INTERFACE
–JOURNAL OF EUROPEAN LANGUAGES AND LITERATURES
ISSN: 2519-1268

Issue 8 (Spring 2019), pp. 29-67
DOI: 10.6667/interface.8.2019.76

Gateways, Placements, and Grouping: Automating the C-Test for Language Proficiency Ranking

WOLFGANG ODENDAHL
National Taiwan University

Abstract

Foreign language departments with the goal of advanced literacy require optimizing student learning, especially at the initial stages of the program. Current practices for admission and placement mainly rely on students' grades from previous studies, which may be the main reason why intra-group language proficiency often varies dramatically. One essential step for creating an environment that enables students to progress according to their skill level is the development of assessment procedures for admission and placement. Such assessment must prominently include proficiency in the target language. This article promotes the incorporation of an automated C-test into gateway and placement procedures as an instrument that ranks candidates according to general language proficiency. It starts with a review of the literature on aspects of validity of the C-Test construct and contains an outline of the functional design of such an automated C-Test. The article highlights the economic benefits of an automated C-Test platform and the central role of proficiency-based student placement for the success of programs aiming to develop advanced literacy in a foreign language. The findings implicate that developing and using the outlined C-Test platform has the potential to increase student achievement in advanced foreign language instruction significantly.

Keywords: Admission Gateway Testing; Language Proficiency Testing; Homogeneity; Grading; Differentiated Language Instruction; Ability Grouping

http://interface.ntu.edu.tw/

Gateways, Placements, and Grouping: Automating the C-Test for Language Proficiency Ranking

One of the pervasive challenges of foreign language classrooms is the diversity of students' levels of language proficiency (Sun, Fan, & Chin, 2017, p. 249; Cohen & Lotan, 2014, Chapter 2; Daud, Daud, & Kassim, 2005, p. 3 ff. Dutcher, 2018, Chapter 6; Harmer, 2010, pp. 14–19; Reese, 2011; Wunsch, 2009). MA programs admit students from different universities according to their BA grades, and advancing through the stages of any program requires a passing grade in the previous level (Mozgalina & Ryshina-Pankova, 2015, p. 347). However, grade-calculation principles vary between institutions and even between teachers in the same institution (cf. Alderson, 2017; Alderson, Brunfaut, & Harding, 2015, p. 242 ff. Xie, 2015; Gamaroff, 2000). Grades therefore do not reflect an objectively comparable selection factor.

For all students to achieve the same educational goal, proficiency-heterogeneous classrooms require differentiated educational measures according to their different proficiency levels (Tomlinson, 2014; Tomlinson & Imbeau, 2014; Stöger & Ziegler, 2013, p. 7). Provisions against unintended heterogeneity in the admission process, such as requiring a certified advanced Common European Framework of Reference for Languages (CEFR) proficiency level,[1] often fail due to issues with inter-rater reliability (e.g. Huang, Kubelec, Keng, & Hsu, 2018; Deygers, Van Gorp, & Demeester, 2018; Díez-Bedmar, 2012). The results of those tests are considered valid and reliable, but there are significant differences between a passing grade and full points (Dunlea & Figueras, 2012). Furthermore, the question of equivalence of test results and the comparability between testing facilities is still under debate (Alder-

1 In the field of German as a Foreign Language (GFL), the established testing institutions – Goethe, Telc, TestDaf, ÖSD – offer summative German GLP assessment according to the CEFR as a paid service. These widely recognized benchmark-assessments test the four foundational language skill areas of reading, writing, listening, and writing in multi-hour sessions.

son, 2017; Xie, 2015; Newbold, 2012; Knapp, 2011, p. 652). Therefore, whenever language skills are a factor, grouping according to the results of an in-house general language proficiency (GLP) test is a viable alternative (Norouzian & Plonsky, 2018, p. 396). Administering a reliable language test, which produces a ranked list of all candidates according to their current overall language proficiency, would be a consistent instrument for informing admission decisions regarding a candidate's language skills and allow grouping admitted students accordingly (cf. Norouzian & Plonsky, 2018, p. 396; Mozgalina & Ryshina-Pankova, 2015). Selecting the most suitable candidates according to GLP could simultaneously raise admission fairness and have positive effects on students' achievement. Students' general language proficiency is the key study tool for any foreign language-related program, regardless of the program's specialization, be it literary studies, teacher training, or translation studies.

One reason why foreign language departments, especially smaller ones teaching other than the mainstream languages, shy away from testing candidates in-house may be rooted in test economy:[2] Designing a reliable and valid test each year is a very demanding and specialized task. Deploying and grading tests also consumes resources, even if – for the sake of economy – they do not contain lengthy written parts.

This article proposes an online GLP test as an economical solution to the gateway problem. Taking its cue from the writing section of the TestDaF, one of the most prominent German-language tests,[3] it outlines

2 Test economy here refers to the cost/benefit ratio in testing. For detailed definitions of terms used in this article, refer to section 1.4 below.

3 The 'Test Deutsch als Fremdsprache' (TestDaF) is a standardized language test for foreign students applying for entry to an institution of higher education in Germany. Non-native speakers planning to study at a German university have to pass either TestDaF or DSH. TestDaF is offered in 96 countries worldwide and counted more than 44,000 test takers in 2016 (Norris & Drackert, 2018, p. 149). Normally level 4, the second of three levels (with 3 being the lowest and 5 the highest), is sufficient for passing the language requirements of German universities (Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V. & TestDaF-Institut, 2017b). The same institution develops the 'Online Language Placement Test' (onSET formerly onDaF), an online placement test based on the C-Test. Currently available for German and English, it aims to "offer online placement tests for a whole range of modern languages at as many university language centres as possible" (Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V. & TestDaF-Institut, 2017a).

the design of an automated C-Test platform for gateway and placement testing. Relying on an expandable corpus of texts, the proposed platform produces a vast quantity of different C-Tests for any number of candidates while minimizing administrational time and effort. Supervision and test administration modalities can be adapted according to the stakes in the outcome. High- and medium-stakes testing, such as gateway and placement testing, need to have some level of supervision as precaution against cheating. Individual students may take the test unsupervised at their leisure as an economical screening-test.[4]

1 Introductory Considerations

1.1 Research Question

In order to solve the above outlined problem of test economy, the question this article aims to answer is how to design an automated test that improves the existing gateway and placement system of foreign language departments. It departs on the proposition that in a situation where language proficiency is the major factor for students enrolled in an advanced language program being able to graduate, a test for general language proficiency is an adequate basis for admission to and for placement in language programs. Bachman (2005, pp. 18–21) states that as long as the test's ability to measure what the program requires, testing is a superior alternative to relying on previous grades. This paper's base hypothesis is that test economy is the major factor preventing institutions from using in-house generated test data (Bachman, 2005, p. 24). Administering and grading tests is work intensive and time consuming – especially if one needs to create a new test every time the program admits or places students. Only extensive, long-term qualitative studies may provide an answer to the question whether institutions and teachers will actually revert to testing if the tests are economical, i.e. easy to administer and effortless to grade. Here the focus is set on the question:

4 Before taking an expensive official test, some students wish to confirm their proficiency with the results of an independent and objective C-Test.

How can language proficiency ranking be automated to the point of effortless efficiency?

1.2 Method

Based on the hypothesis that institutions avoid using their own tests in gateway and advancement decisions for reasons of test economy, this article analyzes the bottlenecks in testing by means of literature review. The problematic points are then individually resolved by sketching out an automated test for general foreign language proficiency.

The aim is to outline the design and functionality of an automated C-Test platform for language proficiency testing for the purposes of gateway testing and student placement in institutions for advanced foreign language studies. Secondary purposes, such as self-evaluation, are not the focus of this article.

The C-Test is a summative test for GLP and not intended for diagnostic or formative purposes. An automated version of the C-Test solves problems of test economy, thereby allowing foreign language institutions to do their own testing. Being able to rank students according to identical standards allows institutions to admit candidates who fit their requirements and place them in groups with peers who show similar language proficiency. This study relies on secondary literature when concerned with the theory of testing in general and the C-Test as measure for GLP in particular. The fundamental outline of the technical aspects of the proposed platform is based on general computer and web programming facts and the author's personal coding experience.

1.3 Outline

This article consists of three parts. The Literature Review summarizes and discusses the theory of general language proficiency testing, the evolution of the C-Test, and its validity as a test of GLP. After intro-

ducing the historical development from cloze to C-Test in terms of construct principles, this section will make the argument that the C-Test is well-fitted for automation, and that with regards to test economy, a web-based, fully automated C-Test platform is an excellent solution for gateway and placement testing (Eckes & Grotjahn, 2006, p. 290). It also discusses the argument for in-house testing as opposed to relying on previous grades.

The second part outlines the functional design of an automated online C-Test platform with usage examples for its practical application. It will give two examples for using the platform, one for institutional testing and the other for individual self-administered assessment, followed by an outline of how academic research may profit from it. The aim for this section is to invoke examples of general usage in institutions teaching foreign languages; readers may be inspired to infer approaches fitting their unique testing needs. The third section discusses several technical details crucial to the functioning of the platform.

1.4 Operational Definitions

This article frequently uses the key terms Ranking, Tracking, and Test Economy, which in different contexts might have different interpretations and need clarification.

