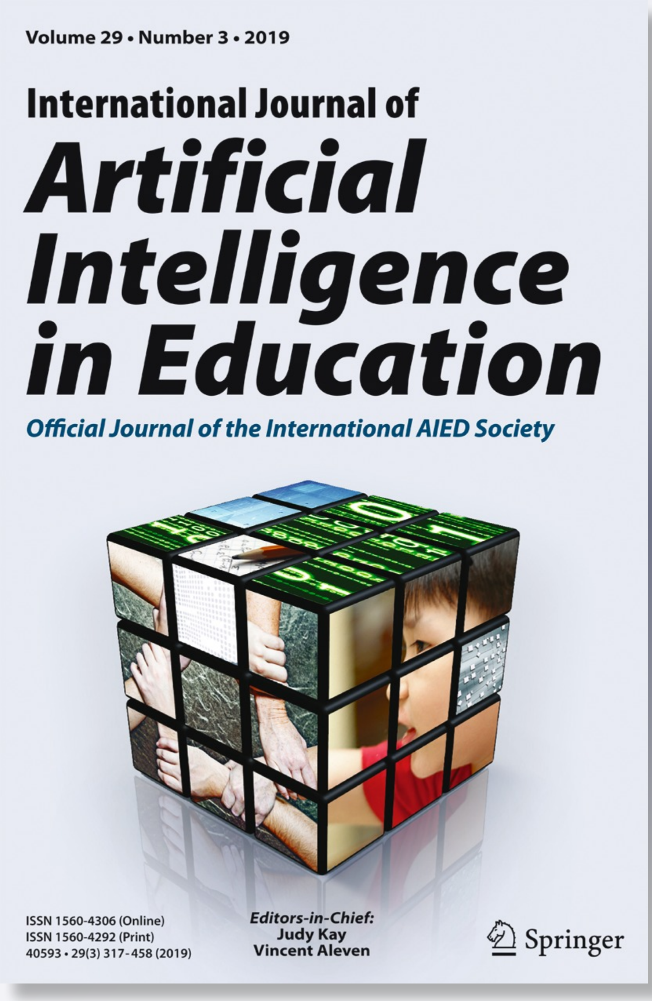
Predicting the Difficulty of Exercise Items for Dynamic Difficulty Adaptation in Adaptive Language Tutoring

Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcène Boubekki, Roger Dale Jones & Ulf Brefeld

International Journal of Artificial Intelligence in Education
Official Journal of the International AIED Society

ISSN 1560-4292
Volume 29
Number 3

Int J Artif Intell Educ (2019) 29:342-367
DOI 10.1007/s40593-019-00180-4



Springer

International Journal of Artificial Intelligence in Education (2019) 29:342–367
https://doi.org/10.1007/s40593-019-00180-4

ARTICLE

Predicting the Difficulty of Exercise Items for Dynamic Difficulty Adaptation in Adaptive Language Tutoring

Irina Pandarova[1] · Torben Schmidt[1] · Johannes Hartig[2] · Ahcène Boubekki[1] · Roger Dale Jones[1,3] · Ulf Brefeld[1]

Published online: 13 May 2019

Abstract
Advances in computer technology and artificial intelligence create opportunities for developing adaptive language learning technologies which are sensitive to individual learner characteristics. This paper focuses on one form of adaptivity in which the difficulty of learning content is dynamically adjusted to the learner's evolving language ability. A pilot study is presented which aims to advance the (semi-)automatic difficulty scoring of grammar exercise items to be used in dynamic difficulty adaptation in an intelligent language tutoring system for practicing English tenses. In it, methods from item response theory and machine learning are combined with linguistic item analysis in order to calibrate the difficulty of an initial exercise pool of cued gap-filling items (CGFIs) and isolate CGFI features predictive of item difficulty. Multiple item features at the gap, context and CGFI levels are tested and relevant predictors are identified at all three levels. Our pilot regression models reach encouraging prediction accuracy levels which could, pending additional validation, enable the dynamic selection of newly generated items ranging from moderately easy to moderately difficult. The paper highlights further applications of the proposed methodology in the area of adapting language tutoring, item design and second language acquisition, and sketches out issues for future research.

Keywords Adaptivity · Intelligent language tutoring systems · Item difficulty prediction · Item response theory · Machine learning · Second language acquisition

Introduction

Recently, there has been a growing interest in the development of digital technologies that offer adaptivity and personalization as a way of supporting and enhancing language learning (Kerr 2015). Among these technologies are

✉ Irina Pandarova
pandarova@leuphana.de

Extended author information available on the last page of the article intelligent language tutoring systems (ILTSs), which use artificial intelligence to capture and analyze learner data and make appropriate adjustments to the instructional process (Shute and Zapata-Rivera 2012; Slavuj et al. 2016). The language proficiency of the individual learner, including type and level of knowledge, skills and misconceptions, typically represents a central source of adaptation. Natural language processing techniques, for instance, make it increasingly possible to perform fine-grained error analysis of learner input and to deliver error-specific feedback and remedial activities (e.g. *E-Tutor* (Heift 2016) and *Tagarela* (Amaral and Meurers 2011)). The intelligent selection, sequencing and mode of presentation of learning content may also be based on long-term performance and can be influenced by further factors such as the learner's age, linguistic background, goals and styles, affective states, disabilities or indeed the learning context itself (Brusilovsky and Millán 2007; Slavuj et al. 2016). This paper focuses on a form of adaptive sequencing sensitive to the developing language ability of the individual learner and achieved by dynamically matching the difficulty of new or remedial content to the learner's current level of ability (Wauters et al. 2010). This process is henceforth referred to as dynamic difficulty adaptation (DDA) and the work reported here is primarily concerned with the development of learning materials, and especially exercise items, which lend themselves to DDA.

This paper presents a pilot study conducted in preparation of an ILTS for practicing English tenses – a grammatical area notoriously challenging for learners – which can be used as a complement to classroom instruction, as well as for collecting data on the development of L2 grammatical ability and researching the effectiveness of adaptive tutoring, including DDA. The pilot study has two main objectives regarding the development of the exercise item pool the ILTS will operate with. The first objective involves calibrating the difficulty of an initial tense exercise pool consisting of cued gap-filling items (CGFIs) using item response theory (IRT). Difficulty calibration is not only crucial for DDA but can also help assess the appropriateness of the initial item pool for the target population (currently 9th and 10th grade learners in Germany) and identify where item pool extensions are necessary.

Motivated by related research in psychological and educational measurement (e.g. Gorin and Embretson 2006; Embretson 1983; Hartig et al. 2012), the second and main study objective is to assess whether – and how well – various observable CGFI features can be used to predict item difficulty. This approach has two potential advantages in the context of ILTSs. First, reliable item difficulty models could substitute prior calibration relying on expensive pilot testing or intuitive ratings, which is commonplace in current ILTSs, and instead aid in the automatic or semi-automatic scoring and generation of unlimited new items with desirable linguistic and psychometric properties (cf. Embretson 1998; Gierl and Haladyna 2012). The second advantage is that such models can provide valuable insights regarding the relative difficulty contribution of learning targets and various other item features. In the future, such data-driven insights can inform the overall structuring of learning content and adaptive sequencing and help predict individual learners' difficulties. More generally, the methodology and findings of this study will also be of interest to test and exercise developers, as well as to researchers of second language acquisition (SLA).

Research Context

Current Approaches to DDA

To date, DDA has been applied predominantly in ILTSs targeting vocabulary and reading skills (e.g. *Dynamic e-book guidance system* (Sung and Wu 2017), *MEL-Enhanced* (Sandberg et al. 2014), *PIMS* (Chen and Hsu 2008), *REAP* (Heilman et al. 2010), *U-Reading* (Wu et al. 2011), Chen and Chung 2008). In contrast, DDA in the area of grammar appears to have been implemented only in two ILTSs, Moeyaert et al.'s (2016) system and *English ABLE* (Zapata-Rivera et al. 2007), both of which focus on formal accuracy only. The former offers exercises targeting a single learning dimension (French verb conjugation), while the latter treats several grammatical features such as subject-verb agreement or pronoun form, each with its own set of (error-correction) exercises.

