Predicting Item Difficulties and Item Dependencies for C-Tests

Alexander Robitzsch & Ina Karius

1 Introduction

The C-Test is an integrative test which is designed to test a person’s command of language by making use of the principle of reduced redundancy (Raatz & Klein-Braley 2002). It is assumed that language is redundant in a way that allows successful communication although possible flaws in the language transmission (unclarity, ambiguity, noise) may impede understanding. The addressee of a message is able to reconstruct the form and meaning of morphologically incomplete words with the help of the local and global context of the message, provided that he or she is familiar with the vocabulary, the grammatical rules and the cultural background of the language used.

Raatz and Klein-Braley (2002) as well as Eckes and Grotjahn (2006) consider a C-Test to measure general language proficiency. However, there are scholars who oppose this view. Cohen, Segal and Weiss Bar-Siman-Tov (1985), for example, holds the opinion that there is a close resemblance between reading an unmutilated text and processing a C-Test. To provide a deeper understanding of validity, we study relationships of C-Test results with the sub domains reading and listening comprehension, writing, orthography and language use by using confirmatory factor analysis.

The smallest units of a C-Test to be analyzed are the damaged words of each single C-Test text. It is generally agreed that these damaged words have local dependencies (Eckes & Grotjahn 2006, Harsch & Hartig 2006). However, the nature of these local dependencies and their implication on the difficulty of a C-Test text is still a matter of research (Sigott 2004, Harsch & Hartig 2006). This contribution aims to quantify and to explain the degree of local stochastic dependencies as well as the item difficulties by linguistic characteristics of both the items and the context they are integrated in. To do this, we focus on a dimensional analysis on the item level. Due to lack of local stochastic independence of the Rasch model we propose a detailed testlet model which

*Korrespondenzadressen: Alexander Robitzsch und Ina Karius, Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Luisenstraße 56, D-10117 Berlin. E-Mail: alexander.robitzsch@IQB.hu-berlin.de bzw. ina.karius@IQB.hu-berlin.de*
models dependence hierarchically, which means that it takes into account that items are nested within sentences and sentences are nested within C-Tests texts.

2 Theoretical Background: Model Assumptions Concerning Validity, Text Difficulty and Local Dependences

2.1 Validity and Dimensionality

According to Raatz and Klein-Braley (2002) a C-Test measures general language proficiency. To explicate the concept of general language proficiency they refer to the model by Bachman and Palmer, 1982 (cf. Bachman, 1990). This model divides language competence into operational competence on the one hand and pragmatic competence on the other. Raatz and Klein-Braley consider Bachman’s operational competence as the superordinate category for lexical, morphological, syntactical and graphological knowledge on the sentence level and for knowledge of cohesion and rhetorical organization on the text level. That is why they believe the general language proficiency tested by the C-Test to be very similar to Bachman’s operational competence. At the same time, they hold the opinion that the C-Test is an unsuitable procedure for measuring pragmatic competence as a separate component. Yet, Raatz and Klein-Braley claim that sufficient general language proficiency as tested by the C-Test is a basic condition for performing successfully in the area of sociolinguistic competence.

Similarly, Eckes and Grotjahn (2006) hold that first a C-Test is unidimensional and second that a C-Test measures general language proficiency. By the term general language proficiency they understand an ability comprising both knowledge and skills. To give proof, Eckes and Grotjahn draw attention to correlation studies which revealed moderate correlations with all four language skills on the one hand and with vocabulary and grammar on the other. Moreover, Eckes and Grotjahn succeeded in finding further support for the thesis of unidimensionality by using confirmatory factor analysis on scores for all language domains. Hence, Eckes and Grotjahn could disprove Cohen’s (1985) thesis that C-Tests are primarily measures of reading ability. Nevertheless, they, in agreement with Spolsky (2001), point out that general language proficiency is multicomponential, which means that every language facet has relevant specific variance.

However, Reed (2000) argues that vocabulary knowledge is an important element of general language proficiency. According to him, this lexical competence involves both knowledge of individual words and the ability to use
contextual clues to restore the damaged word. Additionally, he claims that knowledge of word morphology is particularly important if a language has complex word endings.

Having taken these validity aspects into consideration, we decided to study relationships of C-Test results with the sub domains of reading and listening comprehension, writing, orthography and language use and to, at the same time, pay special attention to the question whether the solution frequency of a mutilated word depends on its morphological markedness.

2.2 Predicting Text and Item Difficulties

2.2.1 Defining Difficulty Predictors

The definition of difficulty predictors is to be based on two types of knowledge: first, the knowledge of the major processes the item solution is based on, and second, the knowledge of manipulable task features corresponding to cognitive processing (Gorin & Embretson, 2006).

Klein-Braley (1985) approached the issue of predicting text difficulty from the perspective of readability researchers, which means that she attempted to relate the statistical characteristics of the text and the empirically determined C-Test difficulties. The two indices which she found to be the best predictors were the type-token ratio and the average length of sentence. The type-token ratio relates the number of different words in a text to the total number of words in the text. Consequently, a low type-token ratio points to a narrow range of vocabulary while a high type-token ratio indicates a wide vocabulary range. If the sentence length does not differ between texts, the text with the wider vocabulary is the more difficult one. The second predictor of text difficulty, the average sentence length, Klein-Braley considers being a rough index of syntactical complexity. Although long sentences can be simple and short sentences highly complex, Klein-Braley holds the opinion that, in general, sentence complexity is the result of more adjectives, dependent phrases and clauses, adverbial additions etc., elements which lengthen the sentence. She also refers to Kintsch and van Dijk (1978), who found out that longer sentences generally contain more propositions or ideas. Hence, a constant type-token ratio provided, a text with longer sentences is more difficult than a text with shorter sentences.