Ranking: The C-Test construct is designed to measure general language proficiency (see sec. 3). Its output can take the form of an absolute statement (see Figure 3: Results of a Self-administered C-Test in sec. 4) or a list that ranks candidates hierarchically in relation to their respective results. The aim of ranking candidates according to their general proficiency in the target language is to draw inferences about each candidate's relative performance in order to select and group students with similar GLP levels. This kind of test is summative in nature, and it is not part of the teaching process (McNamara, 2011, p. 613; Huhta, 2008, p. 473). Contrary to the binary logic of a benchmark test such as level A, B, or C according to CEFR, where the goal is to determine whether

or not the candidate's skills conform to a predetermined standard, ranking candidates according to their skill-level demands an open-ended rating scale of assessment (Jones & Saville, 2008, p. 498; Knapp, 2011, p. 646).[5]

In gateway testing situations where all candidates fulfill the basic requirements and the number of admission slots is limited, ranking allows administrators to fill these slots with candidates who are most proficient in the tested skills. For placement purposes, a ranked list allows administrators to decide on homogeneous or intentional skill-heterogeneous groups. Therefore, in the context of this article, ranking denotes the presentation of a test outcome in the form of an ordered list with the highest scores on top.

In contrast to Steenbergen-Hu et al. (2016, p. 850), this article does not limit the purpose of placement ranking for ability grouping to produce homogeneous learner groups. Planned heterogeneity or "cluster-grouping" (ibid., 851) may be very effective under certain conditions (cf. Odendahl, 2016) and improve peer-assisted learning results (cf. Nesmith, 2018; Odendahl, 2017; Smith, 2017; Tempel-Milner, 2018). Unplanned heterogeneity in language proficiency may result from inadequate selection procedures and is generally undesirable in foreign language classes. Therefore, regardless of the grouping goal, it is imperative to have reliable evidence to base the grouping on.

Ability grouping and tracking: In placement practice, the terms ability grouping and tracking are often used synonymously. Some academics use the term tracking in reference to distributing students into different classes, reserving the term grouping for placement within classes (Loveless, 2013, p. 13). The usage adopted by this article concurs with researchers such as Tieso, who define tracking as "[placement of students] into streams or tracks from which they never escape" (2003, p. 29). By contrast, ability grouping is a more flexible, non-permanent form of distributing students in homogeneous learning groups (Steenbergen-Hu et al., 2016, pp. 850–851). Tracking, under the name

5 For a detailed discussion of proficiency scales and problems of validation see North (2000).

of ability grouping, was widely practiced in U.S. school systems from the 1960s to the 1990s, when vocal criticism from equity advocates,[6] most notably Robert Slavin (1987, 1990, 1993) and Jeannie Oakes (1985, 1986a, 1986b), contributed to its disuse in the public school system. Very similar practices, however, remain in use under different aliases, such as "streaming, setting, sorting, classroom organization or composition, and classroom assignment" (Steenbergen-Hu et al., 2016, p. 856).

Test economy: The overall ratio between cost spent for testing and its benefit is here referred to as test economy.[7] The cost of testing includes aspects of money, time, and effort spent on creating, administering, and grading (cf. Moosbrugger & Kelava, 2011, p. 21; Gnambs, Batinic, & Hertel, 2011, p. 8; Hornke, 2006, p. 434). Furthermore, testing takes not only a toll on teachers and administrative staff, but also on the candidates, who take the test and deal with its outcome. These expenditures have to be matched by the benefits from its outcome. It is therefore of utmost importance to determine exactly what purpose a given test should serve before proceeding. Here, the intended outcome is to find the most fitting students to join an advanced program, thus heightening the chances that participants will be able to graduate. While it seems worthwhile to expend a lot of effort on such an important goal, the most economical ratio always is to spend as little as necessary in order to gain as much as possible. With a favorable ratio, the same test can also be deployed for secondary goals, such as grouping students into skill-homogeneous classrooms. These requirements build on and conform in essence with Lienert and Raatz (1994, p. 12), who define a test as economical if a) its administration requires little time, b) it consumes little material, c) it is easy to handle, d) it may be administered as a group test, and e) its grading is fast and convenient. Hornke (2006, p. 434) enumerates the stakeholders in economical testing as the candidates, the department, the administering staff, and designers.[8]

6 Advocates for equity, or social equality "have opposed the practice [of ability grouping/tracking] on principle as undermining social goals of equity and fairness in our society" (Braddock & Slavin, 1992, p. 5). Relying on Deutsch (1975), Messick (1989, p. 86) discusses the multiple sources of potential injustice which may be salient in any particular setting.

7 For an overview of the impact of language testing and washback effects, see Shohamy (2017); for the impact of computer technology on testing see Chappelle & Voss (2017).

8 Hornke, departing from a standardization standpoint with ISO norms in mind, uses the terms

Therefore, in judging the economic properties of the proposed C-Test platform, the key metrics are availability, reliability, affordability, and convenience.

- In terms of availability, a test needs to be always accessible, using as few tools as possible. For example, an online version of a test will score higher in availability than the paper equivalent, which has to be physically carried around and distributed. It is more available than a specialized computer program or app, which are custom-made for one platform, such as Windows©, Macintosh©, Linux©, iOS© or Android©. This class of computer programs need advance installation on a machine present at the time of testing. In addition, a mobile-accessible user interface scores better on availability than one that can only be accessed on larger computer screens.
- Reliability includes the specialist term from testing research as likelihood of getting the same result when testing several times under the same conditions (cf. Feldt & Brennan, 1989; Haertel, 2006; Dunlea & Figueras, 2012; Newbold, 2012). Here, it also refers to the stability of the test medium – an unreliable computer program or wet paper tests would get a lower score.
- The criterion of affordability includes monetary expenses, labor cost, time, and effort. Affordability affects both administrators and candidates.
- Convenience covers aspects of affordability and availability and applies to administrative aspects as well as the candidate's perspective.

Test economy is a major factor when considering in-house testing. In a medium-stakes situation like admission for a master's program, the admitting institution might be willing to spend considerable time and effort to devise their own test for establishing candidates' language proficiency. However, in order to prevent leaking, such a test would have to be modified for each use, which poses a major drain on resources. This

client and contractor instead of department and administrating staff, and the term researchers instead of designers. In his words, the candidates do not want to be unduly strained with testing, the client does not want to spend more money than necessary, the contractors needs to keep their efforts in relation to a reliable outcome and the researchers need to optimize the test according to their grants (Hornke, 2006, p. 434).

holds true for gateway and placement testing alike.

Homogeneous versus heterogeneous grouping: The practice of placing students homogeneously according to their current proficiency is the topic of an extensive ongoing discussion (Brulles, Saunders, & Cohn, 2010; Henry, 2015; Kim, 2012; Missett, Brunner, Callahan, Moon, & Azano, 2014; Nesmith, 2018; Robinson, 2008; Schofield, 2010; Tempel-Milner, 2018; Vogl & Preckel, 2014). The dispute whether or not homogeneous ability grouping benefits student achievement is mainly a controversy about educational values revolving around equality and is still unresolved. In an oft-cited meta-study, Slavin (1990) concludes there is no evidence for positive or negative effects of ability grouping on student achievement. However, Slavin's sources all use academic achievement as the norm of measurement instead of independent testing with compatible standards. Newer studies, still based on academic achievement, suggest a significant impact of homogeneous ability grouping on students' academic achievement (cf. Steenbergen-Hu et al., 2016).

The most-cited risk of homogeneous placement is the phenomenon of fixed tracking, where students are stuck with a label after one-time placement. On the positive side, as Oakes stated early on, tracking might prevent "less-capable students [from suffering] emotional as well as educational damage from daily classroom contact and competition with their brighter peers" (1986a, pp. 3–4), a claim repeated up to the present time (eg. Glock & Böhmer, 2018, p. 244). On the other hand, Oakes found that "literature suggests that students at all ability levels can achieve at least as well in heterogeneous classrooms" (1986a, pp. 3–4), which also applies to recent research (eg. Francis et al., 2017; Hornstra, van der Veen, Peetsma, & Volman, 2014). A large meta-analysis on the effects of ability grouping indicates that subject grouping and special groups for the gifted have positive effects on the performance of gifted students (Steenbergen-Hu et al., 2016). Negative effects of ability grouping have been shown for low attaining (Francis et al., 2017), socio-economically disadvantaged (Henry, 2015), or ethnic minority (Glock & Böhmer, 2018) students. Permanent tracking not only influences stu-

dents' self-esteem, but also leads to varying teacher expectations, thus perpetuating the initial placement in a vicious cycle (Bernhardt, 2014; Harris, 2012; Oakes, 1985, p. 8).

With regard to equity and equality in education, increasing placement frequency helps avoid the negative effects of tracking. Re-placing students in frequent intervals means increased mobility between groups and counters the negative effects of tracking on student performance. Robinson (2008) found that level-appropriate instruction as the result of homogeneous grouping significantly helped with reading literacy instruction. The same may also apply to the field of foreign languages. With homogeneous groups, administration and teachers can customize their learning environment and progression speed to their students' current proficiency. Differentiated syllabi for parallel classes cater to the student's needs and allow for focused contents. When aiming for teaching efficacy, regardless of one's views on heterogeneous versus homogeneous ability grouping, the ultimate prerogative is having valid data on the current level of students as the deciding placement factor.

In summary, grouping students in classes according to their current proficiency does not necessarily mean homogeneous placement. Since there is ample evidence that low-achieving students' academic performance benefits from interacting with high-achieving classmates (Schofield, 2010, p. 1505), testing for valid and current proficiency data also provides a chance for planned heterogeneity.

2 Literature Review

2.1 The Evolution of the C-Test

The C-Test is a special form of cloze test developed in the beginning of the 1980s by Christine Klein-Braley and Ulrich Raatz (Klein-Braley, 1983; Raatz & Klein-Braley, 1983). It measures GLP by reducing redundancy in texts with fixed-ratio deletion of the second half of every sec-

ond word. The various claims made by this short definition of the C-Test have been the subject of extensive academic discussion and need to be qualified. With the goal of ranking students according to their proficiency in a foreign language in mind, the following paragraphs will discuss the validity of the C-Test construct and match it to its practical application as an automated platform for gateway and placement testing.