DDA has its origins in computer adaptive testing (CAT), where test item difficulty is adapted to quickly and accurately assess test takers' ability level (cf. Van der Linden and Glas 2010). Its purpose in ILTSs goes beyond just assessment and extends to the promotion of learning and motivation (Eggen 2012; Shute et al. 2007). Timms illustrates the reasoning behind this approach as follows:

> [T]he difficulty of the problem has a large effect on how productive the interaction between the student and the learning materials will be. If the student finds the problem too easy, little learning will occur. In contrast, if the problem is too difficult […], they will learn nothing and may also become discouraged. (Timms 2007, p. 213)

Arguably, therefore, the optimal adaptive selection and sequencing of exercise items should ensure that learners are challenged yet capable of succeeding. This idea has been influential in learning theory for some time now, notably in Vygotsky's (1978) zone of proximal development theory, flow theory (Csikszentmihalyi 1991/2008), self-determination theory (Deci and Ryan 1985) and Krashen's (1985) input hypothesis. It has also found support in recent CAT studies which show that DDA can lead to higher achievement, test-relevant motivation and engagement, as well as to more positive subjective test experiences and lower anxiety levels than non-adaptive tests (Fritts and Marszalek 2010; Martin and Lazendic 2018; but see also Ling et al. 2017). There has, however, been very little research on the precise effects of DDA in digital learning environments and results have been mixed. The only controlled study in the area of language learning was conducted on an ILTS targeting a single dimension, French verb conjugation (Moeyaert et al. 2016). Using IRT to estimate item difficulty and learner ability, the study tested five DDA algorithms, each selecting exercise items with a specific success probability range (from 40 to 50% to 80–90%). Results indicate that DDA did not affect learning and motivation significantly in any of the conditions and, furthermore, did not differ from random sequencing, regardless of learners' proficiency level. At the same time, a handful of studies investigating the impact of DDA on learning in intelligent tutors/instructional games outside language learning have reported more positive results (Camp et al. 2001 and Salden et al. 2004 on air traffic control training; Kalyuga and Sweller 2005 on algebra; Yuksel et al. 2016 on music instruction), although some studies point in the opposite direction. Orvis et al.

(2008) on military training and Shute et al. (2007) on algebra, for instance, found no correlation between DDA and increased learning outcomes. In addition, there is some indication that there may not be a one-size-fits-all approach to DDA. Thus, Mitrovic and Martin's (2004) study on SQL programming shows that the positive impact of DDA may be mediated by factors such as learners' proficiency, with advanced learners benefitting the most (see also Orvis et al. 2008, and the CAT studies cited above). Given the lack of unanimity in previous studies and their different domain and task-type foci, it is also possible that DDA may be effective in some learning (sub-)domains or task types but not in others.

While the identification of optimal DDA algorithms – possibly tailored to different learner profiles and especially in ILTSs targeting multidimensional learning domains such as the English tense system – urgently requires more empirical research that would also be of relevance for existing learning theories, this paper concentrates on the more basic issue of implementing DDA in digital learning environments. As Wauters et al. (2012) note, a prerequisite for DDA is having learning materials with a known difficulty level. Yet, the measurement of difficulty in existing ILTSs has some limitations. In some systems, difficulty is evaluated by human raters (system designers, educators, learners) (cf. *REAP*), potentially introducing subjectivity and bias (Impara and Plake 1998; Wauters et al. 2012) and increasing costs. Other systems implement observable item attributes as predictors but with little or no empirical validation (*MEL-Enhanced*; *PIMS*; *U-READING*; Chen and Chung (2008)). When more objective methods (e.g. based on classical test theory or IRT) are used, items require calibration using large numbers of real learners prior to or during system use (Wauters et al. 2012; e.g. *Dynamic e-book guidance system*; *English ABLE*; *MEL-Enhanced*; Moeyaert et al. 2016). This may quickly become infeasible when a subject domain like the English tense system requires a large pool of exercise items.

In light of these challenges, we propose predictive difficulty modelling as a more objective and economical alternative for item pool generation and calibration in DDA-enabling ILTSs. We also argue that this method can inform curriculum design and adaptivity in our ILTS, including DDA at the level of learning targets, which seems to be lacking in most existing ILTSs. We turn to these and related issues next.

Measuring and Predicting Item Difficulty

As mentioned in the "Current Approaches to DDA" section, there are different methods for estimating item difficulty. In psychological and educational measurement, IRT has long been influential. Psychometric models within IRT assume that a person's response to an item depends on qualities of both the person and the item (cf. Embretson and Reise 2000). The simplest model, the one-parameter Rasch model, describes the probability π of a correct answer y to an item i as a logistic function of the difference between the person's ability parameter (θ_p) and the item difficulty parameter (β_i):

$$\pi\left(y_{pi}=1|\theta_p,\beta_i\right)=\frac{\exp\left(\theta_p-\beta_i\right)}{1+\exp\left(\theta_p-\beta_i\right)} \quad (1)$$

This makes it possible to map the ability and difficulty parameters on a common scale and estimate a probability of success for each item and person. For example, if a person with an estimated ability level of 1 logit receives an item with a difficulty of 1 logit, then according to the model, that person has a 50% chance of solving the item correctly. For an item with a difficulty that is 1 logit above the person's ability, the probability of success falls to 27%, while an item with a difficulty 2 logits below the person's ability increases the probability to 88%.

Typically, reliable item difficulty estimates are obtained by piloting items with a large sample of persons (recommendations vary between 50 and 1000; cf. Linacre 1994; Tsutakawa and Johnson 1990). These estimates, together with a choice of a desired probability of success (e.g. 50%, considered neither too easy nor too difficult), can help assess the ability of new persons and serve as the basis for computer-based DDA. While the identification of an optimal probability of success for specific learner groups remains an empirical question (see the preceding section), it should be pointed out that item calibration can also inform item pool development in ILTSs, an issue currently rarely discussed in ILTS literature, which has so far concentrated on technological issues (cf. Vajjala and Meurers 2012). Specifically, if the population sampled for calibration is also the ILTS's target population, the appropriateness of the calibrated item pool can be determined, that is, whether the item pool contains enough items for the range of learner abilities found in that population. This paper addresses this process to some extent in the "Difficulty Calibration" section below, leaving a deeper investigation for future work (for a CAT example, see Reckase 2010).

In psychology and educational measurement, IRT has been instrumental in the study of construct representation and validity (Embretson 1983). It is assumed that if it is possible to formulate hypotheses identifying the cognitive constructs (knowledge, skills and other cognitive characteristics) involved in successful task performance and describe how they are represented by features of individual task items, these hypotheses can be tested empirically. If differences in item difficulties are indeed explained by item features, then empirical support is provided for assumptions about construct representation (e.g. Embretson 1998; Freedle and Kostin 1993; Gorin and Embretson 2006). This is usually done by regressing item difficulties on item features.

The present study also seeks to model the relationship between item features and item difficulty. As detailed in the "Candidate Predictors" section, we consider a range of potential predictors specific to CGFIs targeting the English tenses. However, our long-term goal is not to study construct representation but rather a) to inform curriculum design and adaptivity in our ILTS and b) to predict the difficulty of new exercise items.

Regarding the first goal, some of the features considered, such as the tense and voice prompted by an item, represent primary tutoring targets of our ILTS, while others relate to contexts of use (e.g. conditionals and reported speech) or the overall syntactic and vocabulary complexity of a CGFI (see the "Candidate Predictors" section for details). If, as we expect, the primary targets and context types are influential correlates of item difficulty, this may have implications for the structuring and adaptivity of the learning content. First, data-driven insights may be useful for informing general content ordering. For example, as far as it also makes pedagogical sense, practice material can be organized according to the relative difficulty contribution of each tense, concomitant grammatical features (e.g. active/passive voice, (ir)regular morphology) and other exercise characteristics causing difficulty. Second, an IRT-based sequencing algorithm operating over this general content structure would enable the system to present exercise materials whose difficulty is adapted to learners' ability – both within and across learning targets – or to anticipate where additional scaffolding is necessary (e.g. more or less detailed explanations and hints).[1]

The second, and more immediate, goal is to build a model that can predict the difficulty of future CGFIs based on their features. A further step would be to train the system to identify relevant features in new exercise material and rate difficulty automatically. As a consequence, the system would also be able to generate or recommend exercise items with feature constellations specifically tailored to the current needs of the learner.[2,3]

These goals, however, extend beyond this pilot study. Here, we focus on difficulty prediction alone and evaluate the generalizability of several different item difficulty models using statistical cross-validation. Details on the methods employed are presented in the "Data and Methods" section below. First, however, a description of the item pool and potential difficulty predictors is provided.