Defining difficulty predictors for C-Tests was also an aim of the analyses that went along with DESI, a study that assesses the achievements of German ninth graders in German and English. The measurement concept for the C-Test used in the DESI study and described by Harsch and Hartig (2007) differentiates
between difficulty predictors on text level and on item level. On text level the difficulty of a text is assumed to be determined by the degree of abstractness of the text’s topic, by the degree solution strategies require interpolation techniques and world knowledge and by the text complexity in terms of grammar structures. The latter was geared to curricula since there had not been found a significant interrelation between the type-token ratio and the average sentence length on the one hand and the difficulty of the C-Test text on the other hand. On item level the following difficulty predictors are assumed to be significant: the difficulty of the language phenomenon which is supposed to be measured by the mutilated word, the degree of the item’s semantic accessibility and the focus of the item on a particular linguistic dimension, ranging from lexical or grammar phenomena based on the immediate context to those based on the immediate context or the text in general to items that focus on complex linguistic and extra-linguistic configurations. All difficulty predictors were defined a priori by the DESI team and coded on the basis of either three or four levels. Afterwards regression analysis was used to identify the characteristics most likely to predict difficulties, and to determine the amount of variance explained by all predictors. Item-level predictors and text-level predictors as a whole explain slightly more than half of the observed variance.

A further phenomenon we considered to be of value for the definition of difficulty predictors is the ‘negative-tag’ hypothesis (Wason, 1959, 1961). According to this hypothesis negation is harder to process than affirmation because it requires an additional cognitive operation: first, there has to be made an affirmative representation of the matter of fact, which is called preconception, and second, this affirmative representation has to be negated. However, Wason (1965) proved that negative utterances are less hard to perceive if an appropriate preconception is given. Like Wason, Carpenter and Just (1975) found that sentence negations increase decision time in comparison to sentences without negations. Moreover, Freedle and Kostin (1993), when dealing with the prediction of TOEFL (Test of English as a Foreign Language) reading item difficulty, could prove that the number of negations in that part of the text which is crucial to identifying the correct option of a multiple-choice item is significant. They also found that the more negations in the correct and/or incorrect options the harder the item.

Klein-Braleys studies on sentence length and sentence complexity respectively formed the basis for our determination of context levels. However, we decided to also take into account the syntactical hierarchy of clauses within sentences. From the measurement construct for the C-Test used in the DESI study we especially took into consideration the difficulty predictors on item
level. Attempting to put these predictors on a more objective basis we related, the item’s semantic accessibility to its frequency of use, the item’s focus to the semantic and the grammatical context level needed for completion and the item’s difficulty to whether the word to be completed is morphologically marked or not, and whether it is located in a negated sentence or not.

2.2.2 Explaining and Assessing Difficulty Predictors

Explaining Item Difficulties with Item Attributes

There are two ways of explaining item difficulties: 1. by conducting IRT scaling, and 2. by obtaining item difficulties from IRT scaling and modeling them with a linear logistic test model (LLTM; Fischer, 1973). The LLTM is a restricted Rasch model in the sense that item difficulty parameters are expressed as a linear combination of the cognitive operations required to correctly respond to an item (Yen & Fitzpatrick, 2006). The weights that are put upon each cognitive operation signal reflect their respective importance.

Pioneering work in determining the difficulty of reading items by a set of predictors that reflect the contribution of text structure, item structure and the joint effect of both text and item structure was done by Freedle and Kostin (1993). Their research was based on data from the Test of English as a Foreign Language (TOEFL) and included the hypothesis that many of the significant variables influence reading items difficulty jointly. Freedle and Kostin first examined intercorrelations among all predictor variables in order to find out about collinearity and identified the specific variable within given clusters of variables that accounted for the most variance. Analyses of variances were used to determine main interaction effects between the predictor variables. Theses analyses already appeared to support the construct validity of the TOEFL reading section but were to be considered tentative due to significant intercorrelations and low performances in absolute magnitude. However, by finally conducting stepwise multiple regressions which estimate the multiple-R between difficulty and a set of variables, Freedle and Kostin found adequate support for both of their hypotheses.

Similar approaches have been used by, Gorin and Embretson (2006) and Sonnleitner (2008). Gorin and Embretson (2006) applied regression analyses after having quantified item features for a set of reading comprehension items in order to identify generative components for comprehension items. Likewise, Sonnleitner set up an item-generating system by using hypothesis-driven modelling of item complexity. He applied LLTM to a German reading comprehension test.
Explaining Text Difficulties with Text Attributes

We define text difficulties as the mean of all item difficulties corresponding to a text. In order to state useful relations between attributes and text difficulties, it is crucial that items are generated in the same way for each text, i.e. the items of a text are supposed to be structurally equivalent. If this is not the case, text difficulties are artificially determined by item compositions of different texts. As with LLTM for item difficulties, the analysis of text difficulties can be carried out by a multiple linear regression to determine relevant predictors.

Separating Item and Text Difficulties – a Multilevel Perspective

As previously described, separate analyses of items and texts could confound two sources causing difficulty. Therefore, in order to find out about whether the difficulty in standardized reading tests is given by the passage or the question, multilevel modelling or hierarchical linear modelling (HLM; Raudenbush & Bryk, 2002) is to be applied. This statistical technique determines the contributions made by multiple factors that are located at multiple levels of a nested data structure. Text attributes and item attributes are modelled simultaneously as predictors in a multilevel linear regression model (Gelman & Hill, 2007).

Ozuru, Rowe, O’Reilly and McNamara (2008) analysed the contribution of item and passage features to each item’s difficulty by using hierarchical linear modelling (HLM; Raudenbush & Bryk, 2002). HLM shows the degree to which several factors at different levels influence item difficulty. Ozuru et al. coded the individual questions and their answer options in terms of their relationship with the target passage on the basis of three different classification schemes. First, they classified the questions in terms of relations between passage and question. Second, they defined the abstractness of the information requested by the question. And third, they coded the quality of distractor options with regard to their falsifiability. Passage features were defined by taking into account the number of propositions, the number of words per sentence, the logarithm of the average minimum frequency and the Flesch Reading Ease score. Ozuru et al. (2008) found, among other things, that item difficulty for lower grade level students is influenced largely by passage features.

