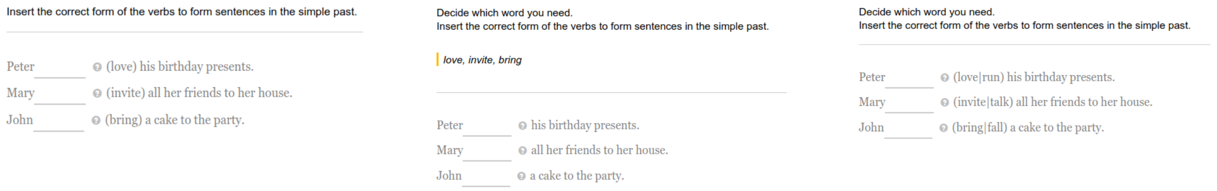
The evaluations are based on data obtained in the context of the Interact4School (I4S) (Parrisius et al., 2022a,b) and the Digbindiff[1] projects. Both studies collected data from 7th grade learners of English in German secondary schools who worked with the Intelligent Language Tutoring System (ILTS) FeedBook over the course of one school year. The system offers practice exercises with intelligent feedback provided to the learners as they work on the exercises. The two versions of the FeedBook used in the studies differ slightly from one another. While the focus in the I4S study was on motivational aspects in a task based setting, the Didi project looked into user-adaptive exercise sequencing.

The exercises in the I4S version of the FeedBook are organized into task-based cycles that each contain multiple linguistically and pedagogically motivated learning targets. The Didi study, on the other hand, groups exercises only according to learning targets. In order to use a common terminology for both projects, we use *chapter* to denote cycles of I4S and learning targets of Didi, and *learning target* when referring to the learning targets of both system versions.

In addition to the submissions of learners to the practice exercises, both studies also collected performance data on C-tests. These were conducted once at the beginning and once at the end of the studies, thus framing the practice exercises. The C-tests used at both test timepoints and in both studies are identical and consist of six parts. Of the 1,360 learners consenting to participate in the studies, 1,102 completed the first and 774 the second C-test. 553 learners completed both C-tests.

The practice exercise types in the systems include FiB, Short Answer (SA), SC, Jumbled Sentences (JS), Mark-the-Words (MtW), Categorization, and Memory exercises. The 201 exercises in the I4S study – excluding listening exercises – attempted by at least one learner were submitted by a mean of 136.13 learners ($\sigma = 112.58$). They are grouped into four chapters and 9 learning targets and contain a total of 1,140 actionable elements. An actionable element can be the blank of a FiB or SC exercise, a sentence of a JS exercise, a clickable chunk in a MtW exercise, an element to sort in a Categorization exercise, a Memory pair, or an answer to a SA exercise. In the Didi study, a mean

[1] http://digbindiff.de



Insert the correct form of the verbs to form sentences in the simple past.

Peter ______ (love) his birthday presents.
Mary ______ (invite) all her friends to her house.
John ______ (bring) a cake to the party.

(a) Lemma in parentheses

Decide which word you need.
Insert the correct form of the verbs to form sentences in the simple past.

love, invite, bring

Peter ______ his birthday presents.
Mary ______ all her friends to her house.
John ______ a cake to the party.

(b) Lemmas as bag of words

Decide which word you need.
Insert the correct form of the verbs to form sentences in the simple past.

Peter ______ (love|run) his birthday presents.
Mary ______ (invite|talk) all her friends to her house.
John ______ (bring|fall) a cake to the party.

(c) Distractor lemma in parentheses

Figure 1: Codings of FiB exercises

of 29.19 learners ($\sigma = 46.00$) attempted each of the 470 exercises with overall 2,003 actionable elements. These numbers differ considerably from those of the I4S study as the macro-adaptive focus of the Didi study resulted in a more varied practice environment adapted to the individual learner. The exercises are grouped into 4 chapters and learning targets. There is no overlap of learners or practice exercises between the two studies.

All data on exercises and learner submissions is stored in a PostgreSQL[2] database and managed through Hibernate[3].

4 Evaluation

We conducted a range of experiments to determine the relevance and learner dependence of co-text complexity for macro-adaptivity. For these analyses, the data was extracted from the databases with utility scripts written in Java which use the Hibernate setup to access the data. For further processing, the extracted learner submission and exercise data was stored in CSV files. Apart from the correctness of each learner's answers to the actionable elements of exercises, meta-information including the associated learning target, the exercise type, the length of the actionable elements, and exercise type specific information was extracted such as the number of chunks for JS or the number of distractors for SC exercises.

In addition to the metadata extracted from the databases, we determined IRT difficulty scores and co-text complexity scores for all exercises. IRT difficulty values b were determined for all actionable elements based on the Rasch model of the `TAM` package for R. Since the datasets of the two studies constitute discrete sets with no overlaps in learners or exercises, we determined the difficulty values independently for each dataset. For performance reasons, the data in addition needed to be split by learning targets. In order to determine co-text complexity of the exercises in the dataset, we extracted the text material from all exercises. This includes prompts as well as all actionable elements and surrounding co-text, but not instructions or any support texts. We approximated co-text complexity for all extracted texts through a number of different readability formulas. In lack of gold standard values for text complexity, we operationalized it as the mean value of normalized[4] readability scores obtained from various readability formulas. Although IRT scores were estimated separately for the learning targets, we used the joint dataset for the readability score determination as text complexity should be independent of exercises and learners.