CGFIs Targeting the English Tenses

Exercise Format

This paper presents a model for predicting the difficulty of CGFIs targeting the English tenses. The term CGFI mimics Purpura's (2004) cued gap-filling tasks, where learners read a short text and fill in the gaps using cues usually consisting of a single word which must be transformed to fit the context. The CGFIs in this study are shorter (spanning two sentences on average) and contain a single gap. As the following examples show, each gap is followed by a bracketed cue, typically a single lexical verb in the infinitive but sometimes also a subject pronoun, an adverb and/or the negative particle *not*.

(1) The Taj Mahal __________________ (build) around 1640.

(2) Laura: Where's Julie? Isn't she here?

Mark: She isn't, I __________ (not, see) her all day.

(3) That man looks familiar. I __________ (definitely, see) him somewhere before.

Though the primary focus of these items is on the form and meaning of the English tenses, the examples above show that a number of epiphenomenal features, including voice,

[1] *English ABLE*, an ILTS targeting grammatical form accuracy via exercises targeting a single category at a time, takes a similar approach. There, exercises within each category are also sequenced using IRT, while "[t]he next category is selected based on a predefined sequence of categories obtained through preliminary difficulty analysis" and the learner's previous performance (Zapata-Rivera et al. 2007, p. 327). The nature of this difficulty analysis is not specified further.

[2] Adaptive item selection can be implemented using the matroid optimization method described in Bengs et al. (2018).

[3] For more on automatic item generation, to date mainly advanced in cognitive testing and STEM disciplines, see, Attali (2018), Bejar et al. (2003), Embretson (1998, 1999, 2005) and Gierl and Haladyna (2012).

polarity, person/number inflection, word order and irregular morphology, are also targeted (see the "Candidate Predictors" section).

CGFIs belong to a group of limited-production exercise formats that have been criticized for their repetitiveness, artificiality and unsuitability for the promotion of broader communicative skills. At the same time, however, even a cursory overview of standard textbooks, practice grammars and online platforms shows that they represent a common and valued exercise format. Probable reasons for this are their relatively straightforward design and utility in targeted grammar practice and assessment (Purpura 2004). From a pedagogical perspective, they can be particularly effective in focusing learners' attention on specific linguistic forms and form-meaning relationships. This approach is seen as especially advantageous for the development of explicit, declarative knowledge, which plays an important role in early second language acquisition (cf. e.g. DeKeyser 2005; Ellis 2012; Norris and Ortega 2000; Schmidt 1995). Another distinguishing feature of CGFIs is that they target both receptive and productive skills. Compared to multiple-choice or error-recognition tasks, for instance, which require recognition or recall of grammatical form and meaning, CGFIs require not only the ability to reconstruct contextually implied grammatical meaning and retrieve the necessary linguistic form, but also to produce this form accurately (Purpura 2004, p. 127). Thus, successful exercise completion depends on the development of each of these abilities. Learner ability, however, is not the only determinant of success. Given that some CGFIs are easy for most learners, while others represent a challenge even for the strongest ones, it follows that variation in item difficulty must also be considered. We discuss CGFI features that may be implicated in this variability next.

Candidate Predictors

To our knowledge, there have been no previous attempts to identify linguistic features affecting or correlating with CGFI difficulty (or cued gap-filling tasks with multiple gaps). To address this problem, we considered relevant candidates in the SLA and psycholinguistic literature, as well as features known to affect the difficulty of related task types (see especially Beinborn 2016, Beinborn et al. 2014, and Svetashova 2015 on C- and X-tests, as well as multiple-choice cloze tests). To systematize the investigation, we distinguish three feature categories: gap-level, context and item-level features. These are discussed next and listed in Table 4 in the appendix.

Gap-Level Features

Gap-level features refer to linguistic properties of the gap solution. The most obvious and presumably strongest candidate predictor is tense.[4] As Table 4 shows, not all English tenses were considered in this study. Due to the practical difficulty of collecting sufficient amounts of relevant learner data at this stage, both future tense forms (i.e. tenses with *will*/*shall*) and future meanings (e.g. the future meaning of the simple present) were excluded, as was the rare conditional perfect progressive (*would have been* V-*ing*). However, the semi-modals *used to* and *was*/*were going to*, which express

[4] The term tense is used here in its everyday sense, covering form-meaning pairings commonly known as, for instance, the simple present or past perfect progressive.

past habituality and past intention/predictability respectively (Declerck et al. 2006), were included on par with the tenses. Despite extensive SLA literature on tense and aspect in L2 English, it is difficult to anticipate the relative effect of these tenses and semi-modals on item difficulty. First, although it is well known that progressive and perfect tenses are generally acquired later than simple tenses (Bardovi-Harlig 2000) and are especially challenging for (German) learners, usage patterns encompass both under- and overuse (cf. e.g. Axelsson and Hahn 2001; Davydova 2011), with overuse entailing that CGFIs targeting other presumably more straightforward tenses (e.g. simple tenses) may be solved incorrectly due to perfect/progressive tense misconceptions.[5] Second, since this study distinguishes between correct and incorrect solutions only to enable the implementation of the Rasch model (see the "Answer Coding" section), it is at this stage impossible to differentiate tense misuses from mere errors in form, a distinction usually made in SLA studies.

As stated in the "Exercise Format" section, the item pool also targeted a number of epiphenomenal grammatical features, including voice, polarity, subject-verb agreement, word order, adverb placement and (ir)regular lexical verb morphology. These features may be correlated with item difficulty since they require distinctive morphological and/or syntactic knowledge and skills. In line with Eckman (1977) and White (1989), we hypothesize that marked realizations (e.g. passive voice, negative polarity, marked person/number morphology, etc.) are more challenging than unmarked or non-realizations.[6]

Because CGFI difficulty may be affected by the interaction between the tense/semi-modal and one or more of the epiphenomenal features described above, two gap-level measures, morpho-syntactic edit distance (MSED) and cue size, were adopted as simple proxies. MSED refers to the number of syntactic and morphological transformations necessary to arrive at a target form or construction.[7] In the present study, the initial stimulus is the bracketed material after the gap, the target may or may not be grammatically composite and MSED in the data ranges from 0 to 7. To illustrate, in 'cue: *go* → solution: *go*' no transformations are required, while that of 'cue: *still*, *consider* → solution: *is still being considered*' requires seven.[8] Since the second target solution is morpho-syntactically more complex than the first, the chances of committing an error may be higher. Hence, we hypothesize that MSED may also predict CGFI difficulty.[9] Cue size is a more rudimentary measure, referring to the number of words in the cue and reflecting increases in syntactic complexity. In the present data, it ranges

[5] Bardovi-Harlig (2000, p. 419) posits the following sequence of emergence of tense forms used expressing past temporality: (simple) present (default) > (simple) past > past progressive > present perfect > past perfect. She also offers some evidence that present and past perfect progressive forms emerge after present and past perfect forms respectively (Bardovi-Harlig 2000, Ch. 3). In addition, Bailey (1987) finds that the present progressive is acquired before the past progressive. We are not aware of studies shedding light on the location of the simple conditional and the conditional progressive in the acquisition sequence.

[6] Following Rice (2007, p. 80), marked forms are "less natural", "more complex", "less common" and/or "harder to articulate" than their counterparts.

[7] This measure represents a slightly modified version of Spada and Tomita's (2010) linguistic complexity measure.

[8] These include: 1) supply the auxiliary *be* 2) in the simple present 3) third person singular; 4) supply the auxiliary *be* 5) in its present participle form; 6) supply the past participle form of the lexical verb; 7) place the adverb between the first and second auxiliary.

[9] As Collins et al. (2009: 339) point out, however, previous research shows that the relationship between the difficulty of a linguistic structure and LC may be mediated by other factors such as input frequency and salience. The same applies to the markedness feature discussed above.

from 1 to 3, including the lexical verb and, optionally, an adverb, the negative particle *not* and/or a pronoun.