Item Characteristic: Person Attribute or Item Attribute or an Interaction of both?
When an examinee is confronted with an item, the probability of a correct response on that item is determined by the examinee’s abilities and specific item characteristics. In any case, response behaviour results from an interaction of person attributes (person side) and item attributes (item side). In the following, a clear distinction between the item side and the person side of the response behaviour has to be made (De Boeck & Wilson, 2004). While the item side provides information about the item difficulty, which is homogenous among examinees, the person side reveals that item characteristics cause different item responses by examinees. For this reason, multidimensionality emerges and, as a consequence, an examinee needs specific skills to solve the item correctly.

Buck and Tatsuoka (1998) share Freedle’s and Kostin’s assumption of language competence being a complex, multidimensional cognitive process. However, when examining second language listening comprehension, Buck and Tatsuoka prefer the so called rule-space methodology for describing the multidimensionality of the person side to multiple regressions for explaining the item side. They do so for several reasons:

1. It is difficult to obtain many significant predictors. Too many predictors and too few cases – Cohen and Cohen (1983) recommended 30 cases for each predictor – might lead to chance results or cause too many restrictions in the analysis. In addition, if item attributes are not experimentally manipulated, causal interpretations of regression parameters should be avoided.

2. Multiple regressions put the emphasis on item characteristics (i.e. on the item side) even though the person performance ability is to be of theoretical interest.

3. Multiple regressions provide information about group performances only. Hence, no information is given about the attributes specific test takers have mastered and about those involved in individual items.

According to Buck and Tatsuoka, rule space methodology provides an opportunity to avoid the drawbacks of multiple regression analysis for the following reasons:

1. Since rule-space methodology is a pattern recognition technique it does not have the same limitations on the number of attributes as multiple regression has.

2. Examinees are given an individual dichotomous score on each attribute. That way, the focus is on personal abilities rather than on item characteristics.

3. It relates attributes to examinees and specific items.
It should be noted that the rule-space methodology (RSM) proposed by Buck and Tatsuoka can be classified as a special cognitive diagnostic model (CDM; Rupp & Templin, 2008). In contrast to other recent CDM families, RSM analyses only deviations from the Rasch model. It, therefore, represents a descriptive approach. In case a multidimensional item response model with dichotomous latent variables is of interest, Rupp and Templin (2008) offer a broad variety of useful models which we favour over RSM.

We argue that the existence of many meaningful item predictors in a regression model does not justify the application of a high dimensional item response model (like CDMs) as new developments in statistical learning (Hastie, Friedman and Tibshirani, 2001) provide alternative solutions. Partial least squares algorithms, for example, simultaneously maximize the correlation with the criterion variable and reduce the number of predictors to a few latent variables (Wang, Liu and Tu, 2005). Nevertheless, it is not possible to make general reliable statements about a large number of predictors if there are too few items or texts.

2.3 Local Dependencies

Local dependencies between items of a C-Test text depend on the context they require in order to be solved. Sigott (2004) defines the amount of context that processing or strategies operate on in relation to a grammatical hierarchy of constituents and in relation to the syntactic structure of the individual sentence. The grammatical hierarchy, according to Sigott, comprises the following levels of organization: word, phrase, clause, sentence and text. Due to the hierarchical organization Sigott assumes that word-level processing is always based on less context than sentence-level processing and that sentence-level processing, in turn, does always require less context than text-level processing. However, Sigott holds the opinion that within a sentence there is no such general hierarchy which he attributes to the syntactic phenomenon of embedding.

Finally, Sigott distinguishes three item types. The first type he calls text-level items. Sigott considers an item to be a text-level item if students attempting to solve this item benefit from the whole passage. Consequently, text-level items are characterized by a significant increase in $p$ from sentence level to text level. The second type of items Sigott calls lower-level items. Subjects who try to solve these items stop benefitting from a more extended context at some level below the text. As a consequence, $p$ reaches a plateau before text level. The third and last item type Sigott calls multi-level items. According to Sigott, an item can be categorized as multilevel if the amount of context is considered increasingly helpful from one level of context to next.
Sigott did neither find a relationship between item type and lexical category nor between item type and difficulty, provided that all item types are fully contextualized in C-Test passages. However, what Sigott did find were implicational relationships between individual items of lower-level items and text-level items. This means that the likelihood of a subject to solve an item of one type can depend on the solution of an item of the other type. These relationships operate in both directions.

According to Harsch and Hartig (2007) local dependencies can be either caused by the amount of context needed to restore the items adequately, then the dependence is called text specific, or by close relationships between items in a text, then the dependence is called item specific. In order to predict these dependencies Harsch and Hartig examined a priori defined text and item characteristics from the German DESI large scale assessment. They selected two item characteristics from the DESI-project which they thought would contribute to the explanation of dependencies, that is, the semantics of items and the focus (language dimension) of items. Both item characteristic they assigned to three context levels.

The context levels for the item characteristic “semantics of items” are described as follows. At level 0 the meaning of a gap is directly and, in most cases, associatively accessible. Hence items at this level are relatively independent from other items and the text as a whole. They correspond to the item type Sigott (2004) calls lower-level items. At level 1 the meaning of a gap has to be derived from the context. These items belong to the item type Sigott calls text-level items. At level 2 an item is embedded in complex semantic relationship. Therefore, the restoration of a damaged word requires the application of a mental model.

The context levels of the item characteristic “focus of items” are as follows. At level 0 gaps have to be filled which focus on lexical or grammatical phenomena. At level 1 there are gaps which focus on lexis, grammar or phenomena of text structuring. Finally, gaps at level 2 require complex information processing. Items at this level focus on complex linguistic and extra-linguistic configurations. In addition to these two item characteristics it has been attempted to identify direct dependencies between items, so called item chains, which, for example, appear if items are part of an idiomatic expression.