Since we assumed that the effect of co-text complexity might only be relevant to some learning targets and to some exercise types, we extracted subsets of exercises for isolated analyses. Each combination of exercise type and learning target resulted in a distinct subset of exercises. In addition, FiB exercises support two possible codings, as illustrated in Figure 1: (1) Specifying the required lemma in parentheses behind the blank (1a) results in mechanical drill exercises. (2) Giving the lemmas as bags of words for the entire exercise (1b) or providing an additional distractor lemma in parentheses (1c) requires top-down skills in the form of parsing the co-text (Nagao, 2002) in order to successfully answer the exercise. Considering that co-text complexity might be less relevant in exercises where correct processing of the text is not essential (Hoshino, 2013), we extracted the co-text sensitive exercises into an additional subset. Some data might not be representative due to the low number of submissions for an exercise. A further subset of core exercises therefore is based

[2] http://postgresql.org

[3] http://hibernate.org

[4] We used the `StandardScaler` of the Python scikit-learn package for scaling of the readability scores of each formula, and the `MinMaxScaler` of the same library to scale the mean readability scores into the range [0,1].

on the number of learner submissions for the exercises. It encompasses all exercises which were submitted by at least 50% of all learners in the respective study. The next three subsets control for exercise difficulty. They consist of exercise items with similar IRT difficulties in the low, intermediate, and high difficulty ranges. Since IRT scores were determined for individual actionable elements instead of for entire exercises, these subsets contain actionable elements as items. In order to maximize the number of items per subset while minimizing the range of difficulty scores, in the intermediate difficulty subset we only included exercises that deviate from the median value in no more than 1%. For the low and high difficulty subsets, we used the same number of exercise items with the lowest and highest difficulty scores respectively. The last three subsets, created in a similar manner based on the scores of the first C-test, control for learner proficiency. They contain only the submission data for exercises attempted by the learners associated with the respective proficiency group.

After thus pre-processing the raw database data into a format independent of the ILTS and enriched with meta-information, we implemented the analyses in Python and R.

4.1 Relationship between C-test and practice performance

C-tests are widely used to assess general language proficiency and have been established to reliably and validly do so (Klein-Braley, 1996). However, more recent critical evaluations show mixed results, ranging from high (e.g. Lei, 2008; Rasoli, 2021) to very low (e.g. Farhady and Jamali, 2006; Mashad, 2008) validity for English. These discrepancies might stem from differences in the participants as Mashad (2008) found C-tests to only be reliable for certain proficiency groups. In order to determine the suitability of determining general language proficiency through C-tests for our target group, we determined the distributions of the C-test scores based on histogram plots. Although Daller and Phelan (2006) point out that C-tests are not necessarily normally distributed, we expect similar distributions for all C-test parts. As a reference point, we determined the overall distribution of C-test scores for both C-tests of the dataset, which was found to have a curved shape. Figure 2 shows that out of the six parts of each C-test, only the second, third and fourth parts reflect this form while the other three parts have monotonically increasing distributions. The meta information available for the C-tests confirms that these parts do indeed not provide representative data: The first part constitutes an example item. The last two parts were attempted by only a small number of learners who managed to complete them within the given time frame, thus presumably being more proficient than the slower learners. In the subsequent evaluations, we therefore only used the results of the second to fourth parts.

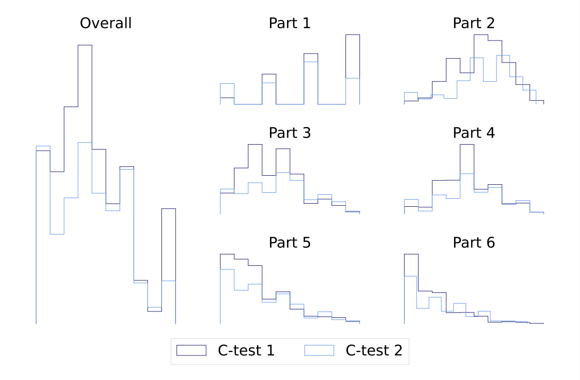


Figure 2: Distributions of C-test scores

The tests can only be indicative of varying performance on exercises if performance on the C-tests is varied across learners. In order to verify that our dataset covers learners of diverse proficiency levels, we determined the range of accuracies obtained on the C-tests. The values are similar for both C-tests with minimum scores of .00 and the highest observed accuracy at .62. When excluding the learners who did not correctly answer any item ($acc = .00$), the lowest score amounts to .01. The study participants thus indeed comprise learners of very low English proficiency who nevertheless made an effort to complete the C-tests. The dataset therefore covers learners with overall English language proficiencies ranging from very weak to moderately strong.

Since we aim to match text complexity to learner proficiency, the scores obtained for both parameters should be equally distributed across exercise texts and learners. We therefore compared the histograms representing the distribution of the text readability scores with that of the overall C-test scores per C-test. Figure 3 illustrates that the curve-shaped distribution of the C-test scores, even more pronounced when excluding the invalid

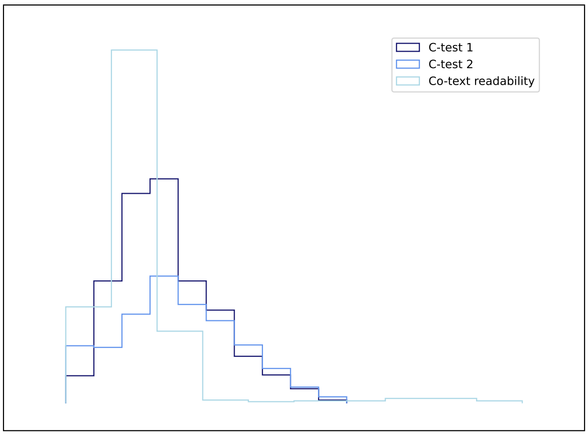


Figure 3: Distributions of C-test and readability scores

parts, is reflected in the histogram for text readability scores. Our dataset thus represents learners and exercises whose global language proficiency, operationalized as C-test scores, and co-text complexity, operationalized as text readability scores, respectively, have compatible distributions.