Having language students[9] fill in blanks in a text is a traditional measure in language teaching and testing. By filling in the right word and adapting the grammatical form of that word to its surroundings, students can demonstrate their grasp of the subject matter, the extent of their vocabulary, and their grammatical prowess.

One of the features of fill-in-the-blanks tests is their adaptability for specific[10] purposes, but the resulting tests often lack the authenticity of natural language. Another major drawback, especially when grading tests with the help of templates, is the occurrence of unplanned ambiguity, i.e. multiple solutions applying to a blank without the designer realizing this during construction. In language testing, carefully designed and targeted fill-in-the-blanks tests can serve in assessing certain specific language phenomena, but fall short when evaluating general language proficiency.[11]

GLP testing is summative[12] in nature and often includes a whole battery of vocabulary, phonetic, and grammatical tests for the different skills, such as reading, listening, writing, and speaking. In order to render

9 In many subjects, teachers design fill-in-the-blanks tests in order to test specific factual knowledge, acquired skills, or proficiency. Where history students might need to provide the exact date of the battle of Hastings, students of German would have to fill in, for example, the gendered articles for nouns, subjunctives, or the declension of adjectives.

10 In theory, very long fill-in-the-blanks tests will eventually present a blank for most morphological and semantic phenomena, thus revealing the candidate's general language proficiency. The practicality of this approach is severely limited by the size of such a test, a fact that has contributed to developing the fixed-ratio approach used by the cloze procedure.

11 Spolsky (1985, p. 180) identifies three areas of language proficiency: Structural proficiency in the form of grammar or structural description of a language, functional proficiency in the various uses to which a language can be put, and general proficiency, which sees language as an indivisible body of knowledge that can be measured in individuals.

12 The purpose of formative testing is to assess a candidate's mastery of a given program's objectives, thereby simultaneously obtaining information on the efficacy of the program itself. In summative evaluation however, the question is whether the candidate can use language efficiently outside the classroom and the limitations of textbooks (cf. Bachman, 1990, p. 62).

summative language testing more economical, Taylor (1953) introduced the cloze test as a single test unit to replace the battery of tests involved before. Cloze is a variation of fill-in-the-blanks tests, where instead of purposefully deleting certain morphologically meaningful entities, every n^{th} word is automatically replaced by a blank. This kind of systematic deletion regardless of morphology or semantics is called fixed-ratio deletion as opposed to rational deletion (cf. Bachman, 1985, p. 536). Cloze tests may be an indicator of lexical and grammatical competence (cf. Jonz, 1990; Alderson, 1979a, 1979b) as well as of discourse competence (John W. Oller & Conrad, 1971; John William Oller, 1979). Although there is "no firm consensus as to what aspects of linguistic competence cloze tests measure, their scores correlate highly with standardized proficiency scores" (Tremblay, 2011, p. 344).

The cloze procedure makes use of one of the "vital truths about language, the fact that language is redundant" (Spolsky, 1968, p. 5). Redundancy in natural language is important in order to convey unequivocal meaning and to overcome disruptions, such as acoustic interferences during a conversation or bad print in written communication. These disruptions, summarily called noise, make comprehension difficult by overlaying meaningful parts of the message and thus causing a reduction in the original amount of redundancy.[13] Spolsky goes on to analyze that the ability of understanding a distorted message can be taken as a sign that the recipient has a thorough understanding of that language and that "someone who doesn't understand the language well […] just cannot function" with distorted or incomplete messages (Spolsky, 1968, p. 9). Thus, being able to understand a distorted message is a strong indicator of language proficiency in learners.

The principle of fixed-ratio deletion simulates naturally occurring communication noise for the purpose of GLP testing. Although cloze tests can produce reliable assessments, they have a considerable number of deficiencies in practical usage (cf. Khoshdel-Niyat, 2017, pp. 1–2). (1) In

13 This lack of redundancy has turned out to be a technical problem for the engineers of early telephone companies (cf. Shannon, 1948), who battled with severe acoustic interference threatening the efficacy of telephone conversations. Having the financial interest of industry backing may have helped motivate further research into the phenomenon of reduced redundancy.

order to have a sufficient number of items, cloze tests need to be very long. (2) Using one long text may turn the test overly specific and thus biased; it might occur that a participant gets a bad result because of her lack of understanding the contents of the text rather than due to her lack of language skills. (3) The blanks in cloze tests still are prone to unplanned ambiguity, which makes scoring time consuming and sometimes subjective. (4) Cloze tests are not automatically valid tests of language proficiency; their difficulty depends on the deleted words rather than on the deletion method (Alderson, 1983, p. 213).

The C-Test was designed to overcome the drawbacks of the cloze procedure (Raatz & Klein-Braley, 1983). Where the cloze eliminates every nth word (usually every 5th) from a given text, the C-Test erases the second half of every second word.

> "Indeed, the weakness that Klein-Braley (1981) spotted in the cloze test was its use of the word, a more or less linguistic unit, as the unit to be deleted, and as she showed, this very fact meant that a specific cloze test was biased towards measuring specific structural features. The new C-Test that Raatz and Klein-Braley (1982) have proposed overcomes this by deleting not words but parts of words; it is thus further from being a measure of structural ability, and so closer to a general measure."
>
> (Spolsky, 1985, p. 188)

A C-Test normally consists of five[14] increasingly difficult, content-neutral, target-group adequate, non-fictional, non-dialogical, authentic short texts of 80-100 words each, each containing approximately 20 blanks, resulting in 100 blanks per test (cf. Klein-Braley, 1997, p. 64).

The increased frequency of deletions allows the combination of five short texts with 20 blanks each, thus reducing problems (1) and (2), i.e. the economy/ bias/ validity complex.[15] The measure of replacing

14 Raatz and Klein-Braley (1985, p. 20) use "six texts with around 60-70 words" which are then turned into C-Tests for calibration with native speakers. After an extensive calibration process, only four texts remain, resulting in a C-Test with 80 blanks.

15 "A classical cloze test using a 5th word deletion rate would have to be at least 500 words long

the second half of words with blanks makes use of the redundancy in natural languages, tests grammatical knowledge simultaneously with vocabulary, and addresses the ambiguity problem (3) of cloze tests.[16]

C-Tests are a summative form of assessment and instructionally insensitive, i.e. they do not per se refer to any specific teaching material, nor do they reflect the teacher's educational skills (cf. Popham et al., 2014, p. 305). C-Tests have been developed for more than 20 languages (Eckes & Baghaei, 2015, p. 85). Among other advantages, the C-Test as a highly computer-adaptable test for GLP is ideal as a ranking tool for purposes of gateway testing and placement testing (Eckes & Baghaei, 2015, p. 85; Klein-Braley, 1997, pp. 65–66).

2.2 C-Test Validity Studies

The question of validity of the C-Test construct has been the topic of papers spanning four decades. Today, there is ample evidence that the C-Test is a valid measure of GLP (Drackert, 2016, p. 184; Sumbling, Viladrich, Doval, & Riera, 2014; Baghaei & Grotjahn, 2014; Tabatabaei & Mirzaei, 2014; Khodadady, 2014; Rouhani, 2008; Eckes & Grotjahn, 2006, pp. 294–300; 315; Chapelle, 1994, p. 175). Strong indicators for this claim are the high correlation between C-Tests and other language tests, factorial structure, and its fit to the Rasch model (Khoshdel-Niyat, 2017; Eckes & Baghaei, 2015; Baghaei, 2010; Eckes & Grotjahn, 2006; Sigott, 2004).

In particular, Eckes & Grotjahn (2006) have shown a significant correlation between C-Tests and other language tests in both receptive and productive skills. Pointing to the consistent correlation of the C-Test's results with other language tests, the vast majority of research confirms the constructs' validity as a test of GLP. Nevertheless, disputes about

to contain 100 items. A C-Test consisting of five texts with 20 half-deleted words would be only approximately half as long." (Klein-Braley, 1997, p. 65).

16 Grotjahn, who did not adhere to the strict rule of deleting exactly half of every other word, encountered several problems with ambiguity in French and Spanish C-Tests, which forced him to increase the number of texts used in the calibration phase. He recommends to "start developing a C-Test with at least twice as many texts as the test will eventually consist of" (Grotjahn, 1987, p. 223).

the validity of C-Test results regarding isolated skills exist (cf. Chapelle, 1994). Roos (1996a) tried with limited success to adapt the C-Test to the agglutinating language Japanese. Arras and Grotjahn (1994) found that Chinese C-Tests tend to test rather the ability of reading and writing Chinese characters than GLP, a result that Roos (1996b) reproduced for Japanese kanji characters. Jafarpur (1995) criticizes a lack of face validity, when his candidates compared the appearance of a C-Test to a puzzle rather than a language test. Therefore, the C-Test construct is valid for testing GLP in inflected languages, but if the goal is assessing isolated skills in listening, speaking, or grammar, one should resort to specialized tests (cf. Dresemann & Traxel, 2005, p. 278).

3 The Case for an Online C-Test Platform

The proposed automated platform can generate, administer, and grade unique C-Tests for any number of candidates. Generating a new C-Test only requires filling in five items of information, which should take less than a minute (see Figure 1: Gateway Testing – Creating a Unique C-Test in the Platform below). The preset testing time for a standard five-text, hundred-item test is 40 minutes, and grading is instantaneous. It is therefore an economical solution for ranking large or small groups of candidates according to their general language proficiency. As Dresemann and Traxl (2005, p. 277) pointed out, many teachers shy away from testing because of a lack of time or a perceived lack in competence. Once testing does not take much time and very little effort, the C-Test platform could also help individual teachers with in-class grouping,[17] assess the overall success of a course, serve individual students as an indicator of personal learning progress, and help students decide whether to commit to a fee-based official assessment test.