Harsch and Hartig assume that higher dependencies occur in texts which contain several items at a high context level and which have several item specific dependencies. To find out about specific dependencies between the gaps they applied testlet models to C-Tests. These models implicitly assume that there is an equal extend of dependencies within and between sentences in a text.
However, our preliminary analysis shows that local dependencies vary substantially between sentences. For this reason a more detailed testlet model which models dependence of the gaps hierarchically might be more appropriate.

2.4 General C-Test Framework

The C-Test framework provided by this contribution involves students on the one hand and items on the other. Furthermore, a hierarchy on item side is assumed, which comprises the levels text, sentence, clause and item. Accordingly, there is not only an interaction between students and the C-Test-text as a whole, but also between students and each hierarchy level as well as between the hierarchy levels themselves (Figure 1). Against that background, any statistical analysis based on item level alone neglects important levels and interactions.

![Diagram of different hierarchy levels of data structure of a C-Test](image)

*Figure 1: Different Hierarchy Levels of Data Structure of a C-Test*

3 Research Questions

Our theoretical considerations led us to three major research questions. First of all, we wanted to gain a deeper understanding about the dimensionality of C-Tests, and that with regard to C-Test texts as a whole and to the items within a C-Test text. Second, we aimed at identifying and explaining difficulty predictors on text level as well as on item level. Third, assuming that each mutilated word is to be regarded as an item itself, we intended to find relevant predictors for local dependencies within a C-Test text.
4 Material

4.1 The Sample and the C-Test Texts

In the course of the measurement of the educational standards items were developed to assess the students’ competences in reading and listening comprehension, writing, orthography and language use. The assessment part on language use contained, among other items, C-Test texts. Altogether ten different C-Tests were used in a sample of 560 students of all secondary school types ranging from grade level eight to ten (14-17 years old). Every student filled in four C-Test texts which resulted in a complete balanced multi-matrix sampling design.

The C-Test texts comprised four literary texts and six scientific (non-literary) texts. Apart from few exceptions, the deletion procedure was carried out according to the canonical C-Test principle, which means that the deletion started with the second word of the second sentence and that the second half of every second word was deleted. Each C-Test text contained 20 blanks and students were given five minutes to complete the damaged words of a text.

To get further information about the C-Test texts we conducted an additional study. All in all 147 students of all school types ranging from grade eight to nine were asked to work on three parts: 1) damaged words in a text, 2) damaged words in single sentences and 3) damaged words on their own. That way we hoped to find an answer to the question whether the solution frequency increases with a higher amount of context.

Semantically and grammatically acceptable alternative solutions to the intended word were scored as correct. In contrast, spelling errors were counted as incorrect. A different scoring was applied to the mutilated words of part three of the additional study. Here, we did not differentiate between correct and incorrect answers, but took record of the respective solution provided by each student.

4.2 Coding of Item Characteristics

First of all, we determined the amount of letters of the original length of the mutilated words. Then we classified the damaged words according to their lexical category, assigned the words to content words and to function words respectively and determined whether a damaged word is morphologically marked according to gender, number, tense etc. The latter we found important with regard to grammatical relations between words. Against that background we also took into account whether a mutilated word depends on a word that is
damaged too. To give an example, in the clause “jedes Mal, wenn ich ihn besuche” (each time I come to see him), the correct grammatical form of the verb “besuche” depends on the identification of the correct pronoun “ich”.

To identify semantic relations between words we used the services provided by “Wortschatz Universität Leipzig”. That way we learned about statistically significant co-occurrences of the input word with one or several other words as well as about statistically significant right and left neighbours, meaning words co-occurring immediately to the right or the left of the input word. The degree of semantic accessibility of a mutilated word we related to its frequency of use delivered by “Wortschatz Universität Leipzig”.

With regard to possible predictors of item difficulty, we also examined whether a mutilated word had been cut at syllable border or at morpheme border and whether it is located in a sentence that contains a negation.

In order to find out the amount of context students need to restore the damaged words, we determined whether a mutilated word is located in a main clause or in a subordinate clause, and whether the corresponding clauses are part of a clause linkage (non-hierarchical relationship) or a complex clause (hierarchical relationship). Furthermore, we identified semantic hints which are helpful to solve the item and which are located outside the clause that contains the damaged word. Finally, we determined five hypothesized levels of context needed to solve the item: 1) within a clause, 2) across clauses within a clause linkage, 3) across clauses within a complex sentence, 4) across sentences within the text, 5) extratextual. This determination represents an extension of the classification of context levels which Bachman (1985) applied to the analysis of Cloze Tests and which comprise only four levels. That way, we took into account that a damaged word is nested within a clause that the clause containing the damaged word is nested either within a clause linkage or, more intricate, a complex sentence, and that sentences are nested within a C-Test text.

Moreover, we differentiated between semantic and grammatical context levels to determine the special nature of local dependencies.

5 Statistical Analyses

5.1 Dimensionality

Dimensionality was determined on the basis of Principal Component Analysis (PCA) (Hattie 1985, Tate 2003). PCAs were conducted on two levels, on text level to find out about the dimensionality of the text as a whole, and on item level to learn about the dimensionality of the items within a C-Test text. When doing so, we took into account that on the two levels scoring took place
differently. That is, while C-Test texts were polytomously scored by simple sum scores of the number of correct items for each student, the mutilated words of each C-Test text are dichotomously scored items. Accordingly, correlations of sum scores were used for PCAs on text level, and, in order to avoid artificial difficulty factors (Hambleton, Swaminathan, & Rogers, 1991), tetrachoric correlations were used for PCAs on item level. We performed separate Principal Component Analyses for each text to investigate structural interdependencies between items within texts and to keep the within dimensionality of a text distinct from the dimensionality between the texts.