After establishing the validity of the C-tests in themselves as well as the possibility to map the scores to co-text complexity, we can effectively use them to operationalize a learner's general language proficiency. This learner characteristic can only impact exercise difficulty if there is any relationship between the operationalizations of both. In order to determine whether this is the case for our dataset, we calculated Pearson's correlation ρ between the learners' performance on the C-tests and that on practice exercises. C-test performance was defined as the accuracy on all items of the valid C-tests. Practice performance was defined as the accuracy on the actionable elements of all practice exercises. In addition to global correlation, we also looked at the correlations within the subsets representing combinations of exercise types and learning targets. This allowed us to determine whether C-test performance impacts exercise difficulty for only certain exercise types or learning targets. Table 1 gives an overview of the results. For the first C-test, the Pearson correlation reveals only a weak relationship between C-test accuracy and practice accuracy ($\rho = .28$). It does not increase when only considering core exercises ($\rho = .28$), and only marginally for co-text sensitive exercises ($\rho = .29$). This suggests that the data for the overall exercise pool reflects the picture of the subset most representative of our target group and that general language proficiency is not more relevant for exercises that require processing of the text material. When controlling for exercise difficulty, the relationship is even less pronounced with a weak correlation of $\rho = .27$ for intermediate-difficulty exercises and no relationship for low- ($\rho = .18$) and high-difficulty exercises ($\rho = .15$). When looking at the different learning targets and exercise types separately, correlations are higher for a number of sub-groups covering almost all exercise types and learning targets. The highest – although weak – correlation ($\rho = .47$) is for FiB exercises on *Simple past vs. Present perfect*. The gap filling exercise types FiB and SC, as well as the occasional JS exercise type, have the highest correlations for a number of learning targets. Of these, there is no pattern indicating that any learning target generally has higher correlations between C-test and practice performance than others.

Exercise set	ρ_{c1}	ρ_{c2}
All	.2811	.4070
Core	.2821	.3641
Co-text sensitive	.2887	.3882
Low difficulty	.1773	.2356
Intermediate difficulty	.2674	.2763
High difficulty	.1536	.2465
FiB – *Simple past vs. Pres. perf.*	.4688	.3890
SC – *Conditionals*	.4101	.4392

Table 1: Pearson's correlations of C-test 1 (ρ_{c1}) and C-test 2 (ρ_{c2}) with practice performance

Interestingly, the scores of the second C-test correlate much better with the learners' practice performance, although the relationship is still weak ($\rho = .41$). When looking at the subsets, the pattern is similar to that with the first C-test: Core exercises ($\rho = .36$) and co-text sensitive exercises ($\rho = .38$) have comparable correlations. Correlations for low- ($\rho = .24$) and high-difficulty exercises ($\rho = .25$) are considerably lower again and exercises of intermediate difficulty correlate slightly better with the C-test scores ($\rho = .28$) than the other two subsets, although much less relative to the overall exercise set than for the first C-test. The highest ranked combination of exercise type and learning target of the first C-test again shows a weak correlation ($\rho = .39$), and is only surpassed by one other combination. The correlation between performance on this C-test and practice performance is highest for SC exercises on *Conditionals* ($\rho = .44$). The patterns for specific exercise types and learning targets are similar to those for the first C-test. Since correlations are higher with the second than with the first C-test for all learning targets, the temporal proximity of the test to the practice session does not seem to be the cause of this observation.

In order to better compare the significance of the two C-tests with respect to their predictive power for practice performance, we generated a partial dependence plot based on an AdaBoost classifier trained to predict whether an actionable element is answered correctly depending on the C-test scores. As the probability increases, the colouring turns from purple to green. For the plot given in Figure 4, the colour changes progressively on the vertical axis representing the second C-test, but not on the horizontal axis representing the first C-test. This illustrates that while for the second C-test, the probability of a learner answering an element correctly increases with higher test scores, this is not the case for the first C-test.

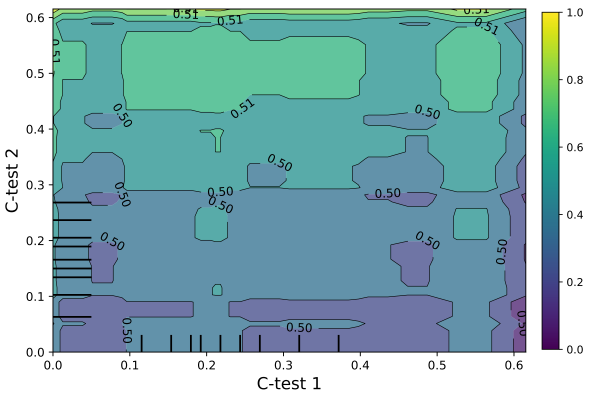


Figure 4: Partial dependence plot for the C-tests when predicting the correctness of a learner's answer