Making the platform web-based further helps with test economy. It satisfies the three key areas of availability, affordability, and convenience:

17 The composition of work groups can be heterogeneous or homogeneous, according to the pedagogical needs of the task (cf. Odendahl, 2016).

Anyone with an internet connection can access it at all times, access can be free of charge, and users can access it using any computer or mobile device with an internet connection.

3.1 Automating the Test: Web-Based General Language Proficiency Testing

Only an economical test that strikes a positive balance between cost and reward has the potential to sway foreign language departments and teachers in favor of testing over traditional gateway and placement procedures.

What is the potential reward from using a GLP test in admissions and placement? Being able to select the students with the best language skills and then grouping them according to the same principle is a very motivating outcome. The cost of a valid and reliable test comprises financial cost as well as time, personnel, and effort spent on design, administering, and grading. As demonstrated by TestDaF and onSET, the C-Test takes little over half an hour and can be administered and graded by computer. The following section will introduce the design principles of the C-Test and the history of its validity debate. It will then proceed to lay out an online C-Test platform which is able to produce, administer, and grade a unique C-Test at the press of a button.

3.2 Institutional Gateway Testing

Gateway testing for an advanced language program at university level, such as admission to an MA program, can be classified as high-stakes testing, where candidates need to be reasonably supervised to verify identity and prevent cheating (American Educational Research Association, 2014, p. 188).

In this setting, a staff member needs to spend a few minutes before the test to fill in (a) the test name and (b) the test date, (c) determine the way

in which candidates identify themselves, and (d) set an access password which allows candidates to take this test.

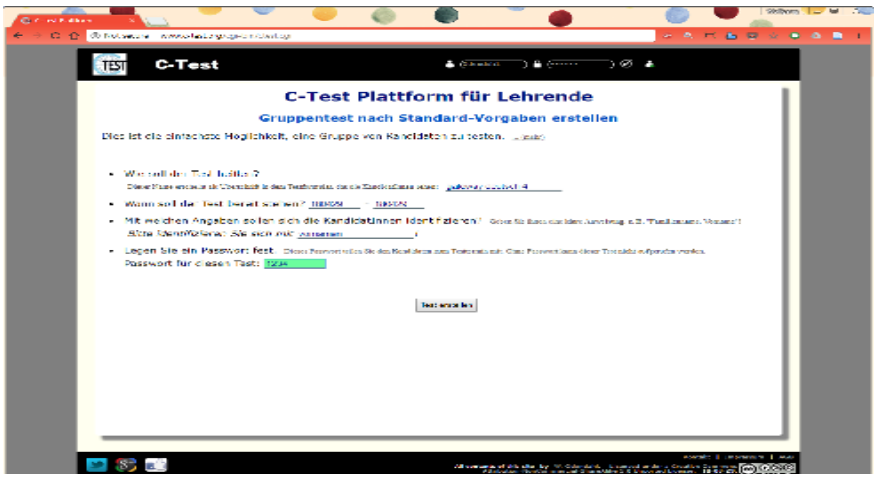
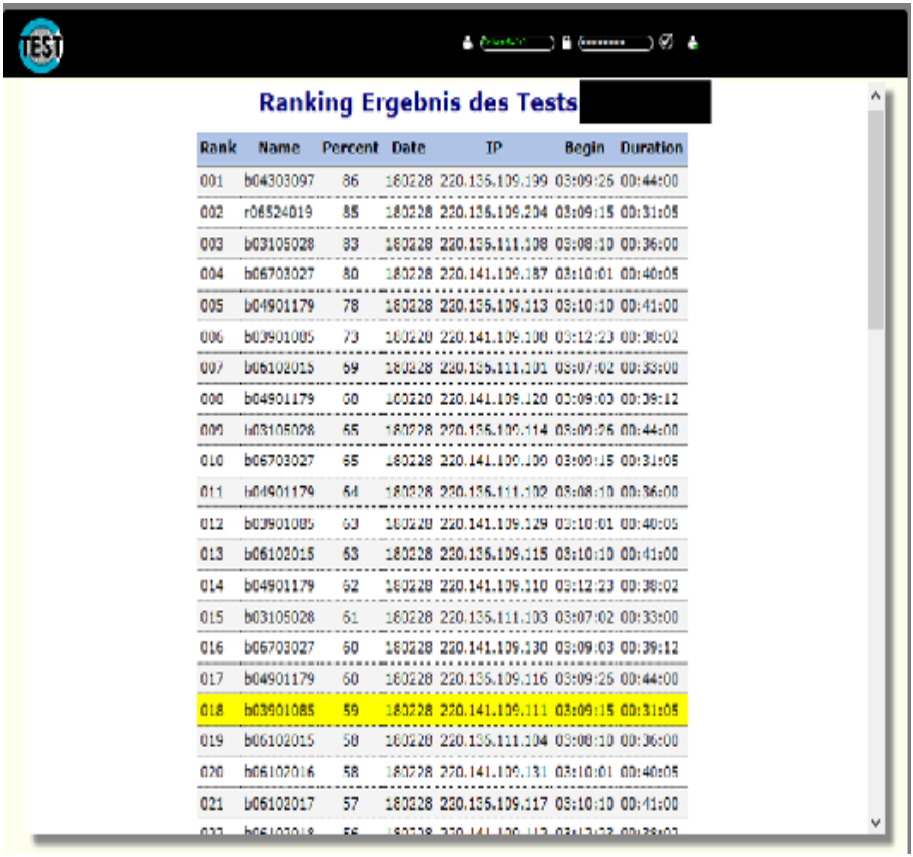


Figure 1: Gateway Testing – Creating a Unique C-Test in the Platform

Based on these variables, the system generates a unique C-Test, which will only be accessible during the predetermined dates and with the correct password. On the test date, the candidates will assemble in a computer-classroom, where they will be instructed about the test modalities. Afterwards, they open a web browser and log in to the test with the URL and the password provided on the blackboard. Each candidate's time will individually start after they successfully log in; in case of computer problems, the candidate may just switch to a different machine without suffering any disadvantages.

While taking the test, the candidate's name is displayed in the upper right corner of the screen. Generally, this feature signals the user that she is logged in; in a medium-stakes test setting such as described above, the teacher might use this information to verify the identity of the candidate taking the test. After the preset amount of time (cf. Figure 1: Gateway Testing – Creating a Unique C-Test in the Platform), a message tells the user to submit their results or suffer overtime deductions. Once every candidate in the room has submitted their test, the teacher may access the ranking list of results.



Ranking Ergebnis des Tests [redacted]

Rank	Name	Percent	Date	IP	Begin	Duration
001	b04303097	86	180228	220.135.109.199	03:09:25	00:44:00
002	r06524019	85	180228	220.135.109.204	03:09:15	00:31:05
003	b03105028	83	180228	220.135.111.108	03:08:10	00:36:00
004	b06703027	80	180228	220.141.109.187	03:10:01	00:40:05
005	b04901179	78	180228	220.135.109.113	03:10:10	00:41:00
006	b03901085	73	180220	220.141.109.100	03:12:23	00:38:02
007	b06102015	69	180228	220.135.111.101	03:07:02	00:33:00
008	b04901179	68	180220	220.141.109.120	03:09:03	00:39:12
009	b03105028	65	180228	220.135.109.114	03:09:25	00:44:00
010	b06703027	65	180228	220.141.109.100	03:09:15	00:31:05
011	b04901179	64	180228	220.135.111.102	03:08:10	00:36:00
012	b03901085	63	180220	220.141.109.129	03:10:01	00:40:05
013	b06102015	63	180228	220.135.109.115	03:10:10	00:41:00
014	b04901179	62	180220	220.141.109.110	03:12:23	00:38:02
015	b03105028	61	180228	220.135.111.103	03:07:02	00:33:00
016	b06703027	60	180228	220.141.109.130	03:09:03	00:39:12
017	b04901179	60	180228	220.135.109.116	03:09:25	00:44:00
018	b03901085	59	180228	220.141.109.111	03:09:15	00:31:05
019	b06102015	58	180220	220.135.111.104	03:08:10	00:36:00
020	b06102016	58	180228	220.141.109.131	03:10:01	00:40:05
021	b06102017	57	180228	220.135.109.117	03:10:10	00:41:00

Figure 2: Ranking List of Test Results

Figure 2: Ranking List of Test Results shows the ranked list of results which the teacher can pull up after the candidates completed a C-Test on the web-platform. Here, the teacher's mouse cursor rests at No. 018, at the boundary of 60 points, highlighting the first candidate who did not achieve this arbitrary limit – or maybe the program the candidates apply for has just 17 slots available which are awarded to the 17 best candidates.

When creating the test (cf. Figure 1: Gateway Testing – Creating a Unique C-Test in the Platform), we asked candidates to identify themselves with their matriculation number, which here is shown in the sec-

ond column, labeled "Name." The percentile refers to the candidate's correct answers. The following columns show the same date, similar IP addresses, and starting times for all candidates. This is owed to the test setting in a computer classroom. Candidates No. 006 and 014 seem to have started significantly later than the others, which might point to problems with their original computers. The time limit was set to 40 minutes, so candidates 001, 005, 009, 013, and 017 went overtime and had points deducted for each minute they delayed submitting their results. It is remarkable that the highest-scoring candidate still holds first place even after having been fined for overtime.