5.2 IRT Variance Decomposition

We used the unidimensional Rasch model, which decomposes manifest responses on a latent (logit) IRT scale into a person effect and an item effect, in order to determine the amount of variance that can be attributed to each of the hierarchy levels shown in Figure 1. In the Rasch model, the probability of a correct solution of an item \(i\) by a person \(p\) is given by

\[
\text{logit}[P(X_{pi} = 1)] = \theta_p - b_i
\]

where \(\text{logit}\) is the logistic transformation of the response probability \(P(X_{pi} = 1)\), where \(\theta_p\) refers to the person ability and \(b_i\) denotes the item difficulty of item \(i\). In C-Tests, local dependencies correspond to secondary person traits which represent interactions between persons and texts or clauses. These are the person traits based on the respective text, the respective sentence and the respective clause.

Correspondingly, the general person parameter \(\theta_p\) in the Rasch model is substituted by a person parameter on text \((t)\), on sentence \((s)\) and on clause \((c)\) level.

\[
\theta_p + \theta_{pt} + \theta_{pts} + \theta_{ptsc}
\]

This formula represents the left part of the hierarchy in Figure 1. Beside the general performance \(\theta_p\) all interactions between the person \(p\) and the text and the sentence and the clause level are modelled. Likewise, the item difficulty \(b_i\) is decomposed into text, sentence, clause and (residual) item difficulty. It represents the right part of Figure 1:

\[
b_t + b_{ts} + b_{tsc} + b_{tsci}
\]

Putting the two pieces of the formula together, the Rasch model (1) is extended to an IRT variance decomposition model (Van den Berg, Glas, & Boosma, 2007)

\[
\text{logit}[P(X_{pi} = 1)] = \theta_p + \theta_{pt} + \theta_{pts} + \theta_{ptsc} - b_t - b_{ts} - b_{tsc} - b_{tsci}
\]
Essentially, this is a variance component model which is usually employed in Generalizability theory (G theory; Brennan, 2001). The one and only difference is that model (2) operates on logit-transformed probabilities whereas G theory ordinarily employs raw scores and therefore untransformed probabilities.

Moreover, we assumed homogenous and uncorrelated variances of all person and item random effects, that is, variances were supposed to be invariant across texts, sentences and clauses. The variances of all $\theta$- and $b$-effects indicate the relative importance of each facet in model (2).

Van den Berg, Glas and Boomsma (2007) provide a statistical framework of handling such different levels of variance components. They propose a combination of an IRT approach with Markov chain Monte Carlo (MCMC) estimation. As in van den Berg et al. (2007), our approach aims at taking into account all hierarchy levels and its interactions.

The model was estimated in WinBUGS (Spiegelhalter et al., 2003) using 5000 iterations with a Markov Chain Monte Carlo estimation procedure.

5.3 Explaining Item Difficulty
5.3.1 Explaining Item Difficulties by Linear Regression

The item difficulties \(b_i\) obtained from a unidimensional Rasch model were put as a dependent variable in a linear regression with item attributes as predictors. This linear regression can be written as

\[
b_i = \eta_0 + \sum_{k=1}^{K} \eta_k q_{ik} + \epsilon_i
\]

Accordingly, \(b_i\) is the difficulty of the \(i\)-th item, \(\eta_k\) refers to the difficulty of the cognitive operation \(k\), \(q_{ik}\) is a weight that reflects the importance of cognitive operation \(k\) in determining the difficulty of the item \(i\), and \(\eta_0\) is an arbitrary scaling constant. Here, the term cognitive operation is synonymously used with the term item attribute.

A general drawback of this approach is that the dependent variable “item difficulty” is, due to the Rasch scaling in the first stage of the analysis, prone to measurement. Ignoring this error leads to unbiased regression parameters, but causes an underestimation of the $R^2$ value. However, this disadvantage can be compensated by using so called variance known models, which are frequently used in meta analysis that include standard errors estimated from IRT scaling as the measurement error in the dependent variable (Viechtbauer, 2005; Raudenbush & Bryk, 2002). Nevertheless, if predictors are contaminated by measurement error, in principle, an underestimation of regression parameters is being found (Ruppert, Carroll, & Stefanski, 2006). In the following analyses, we
ignored measurement error in the dependent variable, but we were aware of the fact that $R^2$ values were slightly underestimated. Even though software programs such as HLM and WinBUGS allow the estimation of variance known models, we opted for ordinary least square methods for estimating the linear regression model.

A further problem, which is reflected by effects that are generally low to moderate, arises if there is a high number of predictors and an only moderate number of items. For studies with a low number of items, this problem leads to small statistical power. Furthermore, only few predictors are statistically significant, which is a circumstance that causes high standard errors. The issue is addressed by several statistical methods, among them statistical learning (regularization methods) (Hastie, Tibshirani, & Friedman, 2001) and Bayesian methods respectively (Gelman & Hill, 2007).

Another problem is the distinction between significant and nonsignificant predictors. With many predictors and a moderate number of cases, it is difficult to obtain significant effects. Therefore, effect-size-oriented approaches like variance importance effect size oriented approaches can be used to quantify the relative importance of regressors in a multiple regression model (Grömping, 2007). These statistics take into account that with correlated regressors it is not possible to just break down the model $R^2$ into shares from individual regressors. The approach applied to our data set was the statistic proposed by Lindeman, Merenda and Gold (1980). The LMG statistic is to be interpreted as the average squared semi-partial correlation coefficient for a predictor; the averaging takes place over all possible permutational orderings of the predictor variable within the set of all predictors. That way, even nonsignificant predictors could become important because high portions of variances of one predictor can be shared with others.

5.3.2 Explaining Item Difficulties by Multilevel Linear Regression

In multilevel modelling (Gelman & Hill, 2007), the nesting of statistical units within different levels is taken into account for model building. In the case of a C-Test, items are nested within clauses, clauses are nested within sentences and sentences are nested within texts. Item difficulties can be decomposed in exactly the same way as in the IRT variance decomposition model with the empty model

$$b_{icst} = \eta_0 + u_i + u_{st} + u_{cst} + e_{icst}$$

where $\eta_0$ is the grand mean item difficulty, $u_i$ denotes the random text effect, $u_{st}$ the sentence effect $s$ within text $t$, $u_{cst}$ the effect of clause $c$ in sentence $s$ and
text $t$ and $e_{icst}$ the residual effect of the item. The variances of the different levels taking together with the random effects provide insight to the importance of the levels with respect to differences in item difficulties.