The approach to match co-text complexity to a learner's global language proficiency in order to improve the learner's performance on practice exercises requires valid indicators of learner proficiency from which to calculate the match. As a learner's general language proficiency may change during their involvement with the system, the validity of the initially elicited proficiency score might decrease over time. In order to determine whether this is the case for our learner population, we trained an AdaBoost classifier[5] individually for each of the four chapters to predict a learner's performance on an exercise from the C-test scores and co-text complexity. Since the chapter index represents the exercises' relative practice timepoint, the development of the feature importances of the two C-tests relative to each other over the sequence of succeeding chapters can give insights into whether recency of a C-test influences the predictive power of general language proficiency. While the classifier's feature rankings – outlined in Table 2 for the entire dataset – indicate varying priority of one of the two C-tests over the other, a C-test's importance does not monotonically increase with its temporal proximity to the practice unit. This is similar for all data subsets as illustrated in Figure 5, which displays the difference in feature importances between the first and second C-test depending on the chapter. Monotonically decreasing lines would indicate that the first C-test loses importance with later chapters while the second C-test's importance increases. However, this is not the case for any of the subsets. The test timepoint therefore does not seem to play a substantial role in the predictive power of C-tests.

	c1	**c2**	**c1-c2**	**Relative impact**
Chapter 1	.16	.12	.04	1 > 2
Chapter 2	.04	.10	-.06	2 > 1
Chapter 3	.02	.08	-.06	2 > 1
Chapter 4	.14	.10	.04	1 > 2

Table 2: Feature importances of the first (c1) and second (c2) C-tests

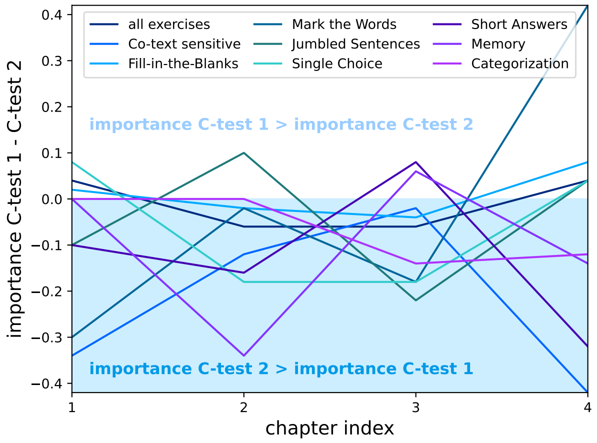


Figure 5: Importance of the C-test scores relative to each other over succeeding chapters

When looking at the development of the learners' C-test scores from one test timepoint to the

[5] The classification was based on the `scikit-learn` (https://scikit-learn.org) implementation for Python.

other, the scatter plot given in Figure 6 reveals that for a considerable number of learners, represented in the shaded area underneath the first bisector, the scores do not show the expected increase, but decrease over time. This also results in an only moderate correlation ($\rho = .5260$) between the two tests. Considering the previous findings that the scores of the second C-test correlate better with practice performance than those of the first C-test, this could indicate that C-tests taken during a learner's first interaction with the system are not entirely representative of their general language proficiency, possibly due to the novelty of the system and the test setup. A tentative conclusion assumes that C-tests do not lose validity over time, at least not within the course of a school year, but that tests are more representative if learners are already familiar with the test platform.

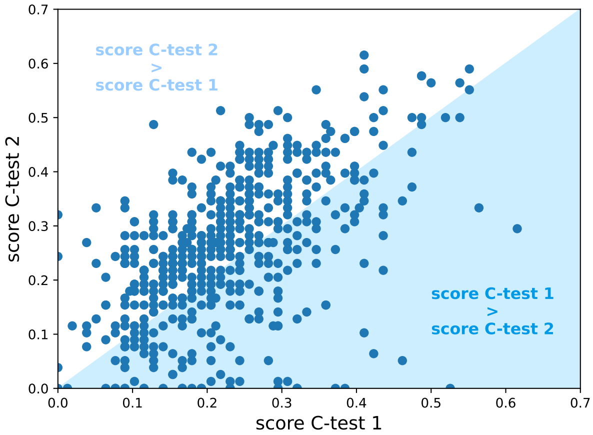


Figure 6: Development of C-test scores between test timepoints

Overall, these results indicate that C-test scores have no or only weak linear relationships with performance on exercises. Although correlations are generally higher for FiB exercises, this is not the case for the co-text sensitive exercises even though they constitute a subset of FiB exercises. Especially for low- and high-difficulty exercises, the relationship of general language proficiency with practice performance, if there is one, does not seem to be linear. C-tests are, however, more predictive of a learner's performance on practice exercises when taken after a period of familiarization with the system.

4.2 Linear relationship between co-text complexity and exercise difficulty

If exercise difficulty increases linearly with increasing co-text complexity, there should be a positive correlation between these two variables. We therefore determined Pearson's correlation between the readability scores and the IRT difficulty scores. Since there might not be a global relationship for all exercise types and learning targets, we calculated correlations for the various subsets in addition to the correlation for the entire dataset.