Lastly, previous studies on the difficulty of related text completion tasks, such as C-tests and cloze tests, have hypothesized that successful gap-filling depends partly on word familiarity (e.g. Beinborn et al. 2014; Svetashova 2015). Since word familiarity cannot be measured directly, word frequency in language corpora is usually employed as a proxy and has been shown to positively correlate with task and/or gap-item difficulty. Unlike cloze and C-tests, CGFIs provide the lexical material needed to fill the gap explicitly and so frequency could be deemed unrelated to CGFI difficulty. Despite this, lexical verb frequency measures were included, since semantic familiarity arguably plays a role in the reconstruction of the propositional meaning expressed by the gapped sentence, including its temporal, aspectual and voice meanings. Frequency information comes from the spoken and written components of the *British National Corpus* (henceforth *BNC-S* and *BNC-W*; Hoffmann et al. 2008) and *SUBTLEXus* (Brysbaert and New 2009). Several related measures were employed (features 37–54 in Table 4). For the *BNC*, these include frequencies of the verb lemma and the specific verb form required, as well as the verb form percentage. The latter two measures were included to capture the likelihood of having encountered a morphological form. The *SUBTLEXus* measures are similar, except that lemma frequencies were replaced with type frequencies. The number/proportion of *SUBTLEXus* documents containing a type was also considered (Brysbaert and New 2009).

Context Features

Several features were defined to examine whether the type of context in which a gap appears affects item difficulty. A first feature group refers to clause type and is categorized according to whether the gapped clause is a simple sentence, part of a compound sentence, superordinate and/or subordinate. The last two contain further grammatical and semantic subcategories:

- superordinate clause: conditional consequent; head of a temporal clause; miscellaneous
- subordinate clause: conditional antecedent; object of the verb *wish*; reported clause; relative clause; temporal clause; miscellaneous

These features were included under the heading of context based on the fact that different clause types (e.g. central conditional antecedents and temporal clauses vs. main clauses) tend to select for different tense/aspect combinations (cf. e.g. Haegeman 2006). We therefore hypothesized that clause type (and the nature of neighboring clauses, if any) can function as a kind of contextual cue to the solution of an item. The classification above is still fairly coarse, however, partly due to the small number of items in the current dataset. Future work should address this.

The second group refers to gap position within the CGFI (beginning, middle or end). Gap position has been used in studies on cloze and C-tests (e.g. Beinborn et al. 2014; Svetashova 2015) on the hypothesis that a gap difficulty is increased by the number of preceding gaps. For the present CGFIs, the reverse could be true: the later the gap appears in the CGFI, the more contextual information will be available to solve it.

Finally, since the item pool contains multiple items representing dialogic exchanges, we included a binary dialogic/monologic feature to assess its effect.

Item-Level Features

Item-level features describe global syntactic and lexical characteristics that may affect CGFI readability and hence difficulty.

Syntactic measures include the number of a) sentences, b) clauses and c) dependent (finite) clauses within a CGFI, as well as the ratio between c) and b). Total CGFI length and mean sentence length in words were also calculated. These features have been found to be good predictors of text readability and C-test difficulty (cf. Beinborn 2016; Svetashova 2015; Vajjala and Meurers 2012).

Several different features describe CGFI vocabulary. Psycholinguistic research shows word frequency plays an important role in language comprehension, with high-frequency vocabulary enabling faster lexical access and therefore increasing readability (Brysbaert and New 2009). To test this hypothesis, we calculate for each CGFI the mean word type frequency and corpus range in *SUBTLEXus*. This is done separately for content words, function words and all words per CGFI (features 83–94). A related measure is McDonald and Shillcock's (2001) contextual distinctiveness score, which represents the co-occurrence probability of a word with 500 highly frequent lemmas in the *BNC*. We include average CGFI scores over all words for this measure.

Age of acquisition (AoA) is another possible predictor that has been shown to explain variance in word recognition and reading (Weekes et al. 2006) and SLA experiments (Izura et al. 2011). We calculate average AoA of CGFI vocabulary using Kuperman et al.'s (2012) database of informant ratings for 30,000 English words. Another measure associated with lexical processing behavior is Brysbaert et al.'s (2014) vocabulary concreteness, involving reference to easily perceptible entities. We obtain scores for content, function and all words from Brysbaert et al.'s database containing ratings of 40,000 words (features 95–97).

Finally, we adopt two lexical features previously tested in readability studies (cf. Vajjala and Meurers 2012): lexical density (percentage of content words) and mean word length in characters. To examine the effect of word length variation, we also include the word length standard deviation.

Having introduced the three categories of features potentially associated with CGFI difficulty, we next provide details of the data and methods employed in this study, followed by the results.

Data and Methods

To estimate item difficulties in the CGFI pool and build a predictive model, a paper-and-pencil test was conducted with a sample of German 9th and 10th grade learners of English. In Germany, students begin learning English as a foreign language in the 3rd grade or earlier and are expected to reach proficiency levels A2 to B1 by the end of the 9th and 10th grades respectively (cf. e.g. Niedersächsisches Kultusministerium 2015a, b). At this stage, all tenses and semi-modals targeted by the initial item pool have been introduced and are reviewed and practiced extensively. The next subsections detail the item pool development, the administration and scoring of the test, the feature extraction procedure and the statistical methods employed.

Item Pool Development and Test Administration

To ensure item pool appropriateness for the target learner group, CGFIs were collected from current print and digital EFL materials for the 9th and 10th grades by the three major education publishers in Germany (*Cornelsen*, *Diesterweg*, *Klett*) and practice grammars for intermediate students. Item quality and possible solutions were evaluated by four native speaker experts and items were modified or discarded if they were topically or stylistically awkward or ambiguous in terms of the range of acceptable solutions (ignoring non-standard/marginal alternatives). In total, 330 items were selected for the test.

After obtaining all necessary permissions and informed consent, the test was administered to 787 9th and 10th graders in two preparatory high schools (*Gymnasium*) and two integrated comprehensive schools (*Integrierte Gesamtschule*) in Lower Saxony. After discarding empty and aborted tests, the number was reduced to 689. Table 1 shows a relatively equal split between school grades but not school forms: approximately 73% of the participants were preparatory high school students.

Test instructions were provided prior to administration, including an example item and answer. Participants were not informed which specific tenses the test would cover but told that the semi-modals *going to* and *used to* were admissible. 40 min were provided to complete the test.

Due to the impracticality of asking students to solve all 330 items, a matrix design consisting of 90 unique booklets containing a subset of 44 items in random order was used. Despite our efforts, 38 items were retrospectively found to permit more than one solution and had to be omitted from the analysis. Each of the remaining 292 items was seen by a mean of 91.68 students (SD = 3.86), with each student working on a mean of 38.86 items (SD = 2.42).

Answer Coding

Test data collection produced 26,772 data points and the coding process distinguished between 'correct' and 'incorrect' answers. Correct answers include a) complete matches (*N* = 7788) and b) correct answers with a spelling mistake unrelated to irregular morphology (e.g. **tought* vs. *thought* or **finaly* vs. *finally*; *N* = 94). Incorrect answers include a) morpho-syntactic inaccuracies (*N* = 17,865), b) unfilled gaps (*N* = 842) and c) illegible, stricken through or unserious responses (*N* = 183). Two authors

Table 1 Test participants according to school type and grade

	Grade nine	Grade ten	Total
Preparatory high school	242	261	503
Integrated comprehensive school	108	78	186
Total	350	339	689

and four student assistants participated in the coding. Interrater reliability was maintained via a coding manual and a workshop, joint discussions of uncertainties and frequent sample checks by one author.[10]

Feature Extraction

The linguistic features tested in the present study were extracted as follows. Features 37–46 were extracted using *BNCweb* (Hoffmann et al. 2008). Features 47–54 were extracted from a *SUBTLEXus* wordlist available at https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus. Features 73–76 and 78–79 were extracted with *WordSmith Tools 6* (Scott 2011). Features 77 and 83–99 were extracted using *Taales 2.2* (Kyle and Crossley 2015). All remaining features were coded and double-checked manually by two authors.

Statistical Analysis

Difficulty calibration was performed based on the Rasch model (Eq. 1) as implemented in the R package TAM. Four items were excluded from the analysis due to low informativeness (they were solved by all or none of the participants). The remaining 288 items were rescaled and their difficulties were subsequently used in the prediction analysis (see the "Difficulty Calibration" section for details).

To model CGFI difficulty, several ridge regression models with different CGFI feature sets were built using the *scikit learn* library for Python. Ridge regression was chosen over other approaches to avoid overfitting and tackle multicollinearity.[11] As a pre-processing step, continuous attributes were scaled to the interval [0,1], while categorical ones were binarized. This resulted in 99 individual features tested in the prediction experiments. No regression intercept was included to obtain difficulty estimates for all features. Nested cross-validation was used for hyper-parameter setting and prediction performance evaluation (five-fold cross-validation repeated 10 times). The following evaluation criteria were used: the average root mean squared error (RMSE)[12] and the Pearson correlation coefficient r between predicted and observed difficulties. Finally, prediction intervals were calculated for the best performing model to estimate its precision on future data.