In a multilevel linear regression model with item predictors $q_{ik}$, the item difficulty of item $i$ in clause $c$, sentence $s$ and text $t$ can be written as

$$b_{icst} = \eta_0 + \sum_{k=1}^{K} \eta_k q_{ik} + u_c + u_s + u_{est} + e_{icst} \quad (4)$$

In principle, the predictors $q_{ik}$ can be text or sentence characteristics, too. Nevertheless, it needs to be pointed out that we do not avoid the specification of different classes of predictors in formulas. Quite the contrary is true as, in comparison to linear regression, random effects of texts, sentences and clauses are included so that a more complicated error structure of item difficulties is being modelled. The size of the variance of the random effects in the models (3) and (4) can be used to define similar $R^2$ measures for each level. Here, the reduction of variance of one level in model (4) in relation to the variance of the empty model (3) determines how much variance is explained by the predictors introduced in the model. Furthermore, a level-wise $R^2$ measure is defined.

In the following analyses, like in linear regression, we ignored the measurement error in the dependent variable. The multilevel regression models were estimated with the R lme4 package.

### 5.4 Assessing and Explaining Item Dependence

Local dependence was investigated by using the $Q_3$ statistic (Yen, 1984). The $Q_3$ statistic estimates the degree of violation of local independence in an unidimensional IRT model (e.g. the Rasch model) for a pair of items in a test. As the $Q_3$ statistic of item pairs is defined as the correlation between the corresponding residuals, all $Q_3$ statistics should be approximately equal to zero if the Rasch model holds.

An approach to obtain local independence despite locally dependent items is provided by testlet response theory models (Wainer and Kiely, 1987) with the term testlet being understood as a group of items that administered and scored as a unit. Bradlow, Wainer and Wang (1999) added a random component to the 2PL model to represent an interaction between examinee and testlet. Wainer, Bradlow and Du (2000) proposed a similar extension of the 3PL model, thus also taking into account the guessing parameter. That way, various dimensions of variance are opened up (Wainer, Bradlow and Wang, 2007).
In order to predict local dependencies, we applied a linear regression to the $Q_3$ statistics of item pairs of the respective C-Test text and again determined the variable importance.

6 Analysis Outcomes

6.1 Dimensionality

The Principal Component Analysis on text level showed that more than 60% of the variance can be explained by the text itself. In other words, due the PCA on text level a clear unidimensionality can be assumed. In contrast, separate PCAs within the C-Test texts, based on a tetrachoric correlation matrix, revealed at least a two-factor solution. There was only one C-Test text that required a four-factor solution. Unlike the other texts, this text is characterized by a very old-fashioned vocabulary. To conclude, whether unidimensionality can be assumed or not depends on the level of statistical analysis, that is whether one chooses a between or a within approach of statistical analysis.

6.2 Variance Decomposition

The IRT variance decomposition model decomposes person abilities and item difficulties which are displayed in Figure 2. On the person side the highest amount of variance, that is 53.6%, is to be attributed to the student. The second highest amount of variance, that is 24.4%, is explained by the interaction between student and clause. Moreover, the interaction between text and sentence on the one hand and the student on the other hand amounts to about 10% of the total variance. This means that on the person side, apart from the examinee’s ability itself, most information is provided on the lowest hierarchy level.

On the other side, a decomposition of item difficulty variance shows that 60.7% of the variance can be ascribed to the item level. However, it is not, as one might expect now, the next higher context level “clause” that contributes the second highest explanation of variance, but the sentence level. And last but not least, the texts do not strongly differ in their item difficulties due to the low proportion of variance of 4.6%.
6.3 Predicting Item and Text Difficulties: A Multilevel Perspective

The prediction of item and text difficulties was done on the basis of linear regression. Regressors assumed to be relevant in terms of item difficulty were the word categories adjective, adverb, article, conjunction, particle, pronoun, noun and verb, number of letters, the context required to solve the item, morphologically markedness of the mutilated word, a cut at syllable border as opposed to a cut at morpheme border, sentence negation, the occurrence of the mutilated word in a simple main clause (the main clause is not part of a clause linkage or a complex clause) as well as dependence of a mutilated word on another mutilated word, the word frequency of the mutilated word, the number of gaps and the number of words within the clause containing a mutilated word.

The results of the linear regression are displayed in Table 1. In the regression model each variable possesses an individual LMG parameter. Since the variable “word category” is a categorical variable comprising all word categories listed up as factors, there is only one LMG parameter value estimated for this variable. “First” in $R^2$(first) refers to the explanation of variance if the corresponding variable is taken first into the regression model and if it is the only predictor.

The table shows that the variables likely to make the biggest contribution towards predicting item difficulty are the variables word category, word frequency, sentence negation and number of letters. In contrast, the context levels above clause linkage and the morphologically markedness of a mutilated word appear to play only a minor role with regard to the difficulty of the item.
Especially the latter goes against our previous assumptions. However, what appears worth having a closer look at is the result concerning sentence negation. As stated before, the table shows that an item which is located in a negated sentence tends to be more difficult than an item in an affirmative sentence. This result confirms Wason’s ‘negative-tag’ hypothesis, which states that affirmation is harder to process than affirmation. Taking into account that the context levels above clause linkage do not significantly contribute to the item’s difficulty, which means that students hardly use the context outside the sentence, it seems appropriate to conclude that even though a C-Test text as a whole might provide a so called appropriate preconception (Wason, 1965) it has only little effect on the item’s difficulty within the respective sentence.
### Table 1: Predicting Item and Text Difficulty by a Linear Regression

The empty model in a multilevel analysis estimates the variance components of different levels (Table 2). The text level variance calculated by this analysis differs from that gained from the IRT variance decomposition model. This is due to several reasons. First, the estimation method used in lme4 (maximum likelihood) leads to an underestimation of variance components in comparison to the MCMC method in the variance decomposition model. Second, measurement error of the dependent variable of the regression model is neglected.