Exercise set	ρ	Sample size
All	.0991	3,104
I4S	.0076	1,101
Didi	.1381	2,003
Future Tenses	-.0094	127
Modals	.7270	34
FiB	.0022	1,849
JS	.3337	444
FiB – *Simple past vs. Present perfect*	-.0231	241
SC – *Conditionals*	.8291	8
Core	.0024	131
Co-text sensitive	.0804	208

Table 3: Pearson's correlation ρ of text readability with exercise difficulty

The results, summarized in Table 3, show that there is no linear relationship between co-text readability and exercise difficulty either for all exercises ($|\rho| = .10$) or for those of the individual I4S ($|\rho| = .01$) and Didi ($\rho = .14$) studies. The values vary considerably between learning targets ($|\rho| = .01$ for *Future Tenses* to $|\rho| = .73$ for *Modals*) and exercise types ($|\rho| = .00$ for FiB to $|\rho| = .33$ for JS). For the subsets comprising combinations of learning targets and exercise types, this variance is equally high ($|\rho| = .02$ for FiB exercises on *Simple past vs. Present perfect* to $|\rho| = .83$ for SC exercises on *Conditionals*[6]). There is no relationship for the subsets containing only core exercises ($|\rho| = .00$) or only co-text sensitive exercises ($|\rho| = .08$). Interestingly, some correlations are negative, suggesting that exercises are more difficult when co-text complexity is lower. While this might be due to insufficiently large sample sizes, it could also indicate

[6] We excluded those combinations with sample sizes of 2, although sample sizes may be too small in most other cases as well (4 - 385) to yield reliable results.

that exercise creators try to compensate some difficulty features with others in order to create exercises of overall approximately similar difficulties. The results, while not entirely conclusive due to data sparseness considering the multitude of parameters influencing exercise difficulty, indicate that co-text complexity does not have the same effect on exercise difficulty for all learning targets and exercise types. There is no overall linear relationship between these two parameters.

For the subsets controlling for exercise difficulty, the difficulty values differ only marginally by definition. We therefore determined the mean as well as the minimum and maximum readability scores within these subsets and compared them between the sets. Following the logic that higher readability scores result in higher exercise difficulties, these metrics should then be lowest for the subset of low-difficulty exercises and highest for the subset of high-difficulty exercises. However, the boxplots in Figure 7 illustrate that readability scores are very similar for all three subsets, with values ranging from .0000 to .4632 ($\mu = .1390$), from .0172 to .3841 ($\mu = .1503$), and from .0074 to 1.0 ($\mu = .1776$) for low-, intermediate-, and high-difficulty items respectively. It should be noted, though, that very high readability scores appear only with high-difficulty exercises, which could indicate that such high text complexities might indeed have an influence on overall exercise difficulty.

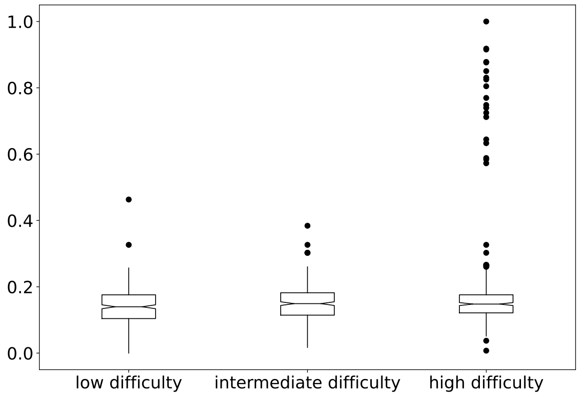


Figure 7: Boxplots of readability score distributions for difficulty controlled subsets

4.3 Non-linear relationships between co-text complexity and exercise difficulty

In order to capture non-linear relationships between co-text complexity and exercise difficulty, we trained various classifiers to predict whether a learner answers an actionable element correctly. The classifiers include a Decision Tree, a Random Forest, and an AdaBoost classifier from the Python `scikit-learn`[7] library, which all provide predictor rankings. As baseline model, we used only simple exercise features such as the exercise type, the number of tokens in the target answer, and the number of other targets in the exercise. We then analyzed a range of model variants for various subsets of the data and with different combinations of additional features targeting IRT difficulty, text readability, and C-test scores. While IRT difficulty scores can be expected to be the most indicative exercise parameter in terms of practice performance, this feature is unknown for new exercises. We therefore analyzed models both with and without the IRT difficulty predictor. All features were encoded as Integer values; not applicable features received the value *zero*. We determined precision, recall, and F1 scores as performance metrics for all model variants in order to evaluate whether adding certain features improves model performance. Precision, recall and F1 scores are comparable for all three classifiers, although the AdaBoost classifier slightly outperforms the others in most experiment settings. For the entire dataset, precision and recall are almost always identical and mirror the F1 scores. We therefore report only F1 scores of the AdaBoost classifier, which are summarized in Table 4. The baseline model already achieves a high F1 score of .72 which increases to .76 when adding the IRT difficulty predictor. When only using text complexity as additional feature, there is almost no increase in performance ($F1 = .72$) as compared to the baseline model. Adding the C-test scores to any of the experiment settings results in a slight increase in F1 scores. Although the best performing model ($F1 = .77$) incorporates all predictors, multiple models with a reduced feature set perform nearly as well. They all include the IRT difficulties as well as C-test scores. The two C-tests result in comparable model performances. The model using all features except for IRT difficulty achieves a F1 score of .73, which constitutes the best performance without IRT difficulties. Adding text complexity as a feature to the best performing models has a small positive effect on performance. F1 scores are gen-