The teacher/staff member may project the results immediately after the test with the candidates still in the testing room, announcing something along the lines of "These are your results. We are now going to take a short break. Candidates 1 through 17, please return after the break for more information about our program. The others may leave at their leisure. Thank you for participating."

The only limit to the number of candidates in such a gateway setting is the number of available computers. If students are allowed to bring their own devices, there is virtually no limit to the number of testees.[18] The four-step effort for preparing and administering tests (cf. Figure 1: Gateway Testing – Creating a Unique C-Test in the Platform) remains the same regardless of the number of candidates.

3.3 Individual Self-Administered Assessment

The second usage example covers self-administered language testing by individual students. In this setting, a student is unsure whether she should invest time and money for an official language test and wants to know her chance for succeeding. Her teachers might encourage her, but she needs an independent and objective assessment of her overall language skills before committing herself.

18 Computing power will go down with increasing numbers of simultaneously submitted test results. For groups exceeding several hundred candidates, advance notice to the technical staff of the server administrators would be advisable.

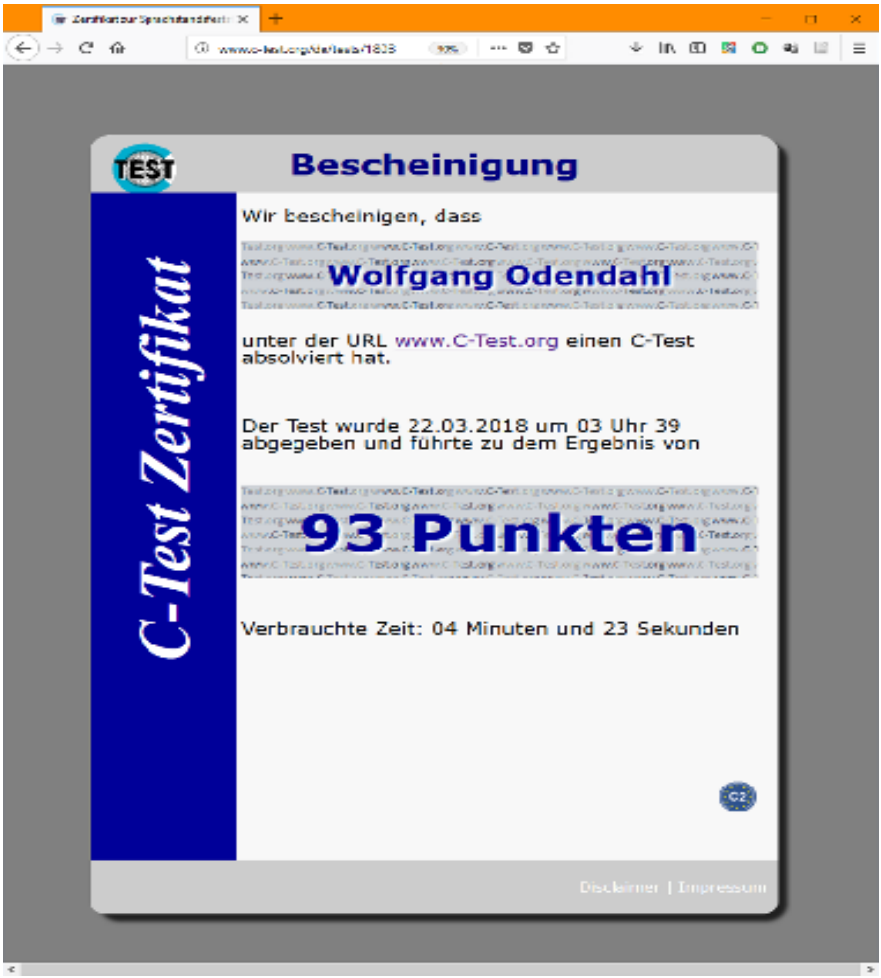


Figure 3: Results of a Self-administered C-Test

In order to get such an independent assessment, she pulls out her smartphone or sits down in front of her computer, accesses the C-Test web platform, skips registration, and directly accesses a test by pressing the "start" button. Immediately after submitting her test, the results display on her screen, giving the achieved percentage points and an estimate of the corresponding language level acording to the CEFR.[19]

19 Matching results from a C-Test to CEFR definitions is a tentative process. It makes use of the fact

In this example, the student just had to press one button in order to have the platform generate and administer a unique C-Test. The printable certificate issued by the server states the data submitted by the candidate and the test results. The mark in the lower right corner indicates an estimate of the equivalent proficiency level according to CEFR. Besides getting a second opinion on their language skills, students might want to independently and objectively track their progress by regularly taking tests in the privacy of their home and at their convenience. They will get a different test every time they choose to take it.

4 Technical Details and Inner Workings of the Platform

4.1 Automated Test Generation

The platform relies on a corpus of edited and calibrated texts indexed by a database. Whenever a user presses the start button on the test webpage, the system randomly picks five texts with three different difficulty levels and arranges them in ascending difficulty. It then iterates through each text, counts the number of words while omitting those marked as exempt from mutilation,[20] randomly determines a starting point between words 15 and 25, splits 20 words in half while replacing the second half with a blank and recording the eliminated part as the solution. The process of randomly selecting and matching five texts from the database in combination with a random starting point for mutilation assures the uniqueness of each different test.

Although there is currently only a German language version with an uncalibrated[21] corpus of 63 texts, the core system is language indepen-

that C-Test results have a high correlation with modular standard tests of general language proficiency (Baghaei, 2010, 2011; Eckes, 2007, 2011; Eckes & Grotjahn, 2006; Khoshdel-Niyat, 2017; Raatz, 1984; Tabatabaei & Mirzaei, 2014; Tremblay, 2011) and assumes that the reduced redundancy principle of C-Tests actually measures general language proficiency (Asano, 2014; Baghaei & Grotjahn, 2014).

20 See the following section for an extended discussion on how to determine which words should not be tested.

21 The calibration of texts for use in C-Tests has been the topic of several academic papers (cf.

dent and theoretically works with any alphabet-based language.[22] Two steps are involved in adding a language set, namely, adapting the user interface and adding a calibrated and edited text corpus.

4.2 Choosing Texts for Use in the C-Test Database

The details of finding texts for use in a C-Test can lead to very complicated problems. What appears to be a rather easy text when read as a whole can become very difficult once the second half of every second word is replaced by a blank. The following section will first discuss the implications of following the C-Test construction principles for choosing texts, and then explain the pragmatic approach in solving these problems.

The construction principles, as laid down by Raatz and Klein-Braley (1985, pp. 20–22), ask for five texts of increasing difficulty with 20 blanks each to constitute one set. Mutilation starts after the first sentence, which is left complete in order to provide some context. Once the predetermined number of blanks is reached, mutilation stops and the text comes to a natural end. The texts should be authentic, short, relevant to the intended user group, and arranged in order of ascending difficulty.

The problems in following these requirements are:

1. Where to find an authentic text of advanced difficulty with only 60-80 words?

Arras, Eckes, & Grotjahn, 2002; Dresemann & Traxel, 2005; Traxel & Dresemann, 2010). Calibration should involve anchor-items of known difficulty as points of reference and several stages of testing with native and non-native speakers. Since the main intended usage for this platform, ranking, can be reliably achieved with uncalibrated texts, the task of calibrating texts from the database will be postponed until its usage has produced sufficient data for analysis.

22 There have been experiments with non-alphabetic languages, such as Japanese (Roos, 1996a, 1996b) and Chinese (Arras & Grotjahn, 1994; Lin, Yuan, & Feng, 2008). However, the construction of C-Tests for these languages requires such a lot of adaptations and deviations from the C-Test principles as laid out by Klein-Braley and Raatz, that it may be argued to be a different testing system altogether. Furthermore, since the written and oral forms of these languages have only little (Japanese) or no (Chinese) relation to each other, the results of such tests cannot be accepted as an indication of general language proficiency.

2. How to ensure enough context if the first sentence is very short?
3. How to handle words that would pose unsolvable difficulties or ambiguity when cut in half?
4. How to determine the difficulty of texts where every second half of every second word is replaced with a blank?

The pragmatic answers to these problems are as follows: Problem (1) is based in Raatz' and Klein-Braley's (2002, p. 75) demand that texts should be non-dialogic and authentic. In terms of content they should pose no difficulties for the target group – here university students – so that the validity of the construct will not be affected by extra-linguistic factors such as personal experience, expertise, or qualifications (Messick, 1989, p. 14). However, finding authentic texts for the database can pose serious obstacles, because authentic texts with a very low readability index[23] are rarely found outside of textbooks. Similarly, at high levels of language competency, authentic texts with only 70-100 words are hard to come by.[24] Cronjaeger et al. (2010, p. 75) argue that authentic texts may altogether be too variable in terms of vocabulary and grammatical structures for use with beginning learners. Although textbook texts could offer the advantage of explicitly being written for learners with a specific level of language skills, copyright issues and the possibility of prior knowledge by some candidates effectively prevent us from using them. Therefore, all texts for consideration in the database originate from authentic sources, but are subject to radical revision, calibration, and partial re-writing before usage.

The pragmatic solution to the second (2) problem, how to ensure enough context if the first sentence is very short, is to not rely on punctuation

23 A common approach to determining the difficulty of texts is readability indices, which use elements like content, style, structure, and design to determine a text's reading ease (DuBay, 2004, p. 18 f.). For German, LIX is a reliable freeware readability index software (Lenhard & Lenhard, 2011). It has to be noted, however, that readability formulas are of limited use when determining the difficulty of texts after they have been mutilated according to the C-Test principles. Aside from language-specific difficulties, such as compound nouns in German, researchers found significant differences in the ability of candidates to solve blanks in content words as opposed to structure words (Chapelle, 1994, p. 176). These would not make a difference in LIX scaling.