In multilevel regression analysis, the same predictors as in linear regression were used and the regression coefficients of both models are of comparable size.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Regression Estimate</th>
<th>Std. Error</th>
<th>LMG in %</th>
<th>R² (first) in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word category: adjective</td>
<td>-0.79*</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: adverb</td>
<td>1.09*</td>
<td>0.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: article</td>
<td>1.70*</td>
<td>0.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: auxiliary verb</td>
<td>1.91*</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: conjunction</td>
<td>0.32</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: particle</td>
<td>0.20</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: preposition</td>
<td>2.06*</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: pronoun</td>
<td>1.20*</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: noun</td>
<td>0.02</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word category: main verb</td>
<td>0.80*</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of letters</td>
<td>0.10</td>
<td>0.16</td>
<td>3.8</td>
<td>7.3</td>
</tr>
<tr>
<td>Context level above context level 2</td>
<td>0.42</td>
<td>0.28</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Morphological markedness</td>
<td>-0.26</td>
<td>0.27</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Cut at syllable border</td>
<td>-0.08</td>
<td>0.21</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Sentence negation</td>
<td>0.94*</td>
<td>0.21</td>
<td>6.6</td>
<td>5.5</td>
</tr>
<tr>
<td>Simple main clause</td>
<td>-0.60*</td>
<td>0.23</td>
<td>3.2</td>
<td>3.5</td>
</tr>
<tr>
<td>Dependence on a mutilated word</td>
<td>0.57</td>
<td>0.32</td>
<td>1.5</td>
<td>1.9</td>
</tr>
<tr>
<td>Word frequency</td>
<td>-1.15*</td>
<td>0.19</td>
<td>15.8</td>
<td>18.0</td>
</tr>
<tr>
<td>Number of gaps within the clause</td>
<td>-0.21</td>
<td>0.21</td>
<td>0.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Number of words within the clause</td>
<td>0.20</td>
<td>0.22</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Σ</td>
<td></td>
<td></td>
<td>43.5</td>
<td>47.2</td>
</tr>
</tbody>
</table>

* : p < 0.05
In addition, in the empty model almost none of the variance of the item difficulty can be ascribed to the task level. Contrary, most of the variance corresponds to the residual item level. What is striking is that 37% of the variance is explained by the predictors, whereas practically all variance on sentence level ($R^2 = 100\%$) is explained by sentence level predictors. Aggregated predictors, which are calculated as means of predictor values at higher levels, could be included in the multilevel regression to study contextual effects and to separate effects with and between tasks (sentences or clauses) (Raudenbush & Bryk, 2002). These analyses will be carried out with a larger sample of texts and students. Moreover, multilevel models with random slopes, which describe (for example) different item predictors effects in different tasks, can be of interest, too.
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Empty Model Regression Estimate</th>
<th>Prediction Model Regression Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word category: adjective</td>
<td>-0.67*</td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>Word category: adverb</td>
<td>0.29*</td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td>Word category: article</td>
<td>0.84*</td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>Word category: auxiliary verb</td>
<td>1.15*</td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>Word category: conjunction</td>
<td>-0.50</td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>Word category: particle</td>
<td>-0.79</td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>Word category: preposition</td>
<td>1.17*</td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>Word category: pronoun</td>
<td>0.50*</td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td>Word category: noun</td>
<td>-0.64</td>
<td></td>
<td>0.35</td>
</tr>
<tr>
<td>Word category: main verb</td>
<td>-0.06*</td>
<td></td>
<td>0.39</td>
</tr>
<tr>
<td>Number of letters</td>
<td>-0.09</td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>Context level above context level 2</td>
<td>0.40</td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Morphological markedness</td>
<td>-0.27</td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Cut at syllable border</td>
<td>-0.10</td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>Sentence negation</td>
<td>0.90*</td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Simple main clause</td>
<td>-0.59*</td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>Dependence on a mutilated word</td>
<td>-0.66</td>
<td></td>
<td>0.32</td>
</tr>
<tr>
<td>Word frequency</td>
<td>-1.08*</td>
<td></td>
<td>0.17</td>
</tr>
<tr>
<td>Number of gaps within the clause</td>
<td>-0.27</td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Number of words within the clause</td>
<td>0.25</td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>Variance components</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R² in %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>0.00</td>
<td>0.00</td>
<td>76</td>
</tr>
<tr>
<td>Sentence</td>
<td>0.21</td>
<td>0.00</td>
<td>100</td>
</tr>
<tr>
<td>Clause</td>
<td>0.63</td>
<td>0.45</td>
<td>29</td>
</tr>
<tr>
<td>Residual</td>
<td>2.07</td>
<td>1.39</td>
<td>33</td>
</tr>
<tr>
<td>Σ</td>
<td>2.91</td>
<td>1.84</td>
<td>37</td>
</tr>
</tbody>
</table>

* : p < 0.05

Table 2: Prediction of Item Difficulty in a Multilevel Model
6.4 Assessing Local Dependence

Local dependence was assessed by making use of $Q_3$ statistic (Yen, 1984) after fitting the Rasch model for all 200 items of the 10 C-Test texts. The image plot (Figure 3) displays all pairwise correlations of residuals and reveals that adjacent blanks (those that are within a C-Test text) are higher correlated than blanks with more space inbetween (blanks that are located in different C-Test texts). Different correlations are displayed in the image plot by different gray contrasts. While the correlation of residuals within C-Test texts (diagonal) averages out at .053, the correlation of residuals between C-Test texts is at average -.026. This means, that there is more local dependence within C-Test texts than between them.