[7] https://scikit-learn.org

Predictors Set of exercises	base-line	+b	+co-text	+b+c1	+co-text +b+c1	+b+c2	+co-text +b+c2	+co-text +c1+c2	all	μ	σ
All	.7238	.7599	.7247	.7655	.7661	.7653	.7664	.7251	.7669	.7515	.0203
Core	.7510	.7612	.7508	.7709	.7784	.7779	.7751	.7630	.7798	.7676	.0115
Co-text sens.	.7070	.7393	.7108	.7431	.7407	.7491	.7505	.7108	.7516	.7337	.0186
$b_{intermediate}$	.7374	.7458	.7408	.7437	.7404	.7429	.7420	.7424	.7437	.7421	.0024
b_{low}	.9450	.9450	.9450	.9467	.9467	.9470	.9470	.9470	.9469	.9463	.0009
b_{high}	.8553	.8538	.8553	.8516	.8516	.8524	.8524	.8545	.8487	.8528	.0021
FiB	.7227	.7750	.7214	.7760	.7755	.7775	.7777	.7197	.7767	.7580	.0276
MtW	.6585	.6711	.6656	.6985	.6913	.7064	.7082	.7146	.7093	.6915	.0211
JS	.8350	.8549	.8362	.8531	.8538	.8587	.8527	.8343	.8549	.8482	.0099
SC	.7760	.7820	.7760	.7844	.7830	.7848	.7855	.7741	.7855	.7813	.0046
SA	.7277	.7652	.7256	.7562	.7546	.7578	.7620	.7361	.7657	.7501	.0159
Memory	.9535	.9535	.9535	.9535	.9535	.9535	.9535	.9581	.9628	.9550	.0033
Categorization	.6949	.6949	.6949	.7190	.7190	.6949	.6979	.7160	.7009	.7036	.0110

Table 4: Classifier performance

erally slightly higher for the subsets of core exercises ($\mu_{F1} = .77, \sigma_{F1} = .01$) and exercises of intermediate difficulty ($\mu_{F1} = .74, \sigma_{F1} = .00$), and marginally lower for co-text sensitive exercises ($\mu_{F1} = .73, \sigma_{F1} = .02$). For high-difficulty exercises, they are considerably higher ($\mu_{F1} = .85, \sigma_{F1} = .00$) and even more so for low-difficulty exercises ($\mu_{F1} = .95, \sigma_{F1} = .00$). The standard deviations show that there are almost no differences in F1 scores between the model variants of exercise sets with controlled difficulty, which highlights the high relevance of the IRT difficulty feature once again.

In addition, we analyzed the feature importances provided by the classifiers, which allow to estimate the relevance of the individual features to the models' predictions. While model performance metrics indicate that co-text complexity has only little impact on a learner's performance on exercises, the feature rankings, illustrated in the heatmaps in Figure 8, show that this parameter holds substantial predictive power. Not surprisingly, exercise difficulty is the overall most predictive feature. It is, however, followed by co-text complexity in most models integrating this feature and ranked highest in models not including IRT difficulty. The feature rankings for the analyzed features – IRT difficulty, text readability and C-test scores – are similar for all subsets of exercises in terms of relative rankings, although absolute values vary. Differences in the rankings concern mostly the simple exercise features and are quite pronounced between the different exercise types. However, co-text complexity also features greater importance for FiB, and most particularly co-text sensitive exercises, SC, and JS exercises compared to the other exercise types. This on the one hand supports the findings of Section 4.2 in terms of exercise types for which co-text plays a role, and on the other hand reveals that it is particularly relevant with co-text sensitive exercises after all. In addition, the relevance of C-test scores varies considerably from one exercise type to the other. According to the predictor rankings, general language proficiency is highly relevant – even more relevant than IRT difficulty – with Memory and Categorization exercises, and less so with JS, SC, SA, MtW, and particularly FiB exercises.

Overall, the classification experiments reveal that co-text complexity does have predictive power with respect to a learner's performance on an exercise.

4.4 Learner dependence of co-text complexity predictiveness

By comparing the performance of classifiers for the subsets of controlled learner proficiency using co-text complexity as a single predictor, we aimed to determine whether co-text complexity is a learner dependent or independent parameter. If the predictive power of co-text complexity varies with the learners' proficiency levels, we expect performance to differ between the subsets. The results indeed show differences in model performance, which is best for high learner proficiency ($F1 = .7755$) and lowest for low proficiency ($F1 = .6627$). Co-text complexity is therefore a good predictor of practice performance for high-proficiency learners, but less so for low-proficiency learners. This could indicate that less proficient learners do not process an exercise's co-text, either because they do not attempt to do so or