24 The CEFR defines competent language use at level C1 in reading comprehension explicitly by stating that the learner "… can understand a wide range of demanding, longer texts […]" (Trim, North, & Coste, 2009, Chapter 3.3).

as end-of-sentence markers, but to randomly assign between 15 and 25 words as an introductory passage. As an added benefit, this solution helps randomize the C-Tests generated from the database, effectively increasing the number of possible C-Test passages created from each source text 11-fold.

The third (3) problem concerns unsolvable difficulties and ambiguity created by eliminating the second half of a word. Research indicates that in C-Tests, blanks resulting from certain word-groups are easier to solve than others. In German, unintended ambiguity may arise with words containing prefixes or suffixes and with combined nouns (cf. Arras, Eckes, & Grotjahn, 2002, p. 184). Furthermore, there is a difference in the difficulty of content words versus structure words. Correctly restoring content words requires knowledge of the formal features of the word as well as processes for composing the morphologically fitting form for a given context (Chapelle, 1994, p. 176). It seems that the ability to solve mutilated content words in C-Tests is a better measure for the general language proficiency of more advanced language learners, whereas weaker candidates tend to show differences in the ability of solving structure/function words (Eckes & Grotjahn, 2006, p. 294). Since in ranking, differences in test performance are the decisive factor, both content- and structure words can be part of C-Tests. In order to further quantify the question of how to adapt C-Tests or candidates of different levels of language proficiency, researchers could create a specialized databank with texts intentionally tweaking the amount of structure- versus content words and comparing the results of different learner groups.

Concerning the usage in C-Tests, the question of text difficulty (4) naturally follows the third (3) problem. The platform relies on a stock database of texts, the index of which includes topical keywords and the readability level of each text. These texts stem from internet blogs, novels, and newspaper articles and are edited for usage in C-Tests.

Since the main purpose of the C-Test platform is to produce rating scales for language proficiency as a gateway tool, the difficulty of the

undamaged texts is not the most important criterion; even with the occasional ambivalent blank, the ranking hierarchy of candidates still remains valid, because all take the same test (cf. Dresemann & Traxel, 2005, p. 277). Concerning individual assessment, however, having flawed texts in the database will lead to inaccurate individual evaluations, which poses a problem for individual students who use the platform as a screening-test before applying for language certification. For this user group, the texts need to be calibrated by means of monitoring test outcomes and running statistical analyses of problematic blanks to provide data for manual editing and modification. After accumulating a sufficient number of modifications, the validated text corpus will deliver more reliable test results. Another source of calibrated texts could be the project of Dresemann and Traxel (2005; 2010), who have assembled a reliability-calibrated database of German texts for the use in C-Tests.

The pragmatic solution for awarding difficulty ratings to the texts is modification rather than calibration. This means that competent native speakers re-write the texts in order to avoid ambiguities and other pitfalls, thereby sacrificing some of the texts' authenticity. During editing, special attention is given to compound nouns, names, and other words considered problematic when mutilated.[25] While rewriting problematic passages in the original text is the most efficient way to eliminate undesirable words, there are two other ways to mark these words for exemption from the automatic mutilation process. First, words with an asterisk at the end are exempt from mutilation, which will make the mutilation process shift one word to the right. The second option is to shift mutilation by one or two letters to the left or right by adding ±n to the end of a word in order to make it solvable: The German combined noun Schifffahrt, for example, contains eleven letters and would regularly be mutilated to Schiff_____, which can be solved in several semantically fitting ways. The editor would therefore change the original to Schifffahrt+1, which tells the system to leave one more letter and mutilate the word to the non-ambiguous Schifff____.[26]

25 In C-Test research, the process of deleting the second half of words is commonly referred to as mutilation (Klein-Braley & Raatz, 1984; Raatz & Klein-Braley, 1985; Grotjahn, 1987; Klein-Braley, 1997; Babaii & Ansary, 2001; Baghaei & Tabatabaee, 2015; Khoshdel-Niyat, 2017)

26 One of the most basic rules in creating C-Test blanks calls for deleting half of the word. In the

5 Conclusion

This article shows that a web-based C-Test platform offers an economical alternative to accepting candidates' previous grades as the basis for gateway testing. It further argues that ranking students by general language proficiency also allows for meaningful grouping in other settings, such as class placement and in-class grouping. The article rebukes the allegation of tracking by the argument that knowing the skill level of students allows for homogeneous grouping as well as according to patterns of planned heterogeneity. Increasing the frequency of placement and regrouping helps to avoid restricting students to a fixed group and promotes mobility according to their current language skills.

In gateway testing, using the same test for all students will set objective standards for admission. Frequent placement tests and regrouping increases the efficiency of instruction by assembling learner groups according to their actual and current language skills. The C-Test construct is an adequate, valid, and reliable means of testing general language proficiency. Conforming to the definition in section 1.4 of this article, the platform proposed here is an economical testing tool. It presents the results of individual tests as a printable diploma, and groups tests in list form, ranking candidates according to their test results. An automated C-Test generating internet platform makes testing universally available with very little preparation, minimum time loss, and considerable benefits.

The data derived from an automated C-Test platform can support research in numerous fields, including the C-Test construct, general language proficiency testing, autonomous language acquisition monitoring, and others. Metadata, like geographical user distribution, frequency of deployment in different circumstances, and the perception of C-Tests by users and administrators, provides answers to a wide array of questions concerning foreign language acquisition.

An interesting area of research will be TestDaF, the admission test for

---

case of Schifffahrt, the word has 11 letters. The division in halves would therefore result in 5 letters and 6 blanks or 6 letters and 5 blanks, i.e. Schif/ffahrt

German universities, which includes a C-Test. Will students who are acquainted with C-Tests from the platform do significantly better with TestDaF? Also based on statistical analyses, another valid question concerns the significance of the mistakes candidates make in filling in the blanks. In concurrence with Klein-Braley, who states "wrong answers provide us with more insights into text processing strategies than right answers do" (1996, p. 39), an analysis of a large number of wrong answers from language students might reveal new insights for test validity, reduced redundancy, and – more generally – basic processes involved in language testing and learning.

References

Alderson, J. C. (1979a). "The Cloze Procedure and Proficiency in English as a Foreign Language". *TESOL Quarterly*, *13*(2), 219–227.

———. (1979b). "The Effect on the Cloze Test of Changes in Deletion Frequency". *Journal of Research in Reading*, *2*(2).

———. (1983). "The Cloze Procedure and Proficiency in English as a Foreign Language", in Oller, J. W. (ed.), *Issues in Language Testing Research* (pp. 205–217). Rowley, Mass: Newbury House.

———. (2017). "Foreword to the Special Issue "The Common European Framework of Reference for Languages (CEFR) for English Language Assessment in China" of Language Testing in Asia". *Language Testing in Asia*, *7*(1), 20.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). "Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice across Diverse Fields". *Applied Linguistics*, *36*(2), 236–260.

American Educational Research Association. (2014). *Standards for Educational and Psychological Testing.* (American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.), Eds.). Washington, DC: American Educational Research Association.

Arras, U., Eckes, T., & Grotjahn, R. (2002). "C-Tests im Rahmen des "Test Deutsch als Fremdsprache" (TestDaF): Erste Forschungsergebnisse" in Grotjahn, R. (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 175–209). Bochum: AKS.

Arras, U., & Grotjahn, R. (1994). "Der C-Test im Chinesischen", in Grotjahn, R. (ed.), Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Vol. 2, pp. 1–60). Bochum: Brockmeyer.

Asano, Y. (2014). "Nähere Betrachtung des Konstrukts: Allgemeine Sprachkompetenz", in Grotjahn, R. (ed.), *Der C-Test: Aktuelle Tendenzen* (pp. 41–54). Frankfurt / M.: Lang.

Babaii, E., & Ansary, H. (2001). "The C-Test: A Valid Operationalization of Reduced Redundancy Principle?" *System*, *29*(2), 209–219.

Bachman, L. F. (1985). "Performance on Cloze Tests with Fixed-Ratio and Rational Deletions". *Tesol Quarterly*, *19*(3), 535–556.

———. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

———. (2005). "Building and Supporting a Case for Test Use". *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34.

Baghaei, P. (2010). "An Investigation of the Invariance of Rasch Item and Person Measures in a C-Test", in Grotjahn, R. (ed.), *Der C-Test: Beiträge aus der aktuellen Forschung* (pp. 71–100). Frankfurt / M.: Lang.

———. (2011). "Optimal Number of Gaps in C-Test Passages". *International Education Studies*, *4*(1), 166–171.

Baghaei, P., & Grotjahn, R. (2014). "Establishing the Construct Validity of Conversational C-Tests Using a Multidimensional Rasch Model". *Psychological Test and Assessment Modeling*, *56*(1), 60–82.

Baghaei, P., & Tabatabaee, M. (2015). "The C-Test: An Integrative Measure of Crystallized Intelligence". *Journal of Intelligence*, *3*(2), 46–58.

Bernhardt, P. E. (2014). "Making Decisions about Academic Trajectories: A Qualitative Study of Teachers' Course Recommendation Practices". *American Secondary Education*, *42*(2), 33–50.

Braddock, J. H., & Slavin, R. E. (1992). "Why Ability Grouping Must End: Achieving Excellence and Equity in American Education". Presented at the Common Destiny Conference, EDRS.

Brulles, D., Saunders, R., & Cohn, S. J. (2010). "Improving Performance for Gifted Students in a Cluster Grouping Model". *Journal for the Education of the Gifted*, *34*(2), 327–350.

Chapelle, C. A. (1994). "Are C-tests Valid Measures for L2 Vocabulary Research?" *Second Language Research*, *10*(2), 157–187.