![Image Plot of $Q_3$ Statistic](image.png)

*Figure 3: Image Plot of $Q_3$ Statistic*

Correspondingly, $Q_3$ values represent a function of word distance (Figure 4). Analyses show that the correlation of $Q_3$ statistics exponentially decreases...
the longer distances \((d)\) between the blanks becomes. The distance \(d\) is transformed by dividing one over the square root of \(d\) which corresponds to the power \(a\) of \(d\) which approximately maximizes the correlation of \(d^a\) with \(Q_3\). If the distance between words equals 1 (which is equal to the transformed distance of 1), the \(Q_3\) value adds up to .16. Looking at the other extreme, the plot shows that if the distance between words equals 20, the \(Q_3\) value is about -0.02.

![Figure 4: Relationship of Local Dependence (\(Q_3\) Statistic) and Transformed Word Distances](image)

**6.5 Predicting Local Dependence by Item Covariates**

The prediction of local dependence was done by applying a linear regression to the local dependencies gained from \(Q_3\). Putting it differently, the relative variable importance was measured (Table 3). Covariates assumed to be relevant for predicting local dependence are the transformed distance, the condition that the respective items are within the same clause, the condition that the respective items are within the same sentence, the condition that the respective items are structure words and located in different clauses, the sum of the log word frequencies of both words and the product of the log word frequencies of both words. With regard to interactions within the same clause, we also included the predictors number of gaps, number of words, the condition that both items are structure words and the condition that both words are identified on context level 1.
The table shows that the variables likely to make the biggest contribution towards predicting local dependence are the variables “transformed distance” and “items within same clause”. With regard to interactions within the same clause the most important variable for predicting local dependence is “both words of context level 1”. Since context level 1 is defined as the level of context within a clause, this result supports our initial theoretical assumption. Moreover, the number of gaps within a clause also seems to play a major role in determining local dependences, which means that the more mutilated words a clause contains the higher are local dependences within a clause. This might go along with the circumstance that the more gaps a clause contains the higher the probability that a mutilated words depends on a word that is mutilated too. Taking into account that local dependence is at highest within a clause, it seems appropriate to conclude that information outside the respective clause are hardly needed in order to complete the mutilated word within that clause. This result confirms the results from our analysis of item difficulty according to which context levels above clause linkage (context level 2) only have a minor influence on the difficulty of the mutilated word. The inclusion of dummy variables of all texts as predictors indicates that the different tasks induce different amount of local dependencies.
Table 3: Prediction of Local Dependency (measured by the $Q_3$ Statistic) using a Linear Regression Model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Regression Estimate</th>
<th>Std. Error</th>
<th>LMG in %</th>
<th>$R^2$ (first) in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Test G315, …, C-Test G312</td>
<td>-0.071* 0.034*</td>
<td>0.01 0.01</td>
<td>6.2 5.7</td>
<td></td>
</tr>
<tr>
<td>transformed distance</td>
<td>0.142*</td>
<td>0.01</td>
<td>9.9 19.2</td>
<td></td>
</tr>
<tr>
<td>items within same clause</td>
<td>0.067*</td>
<td>0.01</td>
<td>6.7 14.4</td>
<td></td>
</tr>
<tr>
<td>items within same sentence</td>
<td>0.030*</td>
<td>0.01</td>
<td>3.4 7.9</td>
<td></td>
</tr>
<tr>
<td>sum of log word frequencies of both words†</td>
<td>0.009*</td>
<td>0.01</td>
<td>0.7 0.5</td>
<td></td>
</tr>
<tr>
<td>product of log word frequencies of both words†</td>
<td>0.006*</td>
<td>0.00</td>
<td>0.1 0.0</td>
<td></td>
</tr>
<tr>
<td>both structure words and in different clauses</td>
<td>-0.011</td>
<td>0.00</td>
<td>0.3 0.5</td>
<td></td>
</tr>
<tr>
<td><strong>Interactions within same clause</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of gaps†</td>
<td>0.023</td>
<td>0.02</td>
<td>0.3 0.1</td>
<td></td>
</tr>
<tr>
<td>number of words†</td>
<td>-0.058*</td>
<td>0.02</td>
<td>0.6 0.2</td>
<td></td>
</tr>
<tr>
<td>both structure words</td>
<td>-0.012</td>
<td>0.01</td>
<td>0.6 2.4</td>
<td></td>
</tr>
<tr>
<td>both words of contextual level 1</td>
<td>0.030*</td>
<td>0.02</td>
<td>0.8 2.7</td>
<td></td>
</tr>
<tr>
<td>Σ</td>
<td></td>
<td></td>
<td>29.6 53.6</td>
<td></td>
</tr>
</tbody>
</table>

*: p < .05 ; †: z-standardized variables

7 Discussion

The results of this article show that the violation of local independence is substantial and permits the usual application of Rasch model of C-Tests on the item level from the perspective of a good model fit. On the other side, the consideration of different levels within a C-Test under a multilevel perspective could give fruitful insights of relevant student interactions with test material and sharpens the C-Test construct in the sense of determining relevant features of deviations from unidimensionality. The proposed multilevel linear regression analysis could be a reasonable starting point to investigate causes of item difficulties on different levels. Further studies should include more experimentally manipulated variations of predictors on all levels and more tasks (in total numbers) to allow more reliable inferences on the task level.

From a methodological perspective, the linear regression and multilevel linear regression could be estimated in an integrated model where abilities and difficulties are jointly modelled like in the IRT variance decomposition model. The prediction of local dependence can be performed with regressions on dependency parameters obtained in testlet models instead of using regression on $Q_3$ statistics, which are also prone to measurement error and dependent from
item difficulties in the item pairs. Nevertheless, the main statements of the models should be very similar to the ones obtained with $Q_3$ statistics.

Taking everything into consideration, our approach might not answer the fundamental question about the validity of the of the C-Test’s construct, which is the proper degree of local dependency of a “good C-Test”? Moreover, it has to be discussed whether it is appropriate to manipulate item and task predictors in order to maximize predictive validity in terms of general language proficiency?

8 Literature