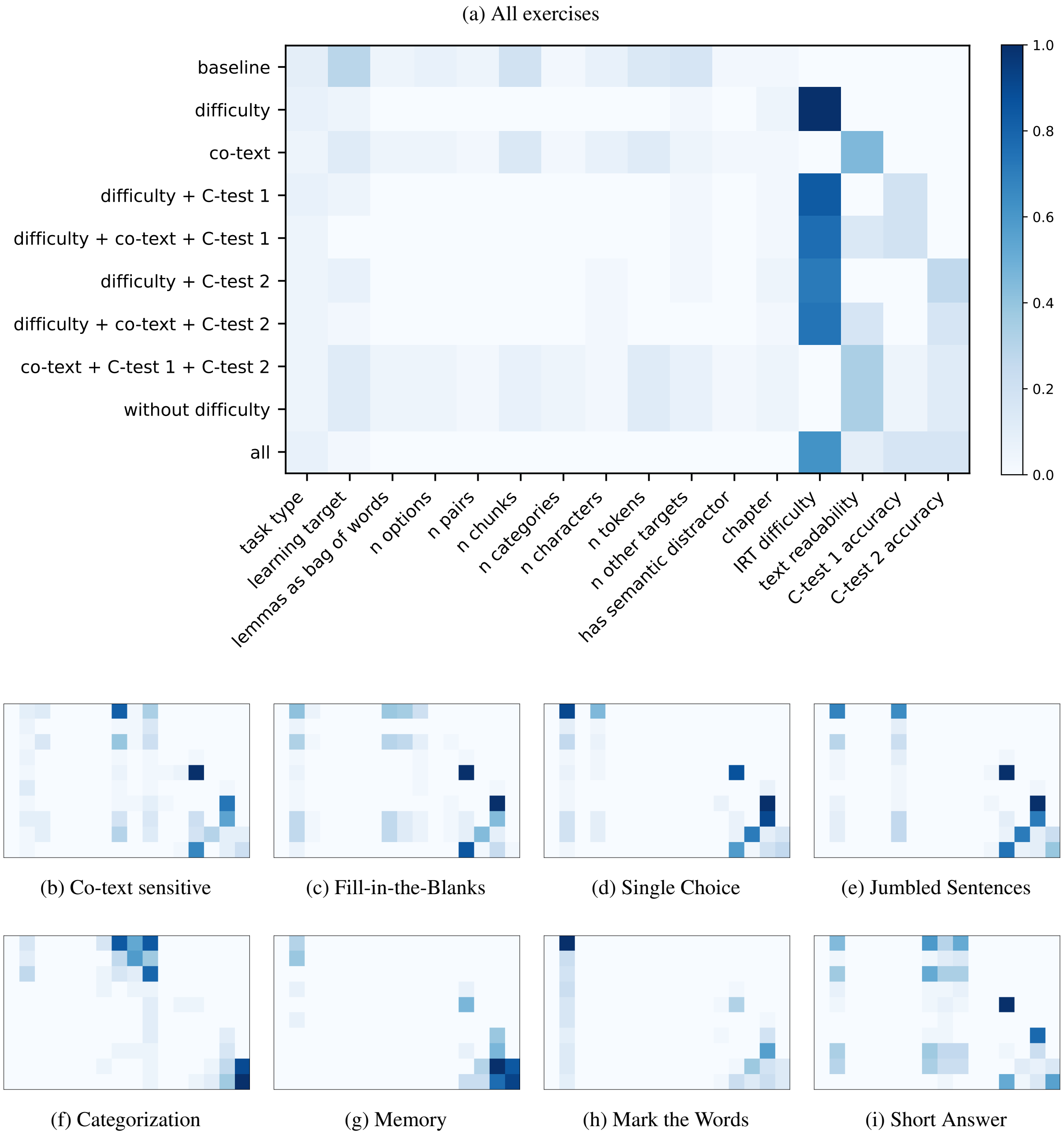


Figure 8: Feature importances

because even the easier texts are too challenging for them, so that this parameter has less impact on their practice performance. Co-text complexity thus seems to be a learner dependent parameter which holds more predictive power the higher the learner's proficiency.

5 Conclusion

We presented an extensive evaluation of the relevance of co-text complexity to exercise difficulty and its dependence on an individual learner's global language proficiency. The analyses cover seven exercise types that differ in the relevance of understanding the co-text in order to successfully answer them. We showed that while there is generally no linear relationship between co-text complexity and a learner's performance on the exercise, statistical models can capture the predictive power of this parameter in combination with other exercise and learner specific features. This is especially true for exercises going beyond mechanical drills, where the co-text provides guidance to successfully answer the exercise. However, its predictive power varies with a learner's proficiency. More proficient learners seem to make use of top-down skills, while less proficient learners use more local clues to solve grammar exercises. Co-text complexity should therefore be considered as a dynamic parameter in adaptive exercise selection in conjunction with a learner's general language proficiency.

We also acknowledge some limitations to our evaluations. Although the C-test scores cover a considerable range, our learners might still constitute a more homogeneous group than in other ILTS where learners do not follow the same curriculum and workbook. Similarly, since the exercises were created from manually composed texts, they do not represent the variability found in authentic texts, especially concerning higher complexities. In addition, readability formulas constitute easy-to-use measures of linguistic complexity thanks to their numerical output scores. However, they do not cover the entire spectrum of linguistic properties relevant to complexity which can be considered in more sophisticated approaches. These should also differentiate between different scopes of the features since for some exercises it might be sufficient to consider the linguistic constructs in the sentence of the actionable element instead of in the entire exercise's co-text.

Future work will need to determine the threshold defining high general language proficiency so that co-text complexity can be considered exclusively for those learners for whom it does make a difference.

References

Roberta G. Abraham and Carol A. Chapelle. 1992. The Meaning of Cloze Test Scores: An Item Difficulty Perspective. *Modern Language Journal*, 76(4):468–479.

Yigal Attali and Maya Bar-Hillel. 2003. Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable. *Journal of Educational Measurement*, 40:109–128.

Lisa Beinborn. 2016. *Predicting and Manipulating the Difficulty of Text-Completion Exercises for Language Learning*. Ph.D. thesis, Technische Universität Darmstadt.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the Difficulty of Language Proficiency Tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.

James Dean Brown. 1989. Cloze item difficulty. *JALT journal*, 11(1):46–67.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4):418–447.

Helmut Daller and David Phelan. 2006. The C-test and TOEIC ® as measures of students' progress in intensive short courses in EFL. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 13(2):101–119.

Anastasia Drackert and Anna Timukova. 2020. What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing*, 37(1):107–132.