Results

Difficulty Calibration

The results of the item pool calibration are displayed in Fig. 1. The overall reliability of the items in measuring tense-related grammatical ability was high ($\mathrm{Rel}_{\mathrm{EAP}} = 0.88$).

[10] As a slight drawback, no rater consistency measures were calculated. However, the handful of coding mistakes that were made (approximately one hundred in total) were corrected during an independent learner error analysis study belonging to a different strand of the research project out of which this study originates.

[11] Lasso regression and decision trees yielded similar results.

[12] RMSE represents the square root of the mean of the squared prediction errors. Here, it expresses the standard deviation of predicted from observed difficulties.

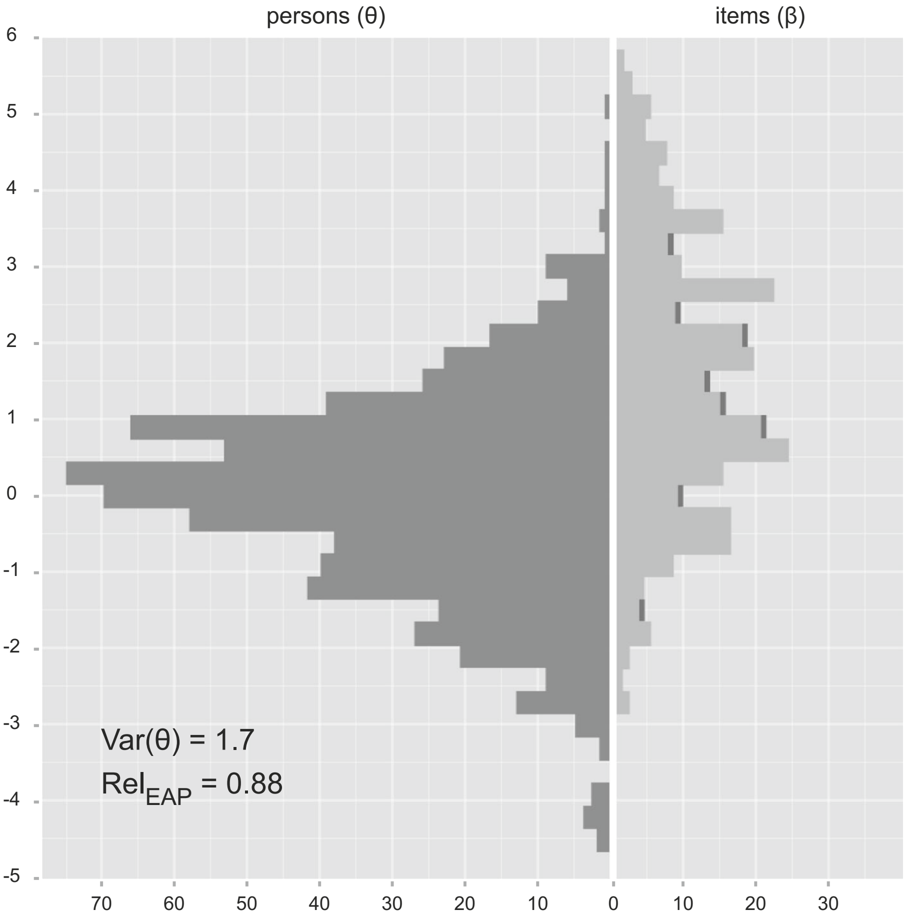


Fig. 1 Wright map displaying the distribution of estimated test participants' abilities (left) and item difficulties (right). Misfitting items (*N* = 8), excluded from subsequent analysis, are marked in dark grey

However, eight items did not fit the model well (infit mean square > 1.33; cf. Wilson 2004) and were excluded from subsequent analysis.

For the remaining 280 items, difficulties ranged from −2.56 to +5.78 logits. As visible from Fig. 1, the overlap between student abilities (M = 0.00, SD = 1.30) and item difficulties (M = 1.51, SD = 1.78) is not ideal. Overall, the latter significantly exceed the former (two-sample t = 14.645, df = 967, $p < 0.0001$). It is also noticeable that the item pool contains a number of items that were too difficult for most students, and very few items suitable for learners at the bottom of the scale. This suggests that the item pool tends to be overly demanding for German 9th and 10th graders and that the most pressing need for additional items is at the lower end of the scale. These findings, however, do not reveal what item content characteristics affect item difficulty and, hence, offer no guidelines for designing future item pool extensions. In order to begin bridging this gap, we next examine the ability of a number of CGFI features to predict item difficulty.

Difficulty Prediction

The prediction experiments are based on the CGFI difficulty estimates obtained from the calibration and the three feature groups described earlier: gap-level, context and item-level features. We compare the predictive performance of several feature combinations to a baseline model (a featureless model always predicting the overall mean difficulty) to identify the best approach.

Table 2 shows that feature models B–K all represent an improvement over the baseline. Model B, which contains only the 11 tenses and two semi-modals, checks how well these features can predict item difficulty by themselves, given that they represent core targets of the CGFIs. Indeed, this model clearly outperforms the baseline, with a considerably lower RMSE and a stronger correlation between predicted and observed item difficulties. Adding the remaining gap-level features improves the results somewhat (Model C). Still, this improvement is fairly small relative to the much larger number of features, suggesting that most of the additional features have low predictive power. The same holds for context and item-level features, which perform poorly both on their own and in combination (D, E and H), yielding RMSE values close to the baseline. However, the moderate correlation with observed difficulties means that they capture some variability in the data.

To better understand the usefulness of the three feature categories, we next assess the performance of different feature combinations. The first, F, contains all 99 features. As shown in Table 2, the full model delivers considerably better results than C (a 18% decrease in RMSE and a 6% correlation increase), confirming that context and item-level features hold some promise. We also check whether a sparser model can be obtained. Models G–I each omit a different category. G, containing gap-level and context features, is the best of the three. Interestingly, it delivers results equivalent to those of the full model but with almost 30% less features. It also represents a considerable improvement over gap-level features alone. H and I are substantially worse, with H performing slightly better than gap-level features alone and I being the

Table 2 Cross-validation results for eleven predictive models with different feature combinations.

	Models	Features	RMSE	*r*
A	Baseline	–	1.78	−0.10
B	Tenses and semi-modals	13	1.08	0.79
C	Gap-level features	54	0.95	0.85
D	Context features	18	1.62	0.41
E	Item-level features	27	1.64	0.39
F	Full model	99	0.78	0.90
G	Gap-level and context features	72	0.78	0.90
H	Gap-level and item-level features	81	0.93	0.86
I	Context and item-level features	45	1.55	0.50
J	Recursive feature elimination I	56	0.75	0.91
K	Recursive feature elimination II	36	0.77	0.90

All Pearson correlation values are significant at the $p < 0.0001$ level weakest by far. Taken together, these findings suggest that gap-level features have the strongest predictive power, followed by context and item-level features.

Removing entire blocks of features as described above, however, may obscure the predictive power of individual features. To examine this possibility, we perform recursive feature elimination of the least influential features. Table 2 reports two of the obtained models, J and K, with 56 and 36 features respectively. Model J delivers the best results, reducing the RMSE by 5% and increasing the Pearson correlation by 1% compared to F and G. Model K is the sparsest model within a 3% tolerance of Model J. It performs virtually equally well as F and G but with far fewer features.

The results of Model K are displayed in Table 3, which shows that features from all three categories contribute to CGFI difficulty/ease. However, the gap-level category is clearly the most important, accounting for 72% of all features in the model. As expected, the type of tense/semi-modal required has a major impact (apart from the past progressive and past perfect).[13] The effect of epiphenomenal grammatical features is confirmed too, though of these, subject-verb agreement appears to play no role whatsoever.[14] Interestingly, the selection of MSED and cue size indicate that there are important interactions between (and within) tenses/semi-modals and epiphenomenal features. Finally, the frequency of the lexical verb form, lemma or type also appear influential, suggesting that our CGFIs require not only grammatical but also lexical knowledge.

Further influence is exerted by clause type and especially certain subordinate clauses associated with special tense constraints (e.g. reported clauses, objects of *wish*). Gap position and dialogic/monologic context are deselected, although they are present in the larger Model J. Item-level features include syntactic embedding (number of dependent clauses), word length and word length variation, AoA, concreteness and mean document range. The latter five suggest once again that various aspects of vocabulary quality play a role in CGFI processing.