Chapelle, C. A., & Voss, E. (2017). "Utilizing Technology in Language Assessment", in Shohamy, E. (ed.), *Language Testing and Assessment* (3rd ed., pp. 149–162). New York: Springer Sci-

ence+Business Media.
Cohen, E. G., & Lotan, R. A. (2014). *Designing Groupwork: Strategies for the Heterogeneous Classroom* (Kindle Edition). New York: Teachers College Press.
Cronjäger, H., Klapheck, K., Krätzschmar, M., & Walter, O. (2010). "Entwicklung eines C-Tests für Lernanfänger der Sek. I mit Methoden der klassischen und probabilistischen Testtheorie", in Grotjahn, R. (ed.), *Der C-Test: Beiträge aus der aktuellen Forschung* (pp. 71–100). Frankfurt / M.: Lang.
Daud, N. S. M., Daud, N. M., & Kassim, N. L. A. (2005). "Second Language Writing Anxiety: Cause or Effect?" *Malaysian Journal of ELT Research*, 1(1), 19.
Deutsch, M. (1975). "Equity, Equality, and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice?" *Journal of Social Issues*, *31*(3), 137–149. https://doi.org/10.1111/j.1540-4560.1975.tb01000.x
Deygers, B., Van Gorp, K., & Demeester, T. (2018). "The B2 Level and the Dream of a Common Standard". *Language Assessment Quarterly*, *15*(1), 44–58.
Díez-Bedmar, M. B. (2012). "The Use of the Common European Framework of Reference for Languages to Evaluate Compositions in the English Exam Section of the University Admission Examination". *Revista de Educación*, *357*, 55–79.
Drackert, A. (2016). *Validating Language Proficiency Assessments in Second Language Acquisition Research*. Frankfurt: Lang. https://doi.org/10.3726/978-3-653-06280-9
Dresemann, B., & Traxel, O. (2005). "Ermittlung von Sprachniveaus mittels kalibrierter C-Tests. Ein Projekt zur Entwicklung einer C-Test Datenbank", in Gebert, D. (ed.), *Innovation aus Tradition: Dokumentation der 23. Arbeitstagung 2004* (pp. 277–283). Bochum: AKS-Verl.
DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa: Impact Information.
Dunlea, J., & Figueras, N. (2012). "Replicating Results from a CEFR Test Comparison Project across Continents", in Tsagari, D. & Csépes, I. (eds.), *Collaboration in Language Testing and As-*

sessment (pp. 31–47). Frankfurt: Lang.
Dutcher, L. R. (2018). *Interaction and Collaboration across Proficiency Levels in the English Language Classroom* (Ph. D. Dissertation). University of Sydney, Sydney.
Eckes, T. (2007). "Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen". Diagnostica, 53(2), 68–82.
———. (2011). "Item banking for C-tests: A polytomous Rasch modeling approach". *Psychological Test and Assessment Modeling*, *53*(4), 414–439.
Eckes, T., & Baghaei, P. (2015). "Using Testlet Response Theory to Examine Local Dependence in C-Tests". *Applied Measurement in Education*, *28*(2), 85–98.
Eckes, T., & Grotjahn, R. (2006). "A Closer Look at the Construct Validity of C-Tests". *Language Testing*, *23*(3), 290–325.
Feldt, L. S., & Brennan, R. L. (1989). "Reliability", in Linn, R. L. (ed.), *Educational Measurement* (pp. 105–146). New York; London: American Council on Education and Collier Macmillan.
Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B., & Travers, M.-C. (2017). "Exploring the Relative Lack of Impact of Research on 'Ability Grouping' in England: A Discourse Analytic Account". *Cambridge Journal of Education*, *47*(1), 1–17.
Gamaroff, R. (2000). "Rater Reliability in Language Assessment: The Bug of All Bears". *System*, *28*(1), 31–53.
Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V., & TestDaF-Institut. (2017a). "About onSET [Corporate]". Retrieved December 31, 2018, from https://www.onset.de/en/language-placement-test-english-onset/about-onset/
———. (2017b). "TestDaF [Corporate]". Retrieved December 31, 2018, from http://www.testdaf.de/fuer-teilnehmende/informationen-zum-testdaf
Glock, S., & Böhmer, I. (2018). "Teachers' and preservice teachers' stereotypes, attitudes, and spontaneous judgments of male ethnic minority students". *Studies in Educational Evaluation*, *59*, 244–255.
Gnambs, T., Batinic, B., & Hertel, G. (2011). "Internetbasierte psychologische Diagnostik [Autorenmanuskript]", in Hornke, L.

F., Amelang, M., Kersting, M. (eds.), *Verfahren zur Leistungs-, Intelligenz- und Verhaltensdiagnostik* (Vol. II/3, pp. 448-498 / 1-62). Göttingen: Hogrefe. Retrieved December 31, from https://timo.gnambs.at/publications

Grotjahn, R. (1987). "How to Construct and Evaluate a C-Test: A Discussion of Some Problems and Some Statistical Analyses", ın Klein-Braley, C., Stevenson, D. K., Grotjahn, R. (eds.), *Taking Their Measure: The Validity and Validation of Language Tests* (pp. 219–253). Bochum: Brockmeyer.

Haertel, E. H. (2006). "Reliability", ın Brennan, R. L. (ed.), *Educational Measurement. Sponsored Jointly by National Council on Measurement in Education and American Council on Education* (4th ed., pp. 65–110). Michigan: Praeger.

Harmer, J. (2010). *How to Teach English* (New ed., 6. impr). Harlow: Pearson Longman.

Harris, D. M. (2012). "Varying Teacher Expectations and Standards Curriculum Differentiation in the Age of Standards-Based Reform". *Education and Urban Society*, *44*(2), 128–150.

Henry, L. (2015). "The Effects of Ability Grouping on the Learning of Children from Low Income Homes: A Systematic Review". *The STeP Journal*, *2*(3), 70–87.

Hornke, L. F. (2006). "Testökonomie: Test Economy", in Petermann F. , Eid, M. (eds.), *Handbuch der Psychologischen Diagnostik* (pp. 434–448). Hogrefe Verlag.

Hornstra, L., van der Veen, I., Peetsma, T., & Volman, M. (2014). "Does Classroom Composition Make a Difference: Effects on Developments in Motivation, Sense of Classroom Belonging, and Achievement in Upper Primary School". *School Effectiveness and School Improvement*, 1–28.

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). "Evaluating CEFR rater performance through the analysis of spoken learner corpora". *Language Testing in Asia*, *8*(1), 1–17.

Huhta, A. (2008). "Diagnostic and Formative Assessment", in Spolsky, B. & Hult, F. M. (eds.), *The Handbook of Educational Linguistics* (pp. 469–482).

Jafarpur, A. (1995). "Is C-testing Superior to Cloze?" *Language Test-*

ing, *12*(2), 194–216.
Jones, N., & Saville, N. (2008). "Scales and Frameworks", in Spolsky, B. & Hult, F. M., *The Handbook of Educational Linguistics* (pp. 496–510).
Jonz, J. (1990). "Another Turn in the Conversation: What Does Cloze Measure?" *TESOL Quarterly*, *24*(1), 61–83.
Khodadady, E. (2014). "Construct Validity of C-tests: A Factorial Approach". *Journal of Language Teaching and Research*, *5*(6), 1353–1362.
Khoshdel-Niyat, F. (2017). "The C-Test: A Valid Measure to Test Second Language Proficiency?" [Preprint]. *HAL Hprints*, *01491274*, 1–30. https://doi.org/10.31219/osf.io/c7sy5
Kim, Y. (2012). "Implementing Ability Grouping in EFL Contexts: Perceptions of Teachers and Students". *Language Teaching Research*, *16*(3), 289–315.
Klein-Braley, C. (1983). "A Cloze is a Cloze is a Question", in Oller, J. W. (ed.), *Issues in Language Testing Research* (pp. 218–230). Rowley, Mass: Newbury House.
———. (1996). "Towards a Theory of C-Test Processing", in Grotjahn, R. (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 23–94). Bochum: Brockmeyer.
———. (1997). "C-Tests in the Context of Reduced Redundancy Testing: An Appraisal". *Language Testing*, *14*(1), 47–84.
Klein-Braley, C., & Raatz, U. (1984). "A Survey of Research on the C-Test". *Language Testing*, *1*(2), 134–146.
Knapp, A. (2011). "Issues in Certification", in Knapp, K., Seidlhofer, B., Widdowson, H. (eds.), *Handbook of Foreign Language Communication and Learning* (pp. 629–662). New York: Mouton de Gruyter.
Lenhard, W., & Lenhard, A. (2011). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. de, Bibergau: Psychometrica. https://doi.org/10.13140/RG.2.1.1512.3447
Lienert, G. A., & Raatz, U. (1994). *Testaufbau und Testanalyse* (5th ed.). Weinheim: Beltz PVU.
Lin, W., Yuan, H., & Feng, H. (2008). "Language Reduced Redundancy Tests: A Reexamination of Cloze Test and C-Test". *Jour-*