Hossein Farhady and Ferdos Jamali. 2006. Varieties of C-test as measures of general language proficiency. In Hossein Farhady, editor, *Twenty-five years of living with applied linguistics: collection of articles*, pages 287–302. Rahnama Press.

Sabrina Galasso. 2018. Automated C-test difficulty prediction: Integrating lexical, sentence, and text features in a multi-lingual perspective. Master's thesis, University of Tübingen, Tübingen.

Françoise Grellet. 1981. *Developing Reading Skills: A Practical Guide to Reading Comprehension Exercises*. Cambridge Language Teaching Library. Cambridge University Press, New York.

Johannes Hartig, Andreas Frey, Gúnter Nold, and Eckhard Klieme. 2012. An Application of Explanatory Item Response Modeling for Model-Based Proficiency Scaling. *Educational and Psychological Measurement*, 72(4):665–686.

Franz Holzknecht, Gareth McCray, Kathrin Eberharter, Benjamin Kremmel, Matthias Zehentner, Richard Spiby, and Jamie Dunlea. 2021. The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, 38(1):41–61.

Yuko Hoshino. 2013. Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia*, 3:16.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *AAAI Conference on Artificial Intelligence*.

Tadamitsu Kamimoto. 1993. Tailoring the Test to Fit the Students : Improvement of the C-Test through Classical Item Analysis. *Language Laboratory*, 30:47–61.

Christine Klein-Braley. 1996. *Towards a theory of C-Test processing*, pages 23–94. R. Grotjahn, Rüdiger.

Klaus D. Kubinger and Christian H. Gottschall. 2007. Item difficulty of multiple choice tests dependant on different item response formats - An experiment in fundamental research on psychological assessment. *Psychology Science*, 49.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Ji-Ung Lee, Erik Schwan, and Christian Meyer. 2019. Manipulating the Difficulty of C-Tests. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 360–370.

Lei Lei. 2008. Validation of the C-Test amongst Chinese ESL Learners. *Journal of Asia TEFL*, pages 117–140.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021a. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115.

Qi Liu, Shuanghong Shen, Zhenya Huang, Enhong Chen, and Yonghe Zheng. 2021b. A Survey of Knowledge Tracing. *Computing Research Repository*, abs/2105.15106.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253, Osaka, Japan. The COLING 2016 Organizing Committee.

Iran Mashad. 2008. Another look at the C-Test: A validation study with Iranian EFL learners. *The Asian EFL Journal*, 10(1):154.

Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-Starting Item Parameters for Adaptive Language Tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hirotaka Nagao. 2002. Using Top-Down Skills to Increase Reading Comprehension. Unpublished Report, ERIC Number ED475744.

Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcene Boubekki, Roger Jones, and Ulf Brefeld. 2019. Predicting the Difficulty of Exercise Items for Dynamic Difficulty Adaptation in Adaptive Language Tutoring. *International Journal of Artificial Intelligence in Education*.

Jung Yeon Park, Frederik Cornillie, Han L. J. van der Maas, and Wim Van Den Noortgate. 2019. A Multidimensional IRT Approach for Dynamically Monitoring Ability Growth in Computerized Practice Environments. *Frontiers in Psychology*, 10.

Cora Parrisius, Ines Pieronczyk, Carolyn Blume, Katharina Wendebourg, Diana Pili-Moss, Mirjam Assmann, Sabine Beilharz, Stephen Bodnar, Leona Colling, Heiko Holz, et al. 2022a. 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.

Cora Parrisius, Katharina Wendebourg, Sven Rieger, Ines Loll, Diana Pili-Moss, Leona Colling, Carolyn Blume, Ines Pieronczyk, Heiko Holz, Stephen Bodnar, et al. 2022b. Effective Features of Feedback in an Intelligent Tutoring System-A Randomized Controlled Field Trial (Pre-Registration).

Ildikó Pilán, Elena Volodina, and Lars Borin. 2017. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues*, 57.

Mohammad Kabir Rasoli. 2021. Validation of C-test Among Afghan Students of English as a foreign Language. *International Journal of Language Testing*, 11(2):109–121.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning-Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Jinnie Shin, Okan Bulut, and Mark J. Gierl. 2020. The Effect of the Most-Attractive-Distractor Location on Multiple-Choice Item Difficulty. *Journal of Experimental Education*, 88(4):643–659.

Günther Sigott. 1995. The C-Test: Some Factors of Difficulty. *Aaa-arbeiten Aus Anglistik Und Amerikanistik*, 20(1):43–53.

Vanja Slavuj, Ana Meštrović, and Božidar Kovačić. 2017. Adaptivity in educational systems for language learning: a review. *Computer Assisted Language Learning*, 30(1-2):64–90.

Yunik Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, 12.

David Swanson, Kathleen Holtzman, Krista Allbee, and Brian Clauser. 2006. Psychometric Characteristics and Response Times for Content-Parallel Extended-Matching and One-Best-Answer Items in Relation to Number of Options. *Academic Medicine*, 81:52–5.

Roumen Vesselinov and John Grego. 2016. The Babbel efficacy study. Babbel White Paper.

Joel Walz. 1989. Context and Contextualized Language Practice in Foreign Language Teaching. *Modern Language Journal*, 73(2):160–168.

Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.

Eve Wilson. 1994. A User-Adaptive Interface for Computer Assisted Language Learning. In *Proceedings of ED-MEDIA 84–World Conference on Educational Multimedia and Hypermedia*.

Wynne Wong and Bill Van Patten. 2003. The Evidence is IN: Drills are OUT. *Foreign Language Annals*, 36(3):403–423.