Lastly, we examine the predictive precision of Model K to assess its potential utility in the context of DDA. Here, we estimate the prediction interval in which individual future observations are expected to fall with a 90% probability

[13] The tenses/semi-modals in Model K are ranked from highest negative impact to highest positive impact as follows: (i) (−)simple past > (−)simple present > (−)simple conditional > (−)present perfect > (−)conditional perfect > (−)present progressive > **{(−)past progressive > (−)past perfect}** > (+)*used to* > (+)conditional progressive > (+)past perfect progressive > (+)present perfect progressive > (+)*was/were going to*.

The past progressive and the past perfect do not feature in Model K due to their negligible (negative) impact. They are included here in brackets to show larger recursive elimination models place them in intermediate position between the present progressive and *used to*. We note two interesting tendencies emerging from this ranking. First, morpho-syntactically simpler forms tend to decrease item difficulty, while more complex forms tend to increase it, a finding which is in line with previous SLA research on linguistic difficulty (cf. e.g. DeKeyser 2005; Hulstijn and De Graaff 1994; Spada and Tomita 2010). It is also noticeable that simple tenses precede perfect tenses, which in turn tend to precede progressive tenses. The progressives do not have gnrammaticalized equivalent in German and so our finding is also very much in line with previous research, which shows that such forms are particularly troublesome for German learners (cf. e.g. Götz 2015; Kämmerer 2012; Rogatcheva 2012). It should be emphasized, however, that these results are still preliminary. As pointed out in the discussion, larger and more balanced datasets are needed to examine their validity and support broader generalizations.

[14] In line with our hypothesis, the passive voice, irregular morphology, *wh*-interrogative word order, which were defined as 'marked', in the "Context Features" section boost item difficulty in comparison to their 'unmarked' counterparts (active voice, regular morphology, declarative word order). Again, the generalizability of these findings remains to be examined in future research.

Table 3 Features selected in Model K (36 features; RMSE = 0.77, $r = 0.91$), ranked according to their impact on item difficulty (+1 = highest positive impact; −1 = highest negative impact)

Feature groups	Selected features	Ranking
Gap-level features ($N = 26$, 72%)	Simple present	−4
	Simple past	−3
	Simple conditional	−5
	Present progressive	−15
	Conditional progressive	7
	Present perfect	−10
	Conditional perfect	−14
	Present perfect progressive	5
	Past perfect progressive	6
	Was/were going to	4
	Used to	11
	Voice: passive	20
	Adverbs: absent	12
	Polarity: positive	13
	Word order: declarative	8
	Word order: *wh*-interrogative	16
	Lexical verb morphology: irregular	17
	MSED: 7	14
	Cue size	1
	BNC-S normalized verb lemma frequency	−13
	BNC-S normalized verb form frequency	−9
	BNC-S $\log_{10}$ normalized verb form frequency	−2
	BNC-W $\log_{10}$ normalized verb lemma frequency	3
	SUBTLEXus normalized type frequency	−1
	SUBTLEXus normalized verb form frequency	2
	SUBTLEXus $\log_{10}$ type document range	−12
Context features ($N = 4$, 11%)	Subordinate clause type: reported clause	19
	Subordinate clause type: object of *wish*	15
	Subordinate clause type: temporal clause	−7
	Simple sentence: no	18
Item-level features ($N = 6$, 17%)	Number of dependent clauses	−8
	Mean word length in characters	9
	Word length SD	−6
	Age of acquisition (AW)	21
	Concreteness (AW)	−11
	SUBTLEXus $\log_{10}$ mean document range (AW)	10

via the formula $\hat{\beta}_i \pm 1.64 * \mathrm{RMSE}$. Thus, for example, the actual value of an item with a predicted difficulty of 1 logit may fall anywhere between −0.22 and 2.22 logits. For a DDA setting, this means that a student with an ability level of 1 logit would have an estimated probability of success on this item ranging between 22% and 78% (using the formula $\pi = e^{\theta-\beta}/1 + e^{\theta-\beta}$). The more accurate but larger Model J would lead to almost the same margin (π=50% ± 27%). With this, we turn to the discussion of the results.

Discussion and Conclusion

One aim of this pilot study was to estimate the difficulty of a set of cued gap-filling exercise items to be used in an ILTS for practicing the English tenses/semi-modals. The IRT calibration showed that the large majority of items measure the same set of abilities required for successful completion and can be ordered on a joint scale. In the future, these items can therefore be deployed in DDA directly. Somewhat surprisingly, the overall difficulty of the item pool tended to surpass a range of abilities found in the learner population (9th and 10th grade students in Germany), even though the items were sourced largely from learning materials intended for their level. This could be explained by the fact that individual exercises in the learning materials typically targeted a small set of tenses explicitly, whereas the participants in this study received no such indication (except for the possibility of using the semi-modals *was*/*were going to* and *used to*). Furthermore, the pool included items from 10th grade textbooks and intermediate practice grammars which might have been too challenging for the 9th graders. In any case, the mismatch highlights that item pool appropriateness relative to a well-defined target group of learners should be taken seriously in the development of pedagogically sound ILTSs. Thus, in our case, there is a clear need for items at the lower end of the difficulty/ability scale that must be addressed.

The second study aim was to attain a better understanding of the factors that affect CGFI difficulty in order to enable a difficulty-oriented item design which could eliminate the need for independent calibration via subjective ratings or costly pilot testing in the future. The prediction results show that CGFI difficulty is associated with a range of SLA, psycholinguistic and some specially formulated features concerning the linguistic properties of the required solution, the context surrounding the gap and, more globally, the syntactic and lexical characteristics of the CGFI as a whole. The results also show that of these, gap-level features are the most predictive, echoing previous findings on related text-completion formats (Beinborn 2016; Beinborn et al. 2014). All grammatical features at the gap level except for subject-verb agreement were shown to be useful predictors. Interestingly, MSED and cue size were also found to be influential, suggesting that item difficulty is affected by the interaction between grammatical categories, which may increase the processing load and lead to more errors. In the future, larger, more balanced datasets will make it possible to explore these interactions further. Finally, even though our CGFIs have been described as limited-production tasks strictly targeting grammatical ability (cf. Purpura 2004), it appears that this is not all they do. In particular, several lexical measures capturing various lexical verb or global vocabulary qualities are relevant for CGFI processing. This is hardly surprising given that inferring the intended grammatical meaning of the gapped verb requires the ability to understand the context. Thus, it is possible that CGFIs targeting the same grammatical features may vary in difficulty due to lexical (and possibly other) factors. This means that when referring to grammatical phenomena in score interpretation, such features need to be controlled for on the one hand. On the other hand, the significant effects of lexical features present the opportunity to include them in score interpretations as well. This would require the systematic inclusion of those features in item construction.

That said, the cross-validation results of the regression experiments are not entirely satisfactory from a DDA perspective. The size of the prediction intervals of our best models cannot guarantee exact ability-difficulty matching. However, the predictions could, upon additional validation, enable a form of DDA in which students are provided with items ranging from moderately easy to moderately difficult. Since the probability of success increases/decreases exponentially with the size of the difficulty/ability mismatch, a much greater number of overly easy and overly difficult items would effectively be filtered out. We consider these results encouraging and possibly even practical, given the dearth of research on what forms of DDA work best (e.g. exact ability/difficulty matching at some to-be-established success probability level or indeed alternating between moderately easy and difficult items). The results should also be seen in the light of state-of-the-art studies on related text-completion formats which report comparatively higher RMSE and lower Pearson correlation values (e.g. Beinborn 2016; Beinborn et al. 2014; Svetashova 2015).

The pilot study was constrained primarily by the amount of data available, notably with regard to the coverage of some binary CGFI features and feature combinations. Despite the use of a matrix design, logistical limitations and the dangers of test fatigue prohibited the inclusion of more test participants and items. In follow-up work, this problem can be addressed easily with an anchor item design, in which a small set of previously calibrated items is incorporated into subsequent tests covering underrepresented features and feature combinations. A simple linking transformation can then be used to express the new item difficulty scale in terms of the existing one (Wauters et al. 2012). Larger datasets will also allow for further statistical validation and model adjustment. Furthermore, it is possible to include additional features in the analysis. NLP tools such as L2SCA (Lu 2010) and TAASSC (Kyle 2016) offer a large number of additional morphological, syntactic and lexical measures that can be tested in the future. On the semantic-pragmatic side, the polyfunctionality/polysemy of the tense forms and semi-modals and the availability of contextual cues pointing to the correct solution, for instance, likely hold high predictive power. The inclusion of the former was constrained by insufficient variability regarding some tenses and will be addressed when more data is available. Incorporating the latter, in contrast, requires additional research into what actually counts as a cue from a learner perspective.