nal of Pan-Pacific Association of Applied Linguistics, *12*(1), 61–79.
Loveless, T. (2013). *How Well Are American Students Learning? With Sections on the Latest International Tests, Tracking and Ability Grouping, and Advanced Math in 8th Grade* (Brown Center Report on American Education No. Vol. 3, No. 2) (p. 36). Washington, DC: Brookings Institution.
McNamara, T. (2011). "Principles of Testing and Assessment", in Knapp, K., Seidlhofer, B., Widdowson, H. (eds.), *Handbook of Foreign Language Communication and Learning* (pp. 607–627). New York: Mouton de Gruyter.
Messick, S. (1989). "Validity", in Linn, R. L. (ed.), *Educational Measurement* (pp. 13–103). New York; London: American Council on Education and Collier Macmillan.
Missett, T. C., Brunner, M. M., Callahan, C. M., Moon, T. R., & Azano, A. P. (2014). "Exploring Teacher Beliefs and Use of Acceleration, Ability Grouping, and Formative Assessment". *Journal for the Education of the Gifted*, *37*(3), 245–268.
Moosbrugger, H., & Kelava, A. (2011). *Testtheorie und Fragebogenkonstruktion* (2nd ed.). Springer-Verlag.
Mozgalina, A., & Ryshina-Pankova, M. (2015). "Meeting the Challenges of Curriculum Construction and Change: Revision and Validity Evaluation of a Placement Test". *The Modern Language Journal*, *99*(2), 346–370.
Nesmith, B. M. (2018). *Deciding on Classroom Composition: Factors Related to Principals' Grouping Practices* (Doctor of Education). Georgia Southern University, Statesboro.
Newbold, D. (2012). "Local Institution, Global Examination: Working Together for a 'Co-certification'", in Tsagari, D., Csépes, I. (eds.), *Collaboration in Language Testing and Assessment* (pp. 127–142). Frankfurt: Lang.
Norouzian, R., & Plonsky, L. (2018). "Correlation and Simple Linear Regression in Applied Linguistics", in Phakiti, A., De Costa, P., Plonsky, L., Starfield, S. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 395–421). London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-59900-1_19

Norris, J., & Drackert, A. (2018). "Test Review: TestDaF". *Language Testing*, *35*(1), 149–157.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. New York: Lang.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*. New Haven: Yale University Press.

———. (1986a). "Keeping Track, Part 1: The Policy and Practice of Curriculum Inequality". *The Phi Delta Kappan*, *68*(1), 12–17.

———. (1986b). "Keeping Track, Part 2: Curriculum Inequality and School Reform". *The Phi Delta Kappan*, *68*(2), 148–154.

Odendahl, W. (2016). "Promoting Student Engagement through Skill-Heterogeneous Peer Tutoring". *Interface - Journal of European Languages and Literatures*, 1, 119–153. https://doi.org/10.6667/interface.1.2016.26

———. (2017). „Individuelle Noten aus kollaborativer Arbeit". *Deutsch-Taiwanische Hefte*, *16*(25), 27–57.

Oller, John William. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.

Oller, John W., & Conrad, C. A. (1971). "The Cloze Technique and ESL Proficiency". *Language Learning*, *21*(2), 183–194.

Popham, W. J., Berliner, D. C., Kingston, N. M., Fuhrman, S. H., Ladd, S. M., Charbonneau, J., & Chatterji, M. (2014). "Can Today's Standardized Achievement Tests Yield Instructionally Useful Data? Challenges, Promises and the State of the Art". *Quality Assurance in Education*, *22*(4), 2–2.

Raatz, U. (1984). *The Factorial Validity of C-Tests*.

Raatz, U., & Klein-Braley, C. (1983). "The C-Test - A Modification of the Cloze Procedure", in Stevenson, D. K., Klein-Braley, C. (eds.), *Practice and Problems in Language Testing. Proceedings of the Fourth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe, held at the University of Essex*, 14-17th September, 1981 (Vol. 4, pp. 113–148). Colchester: Univ. of Essex.

———. (1985). "How to Develop a C-Test". *Fremdsprachen Und Hochschule*, *13*(14), 20–22.

———. (2002). "Introduction to Language Testing and to C-Tests",

in Coleman, J. A., Grotjahn, R., Raatz, U. (eds.), *University Language Testing and the C-test* (pp. 75–91). Bochum: AKS.

Reese, S. (2011). "Differentiation in the Language Classroom". *The Language Educator, 6*(4), 40–46.

Robinson, J. P. (2008). "Evidence of a Differential Effect of Ability Grouping on the Reading Achievement Growth of Language-Minority Hispanics". *Educational Evaluation and Policy Analysis*, *30*(2), 141–180.

Roos, U. (1996a). "The C-Test in Japanese", in Grotjahn, R. (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 2, pp. 61–118). Bochum: Brockmeyer.

———. (1996b). "The Reconstructability of Japanese Characters: Some New Evidence", in Grotjahn, R. (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 139–157). Bochum: Brockmeyer.

Rouhani, M. (2008). "Another Look at the C-Test: A Validation Study With Iranian EFL Learners". *The Asian EFL Journal Quarterly March 2008 Volume 10, Issue*, *10*(1), 154.

Schofield, J. (2010). "International Evidence on Ability Grouping with Curriculum Differentiation and the Achievement Gap in Secondary Schools". *The Teachers College Record*, *112*(5), 8–9.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Shohamy, E. (2017). "Critical Language Testing", in Shohamy, E. (ed.), *Language Testing and Assessment* (3rd ed., pp. 441–454). New York: Springer Science+Business Media.

Sigott, G. (2004). *Towards Identifying the C-Test Construct*. Frankfurt: Lang.

Slavin, R. E. (1987). „Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis". *Review of Educational Research*, *57*(3), 293–336.

———. (1990). "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis". *Review of Educational Research*, *60*(3), 471–499.

———. (1993). "Ability Grouping in the Middle Grades: Achievement Effects and Alternatives." *The Elementary School Journal*,

93(5), 535–552.
Smith, A. L. (2017). *Grouping Structures of Gifted and High Achieving Middle School Students: Teacher Perceptions and Data Analysis of the Impact of Grouping* (Ph. D. Dissertation). Columbus State University. Retrieved December 31, 2018, from https://csuepress.columbusstate.edu/theses_dissertations/224
Spolsky, B. (1968). "What Does It Mean to Know a Language, Or How Do You Get Someone to Perform His Competence?" Presented at the Second Conference on Problems in Foreign Language Testing, University of Southern California: ERIC Clearinghouse.
———. (1985). "What Does It Mean to Know How to Use a Language? An Essay on the Theoretical Basis of Language Testing". *Language Testing*, *2*(2), 180–191.
Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). „What One Hundred Years of Research Says about the Effects of Ability Grouping and Acceleration on K–12 Students' Academic Achievement: Findings of Two Second-Order Meta-Analyses". *Review of Educational Research*, *86*(4), 849–899. https://doi.org/10.3102/0034654316675417
Stöger, H., & Ziegler, A. (2013). "Heterogenität und Inklusion im Unterricht". *Schulpädagogik Heute*, *7*(4), 1–30.
Sumbling, M., Viladrich, C., Doval, E., & Riera, L. (2014). „C-Test as an Indicator of General Language Proficiency in the Context of a CBT (SIMTEST)", in Grotjahn, R. (ed.), *Der C-Test: Aktuelle Tendenzen- The C-Test: Current Trends* (pp. 55–110). Frankfurt: Lang. https://doi.org/10.3726/978-3-653-04578-9
Sun, X., Fan, J., & Chin, C.-K. (2017). "Developing a Speaking Diagnostic Tool for Teachers to Differentiate Instruction for Young Learners of Chinese", in Zhang, D., Lin, C.-H. (eds.), *Chinese as a Second Language Assessment* (pp. 249–270). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-4089-4_12
Tabatabaei, O., & Mirzaei, E. (2014). "Correlational Validation of Cloze Test and C-Test against IELT". *Journal of Educational and Social Research*, 4(1), 345.
Taylor, W. L. (1953). "Cloze Procedure: A New Tool for Measuring

Readability". *Journalism Quarterly, 30*, 415–433.
Tempel-Milner, M. E. (2018). *Implementing Full-Time Gifted and Talented Programs in Title 1 Schools: Reasons, Benefits, Challenges and Opportunity Costs* (Ph. D. Dissertation). Univeristy of Maryland, College Park.
Tieso, C. L. (2003). "Ability Grouping Is Not Just Tracking Anymore". *Roeper Review, 26*(1), 29–36.
Tomlinson, C. A. (2014). *Differentiated Classroom: Responding to the Needs of All Learners*. Alexandria, Va.: Ascd.
Tomlinson, C. A., & Imbeau, M. B. (2014). *Leading and Managing a Differentiated Classroom*. Alexandria, Va.: Ascd.
Traxel, O., & Dresemann, B. (2010). "Collect, Calibrate, Compare: A Practical Approach to Estimating the Difficulty of C-Test Items", in Grotjahn, R. (ed.), *Der C-Test: Beiträge aus der aktuellen Forschung* (pp. 57–69). Frankfurt / M.: Lang.
Tremblay, A. (2011). "Proficiency Assessment Standards in Second Language Acquisition Research: "Clozing" the Gap". *Studies in Second Language Acquisition, 33*(03), 339–372.
Trim, J., North, B., & Coste, D. (2009). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen [Niveau A1, A2, B1, B2, C1, C2].* (Council for Cultural Co-operation, Ed.). Berlin; München; Wien; Zürich; New York NY: Langenscheidt.
Vogl, K., & Preckel, F. (2014). "Full-Time Ability Grouping of Gifted Students: Impacts on Social Self-Concept and School-Related Attitudes". *Gifted Child Quarterly, 58*(1), 51–68.
Wunsch, C. (2009). „Binnendifferenzierung", in Jung, U. O. H. (ed.), *Praktische Handreichung für Fremdsprachenlehrer* (5th ed., pp. 41–47). Frankfurt: Lang.
Xie, Q. (2015). ""I must impress the raters!" An investigation of Chinese test-takers' strategies to manage rater impressions". *Assessing Writing, 25*, 22–37. https://doi.org/10.1016/j.asw.2015.05.001

[received November 23, 2018
accepted January 22, 2019]