Several directions for future research can be identified. First, as noted above, more data is required to evaluate and improve our prediction models and to obtain a larger item pool that is more balanced and covers a wider range of abilities in the target population. Achieving this would, on the one hand, enable (semi-)automatic evaluation and manipulation of the difficulty of newly generated items for use in the ILTS. On the other hand, it would permit us to investigate the precise contribution of grammatical, contextual and other features of CGFIs on item difficulty and its implications for the content structure and adaptivity of the ILTS. Subsequently, different content structure and DDA scenarios can be constructed and tested in order to assess the precise benefits of DDA on learning outcomes and motivation. To our knowledge, no research exists in these areas with regard to learning environments targeting multidimensional content areas.

Second, as Moeyaert et al. (2016) caution, future research should also consider that simple feedback indicating the correct response is likely not enough to promote learning and that the effect of different types of corrective feedback should also be taken into account.

Third, this study focused on a single exercise format and involved administering a test with a fixed length and a uniform set of instructions. From a pedagogical and language-learning perspective, an ILTS should ideally offer a range of different exercises and exercise variants, involving, for instance, different scaffolding techniques, in order to provide learners with more varied practice opportunities. One avenue for future research, therefore, would be to compare multiple exercise types, as well as variations in exercise instructions, modes of presentation and item selections. Indeed, doing this would offer a window into a whole host of non-linguistic, task-related item characteristics that could have an effect on item difficulty. Here, it should also be mentioned that the items studied in this paper were piloted with a paper-and-pencil test for logistical reasons. In subsequent testings, a computer-based setting should be preferred in order to better approximate an ILTS setting.

And fourth, it must be noted that the present analysis employed a two-stage approach in which item difficulties and the effects of item characteristics were estimated separately. It is technically possible to include the item characteristics directly in the difficulty measurement model. However, Hartig et al. (2012) have shown that this approach delivers practically identical results. We chose the two-stage approach since the direct approach is limited with respect to cross-validation techniques. Nevertheless, including the item characteristics within the measurement model (e.g. a multifaceted Rasch model) is a promising perspective for future studies.

To conclude, this pilot study provides initial evidence regarding the possible features of CGFIs targeting the English tenses that affect exercise item difficulty. Despite the limited scope of the study, this approach has much potential in the context of (semi-)automatic difficulty scoring and manipulation for DDA in ILTSs and CAT in general. We believe that it will also be useful for educational content designers and empirical SLA researchers interested in understanding the factors that underlie learner performance.

Acknowledgements We wish to acknowledge the financial support offered for this project by the Fund for Scientific Research (Kleinforschungsprojekt) at Leuphana University Lüneburg (73100024). Thanks to Jana-Luise Bimkiewicz, Svenja Buck, Kim Greife and Mareike Vahlenkamp for their support during data collection. Thanks also to Beatrice Brodesser for her support in the formatting of this article. Last but not least, we are grateful to the three anonymous reviewers for their competent and constructive comments and suggestions. All remaining errors are our own.

Appendix

Table 4 CGFI features used in the prediction experiments

GAP-LEVEL FEATURES		Items per category	Nr.
Tenses and modals	Simple present	51	1
	Simple past	41	2
	Simple conditional	25	3
	Present progressive	19	4
	Past progressive	22	5
	Conditional progressive	10	6
	Present perfect	17	7
	Past perfect	41	8
	Conditional perfect	19	9
	Present perfect progressive	8	10
	Past perfect progressive	7	11
	Was/were going to	9	12
	Used to	11	13
Voice	Active	251	14
	Passive	29	15
Polarity	Positive	232	16
	Negative	48	17
Subject-verb agreement	Unmarked	189	18
	Marked	91	19
Adverbs	Absent	259	20
	Present	21	21
Word order	Declarative	252	22
	Yes/no interrogative	9	23
	Wh-interrogative	9	24
	Imperative	10	25
Past-tense and participial morphology	Regular	202	26
	Irregular	78	27
MSED	0	15	28
	1	41	29
	2	41	30
	3	66	31
	4	73	32
	5	32	33
	6	10	34
	7	2	35
	Cue size	–	36
BNC-S	Normalized verb lemma frequency	–	37
	Log_{10} normalized verb lemma frequency	–	38
	Normalized verb form frequency	–	39
	Log_{10} normalized verb form frequency	–	40
	Verb form percentage	–	41

Table 4 (continued)

GAP-LEVEL FEATURES		Items per category	Nr.
BNC-W	Normalized verb lemma frequency	–	42
	Log_{10} normalized verb lemma frequency	–	43
	Normalized verb form frequency	–	44
	Log_{10} normalized verb form frequency	–	45
	Verb form percentage	–	46
SUBTLEXus	Normalized type frequency	–	47
	Log_{10} normalized type frequency	–	48
	Type document range	–	49
	Log_{10} type document range	–	50
	Type document range (%)	–	51
	Normalized verb form frequency	–	52
	Log_{10} normalized verb form frequency	–	53
	Verb form percentage	–	54
CONTEXT FEATURES			
Gap position	Beginning	92	55
	Middle	97	56
	End	91	57
Subordinate clause type	Conditional antecedent	40	58
	Object of *wish*	20	59
	Relative clause	5	60
	Reported clause	30	61
	Temporal clause	16	62
	Other	5	63
Type of superordinate clause	Conditional consequent	51	64
	Head of temporal clause	25	65
	Other	9	66
Simple sentence	No	63	67
	Yes	217	68
Coordinated clause	No	256	69
	Yes	24	70
Dialogic context	No	248	71
	Yes	32	72
ITEM-LEVEL FEATURES			
SUBTLEXus	CGFI length in words	–	73
	Mean word length in characters	–	74
	Word length standard deviation	–	75
	Type count	–	76
	Lexical density	–	77
	Number of sentences	–	78
	Mean sentence length	–	79
	Number of clauses	–	80
	Number of dependent clauses	–	81
	Dependent clauses/clauses ratio	–	82
	Mean normalized type frequency (AW)	–	83
	Log_{10} mean normalized type frequency (AW)	–	84

Table 4 (continued)

GAP-LEVEL FEATURES		Items per category	Nr.
	Mean document range (AW)	–	85
	Log_{10} mean document range (AW)	–	86
	Mean normalized type frequency (CW)	–	87
	Log_{10} mean normalized type frequency (CW)	–	88
	Mean document range (CW)	–	89
	Log_{10} mean document range (CW)	–	90
	Mean normalized type frequency (FW)	–	91
	Log_{10} mean normalized type frequency (FW)	–	92
	Mean document range (FW)	–	93
	Log_{10} mean document range (FW)	–	94
Concreteness	AW	–	95
	CW	–	96
	FW	–	97
	Age of acquisition (AW)	–	98
	Contextual distinctiveness (AW)	–	99

Word frequency and document range measures are normalized per million words. Abbreviations: *AW* all words, *CW* content words, *FW* function words

References

Amaral, L. A., & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL, 23*, 4–24.

Attali, Y. (2018). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Artificial intelligence in education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018* (pp. 17–29). Cham: Springer.

Axelsson, M., & Hahn, A. (2001). The use of the progressive in Swedish and German advanced learner English: a corpus-based study. *ICAME Journal, 25*, 5–30.

Bailey, N. H. (1987). *The importance of meaning over form in second language system building: An unresolved issue*. Ph.D. dissertation, City University of New York.

Bardovi-Harlig, K. (2000). *Tense and aspect in language acquisition: Form, meaning and use*. Oxford: Blackwell.

Beinborn, L. M. (2016). *Predicting and manipulating the difficulty of text-completion exercises for language learning*. Doctoral dissertation. Darmstadt: Technische Universität Darmstadt.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association of Computational Linguistics, 2*(1), 517–529.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3) Available at: https://ejournals.bc.edu/index.php/jtla/article/view/1663. Accessed 6 May 2019.

Bengs, D., Brefeld, U., & Kröhne, U. (2018). Adaptive item selection under Matroid constraints. *Journal of Computerized Adaptive Testing, 6*(2), 15–36.

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 3–53). Berlin/Heidelberg: Springer.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Camp, G., Paas, F., Rikers, R., & van Merriënboer, J. (2001). Dynamic problem selection in air traffic control training: a comparison between performance, mental effort and mental efficiency. *Computers in Human Behavior, 17*(5/6), 575–595.

Chen, C.-M., & Chung, C.-J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education, 51*, 624–645.

Chen, C.-M., & Hsu, S.-H. (2008). Personalized intelligent mobile learning system for supporting effective English learning. *Educational Technology & Society, 11*, 153–180.

Collins, et al. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal, 93*(3), 336–353.

Csikszentmihalyi, M. (1991/2008). Flow: The psychology of optimal experience. New York Harper Perennial.

Davydova, J. (2011). *The present perfect in non-native Englishes: A corpus-based study of variation. Topics in English linguistics: Vol. 77*. Berlin: de Gruyter.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

Declerck, R. (in collaboration with S. Reed and B. Cappelle). (2006). *The grammar of the English verb phrase, volume 1: The grammar of the English tense system: A comprehensive analysis*. Berlin and New York: Mouton de Gruyter.

DeKeyser, R. (2005). What makes second-language grammar difficult? A review of issues. *Language Learning, 55*, 1), 1–1),25.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, 27*, 315–330.

Eggen, J. H. M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. In J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 11–21). Enschede: RCEC.

Ellis, R. (2012). *Language teaching research and language pedagogy*. Hoboken: Wiley-Blackwell.

Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods, 3*(3), 380–396.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64*(4), 407–433.

Embretson, S. E. (2005). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 251–267). Cambridge: Cambridge University Press.

Embretson, S. E., & Reise, S. P. (Eds.). (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing, 10*, 133–170.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*(3), 441–458.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. New York: Routledge.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394–411.

Götz, S. (2015). Tense and aspect errors in spoken learner language: Implications for language testing and assessment. In M. Callies & S. Götz (Eds.), Learner corpora in language testing and assessment (pp. 191–216). Amsterdam: John Benjamins.

Haegeman, L. (2006). Conditionals, factives and the left periphery. *Lingua, 116*(10), 1651–1669.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*(4), 665–686.

Heift, T. (2016). Web delivery of adaptive and interactive language tutoring: revisited. *International Journal of Artificial Intelligence in Education, 26*(1), 489–503.

Heilman, M., Collins-Thompson, K., Eskenazi, M., Juffs, A., & Wilson, L. (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education, 20*, 73–98.

Hoffmann, S., Evert, S., Smith, N., Lee, D., & Berglund Prytz, Y. (2008). *Corpus linguistics with BNCweb: A practical guide*. Frankfurt: Peter Lang.

Hulstijn, J. H., & De Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review, 11*, 97–112.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81.

Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: a word training study. *Journal of Memory and Language, 64*(1), 32–58.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive elearning. *Educational Technology Research and Development, 53*(3), 83–93.

Kämmerer, S. (2012). Interference in advanced English interlanguage: Scope, detectability and dependency. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora* (pp. 284–297). Brno: Masaryk University Press.

Kerr, P. (2015). Adaptive learning. *ELT Journal, 70*(1), 88–93.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Harlow: Longman.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. PhD dissertation. Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35. Accessed 16 Apr 2019.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.

Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions, 7*(4), 328.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement, 41*(7), 495–511.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295–323.

Mitrovic, A., & Martin, B. (2004). Evaluating adaptive problem selection. In W. Nejdl & P. De Bra (Eds.), *Lecture notes in computer science: Vol. 3137. Adaptive hypermedia and adaptive web-based systems* (pp. 185–194). Berlin: Springer-Verlag.

Moeyaert, M., Wauters, K., Desmet, P., & Van den Noortgate, W. (2016). When easy becomes boring and difficult becomes frustrating: disentangling the effects of item difficulty level and person proficiency on learning and motivation. *Systems, 4*(14), 1–18.

Niedersächsisches Kultusministerium (Hrsg.) (2015a). Kerncurriculum für das Gymnasium. Schuljahrgänge 5–10. Englisch. http://db2.nibis.de/1db/cuvo/datei/en_gym_si_kc_online.pdf. Accessed 6 Feb 2018.

Niedersächsisches Kultusministerium (Hrsg.) (2015b). Kerncurriculum für die Integrierte Gesamtschule. Schuljahrgänge 5–10. Englisch. http://db2.nibis.de/1db/cuvo/datei/en_igs_si_kc_online.pdf. Accessed 6 Feb 2018.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417–528.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior, 24*, 2415–2433.

Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127–141.

Rice, K. (2007). Markedness). In P. de Lacy (Ed.), *The Cambridge handbook of phonology* (pp. 79–97). Cambridge: Cambridge University Press.

Rogatcheva, S. I. (2012). Measuring learner (mis)use: Tense and aspect errors in the Bulgarian and German components of ICLE. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora* (pp. 258–272). Brno: Masaryk University Press.

Salden, R. J. C. M., Paas, F., Broers, N. J., & van Merriënboer, J. (2004). Mental effort and performance as determinants for dynamic selection of learning tasks in air traffic control training. *Instructional Science, 32*, 153–172.

Sandberg, J., Maris, M., & Hoogendoorn, P. (2014). The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning. *Computers & Education, 76*, 119–130.

Schmidt, R. (1995). Consciousness and foreign language: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu: University of Hawaii Press.

Scott, M. (2011). *WordSmith tools version 6*. Liverpool: Lexical Analysis Software.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 7–27). New York: Cambridge University Press.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). Evaluating ACED: The impact of feedback and Adaptivity on learning. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education – building technology rich learning contexts that work* (pp. 230–237). Amsterdam: IOS Press.

Slavuj, V., Meštrović, A., & Kovačić, B. (2016). Adaptivity in educational systems for language learning: a review. *Computer Assisted Language Learning, 30*(1/2), 64–90.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: a meta-analysis. *Language Learning, 60*(2), 263–308.

Sung, T.-W., & Wu, T.-T. (2017). Dynamic e-book guidance system for English reading with learning portfolio analysis. *The Electronic Library, 35*(2), 358–373.

Svetashova, Y. (2015). *C-test item difficulty prediction: Exploring the linguistic characteristics of C-tests using machine learning*. Master's thesis, Eberhard Karls Universität Tübingen.

Timms, M. J. (2007). Using item response theory (IRT) to select hints in an ITS. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education – building technology rich learning contexts that work* (pp. 213–221). Amsterdam: IOS Press.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371–390.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification. *Proceedings of the seventh workshop on innovative use of NLP for building educational applications (BEA7)*. Association for Computational Linguistics. 163–173.

Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning, 26*, 549–562.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: an auspicious collaboration between data and judgment. *Computers & Education, 58*, 1183–1193.

Weekes, B. S., Castles, A. E., & Davies, R. A. (2006). Effects of consistency and age of acquisition on reading and spelling among developing readers. *Reading and Writing, 19*(2), 133–169.

White, L. (1989). *Universal grammar and second language acquisition*. Amsterdam: Benjamins.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Manwah: Lawrence Erlbaum.

Wu, T.-T., Sung, T.-W., Huang, Y.-M., Yang, C.-S., & Yang, J.-T. (2011). Ubiquitous English learning system with dynamic personalized guidance of learning portfolio. *Educational Technology & Society, 14*(4), 164–180.

Yuksel, B. F., Oleson, K. B., Harrison, L, Peck, E. M., Afergan, D., Chang, R., & Jacob, R. J. K. (2016). Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, California, USA*, 5372–5384.

Zapata-Rivera, D., Vanwinkle, W., Shute, V., Underwood, J. S., & Bauer, M. (2007). English ABLE. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education – building technology rich learning contexts that work* (pp. 323–330). Amsterdam: IOS Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Irina Pandarova[1] · Torben Schmidt[1] · Johannes Hartig[2] · Ahcène Boubekki[1] · Roger Dale Jones[1,3] · Ulf Brefeld[1]

Torben Schmidt
torben.schmidt@leuphana.de

Johannes Hartig
hartig@dipf.de

Ahcène Boubekki
ahcene.boubekki@leuphana.de

Roger Dale Jones
r.jones@tu-braunschweig.de

Ulf Brefeld
brefeld@leuphana.de

[1] Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany

[2] Leibniz Institute for Research and Information in Education (DIPF), Rostocker Straße 6, 60323 Frankfurt, Germany

[3] Present address: Technical University Braunschweig, Universitätsplatz 2, 38106 Braunschweig, Germany