relations with the dominant society/language/culture: some degree of direct rejection and/or a kind of approach-avoidance revulsion, both hidden beneath a veneer of collaboration. Tests (as currently conceptualized in IL-speaking communities) are clearly artifacts of the dominant society; therefore, they are either to be rejected or they are to be feared. How communities can be convinced to accept testing in their IL will have to be decided on a case-by-case, perhaps moment-by-moment basis.

While it will always take a lot of chutzpah to make SKs in ILs, it seems that the job can be made less daunting if IL test developers can understand that each individual's language weaknesses or lack of proficiency is not a significant stumbling block. What matters is the aggregate proficiency of the entire team, plus whatever any or all of them learn or learn to do during the test-making process. Taken together, their knowledge and skills are "native." Furthermore, by going (back) into IL-using communities, test developers can make reasonably sure that the criteria represented in their tests are, in essence, "real." And, I think, more cannot be asked of them.

References


Rüdiger Großjahn

Determinants of the difficulty of foreign language reading and listening comprehension tasks: Predicting task difficulty in language tests*

This article begins with a brief discussion of the modelling of foreign language reading and listening comprehension. Thereafter, the problem of measuring and predicting the difficulty of comprehension tasks is dealt with. Two main approaches are distinguished: text-centred and addressee-centred conceptions. The discussion of text-centred approaches treats readability research, the measurement and prediction of task difficulty on the basis of expert ratings, and the a priori determination of task difficulty within the framework of task-based language assessment. In the discussion of addressee-centred conceptions, the focus is on Buck, Tatsuoka & Kostin's (1997) methodologically innovative rule-space analyses. Finally, a tentative list of central determinants of text difficulty based on relevant empirical research is presented.

1. Introduction

This article pursues a topic which Kris Klein-Braley had been dealing with since the early 1980s. In 1981, for example, she held a lecture at the 12th annual conference of the Gesellschaft für Angewandte Linguistik (German association for applied linguistics) entitled: "The determination of text difficulty". Thereafter, in 1983, Kris Klein-Braley dealt with issues in text comprehensibility research in a lecture at the first Colloquium on Quantitative Linguistics at the University of Essen.

* Substantial parts of this article are based on Großjahn (2000b). A much longer version has appeared as Großjahn (2000a). The observations presented here were developed for the most part in connection with my work on TestDaF (Test Deutsch als Fremdsprache – Test of German as a Foreign Language) – the recently introduced German equivalent to IELTS and TOEFL (cf. Projekt-Gruppe TestDaF 2000).
The topic of text difficulty and text comprehensibility with particular reference to the C-Test was also a central issue in her post-doctoral thesis (Klein-Braley 1994).

The question as to which factors determine the difficulty of reading and listening comprehension tasks in foreign language tests has great practical and theoretical relevance. In the phase of theory-guided test construction, item writing would be much easier if on the basis of relatively simple task characteristics it were possible to make at least an approximate prediction of text and item difficulty without an empirical analysis of examinees' responses to the items. In the phase of empirical examination of test takers' actual responses, information on determinants of task difficulty would help to explain differences in item difficulty and discrimination, and thus facilitate the process of item revision; moreover, such information would help in the process of construct validation. Finally, information on specific determinants of task difficulty could also help produce an individual, detailed ability profile for each examinee. This diagnostic information could, for example, be used for learner counselling.

The main focus of the present article is reading comprehension. However, since reading and listening comprehension are based to a considerable extent on similar cognitive processes, a substantial part of what follows is also relevant to listening comprehension.¹

¹ Among other things, the substantial correlations which are usually found between reading and listening comprehension tests lend support to this view. There are also, however, a number of important differences between the two skills: listening comprehension involves much more "real-time processing" than does reading comprehension. Consequently, the listener has much less control of the text than does a reader, who, for example, can read certain portions of the text repeatedly. Furthermore, particular phonetic qualities such as an unfamiliar accent can make it difficult to understand what is being said. Particular lexical and grammatical qualities can also reduce comprehensibility (cf. e.g. Rost 1990, Lund 1991, Rubin 1994, Brindley 1998, Lynch 1998, Salmeecke 1993, 2000). Such differences must be taken into account when findings from reading comprehension research are applied to listening comprehension.

2. Modelling foreign-language reading and comprehension processes

Statements on reading and listening comprehension and on determinants of the difficulty of comprehension tasks only make sense if they can be related to an (explicit) model of the reading and listening process.

A number of levels must be distinguished when constructing a model. As far as reading comprehension is concerned, at least five hierarchical levels are generally recognised: the (a) graphophonetic, (b) lexical, (c) syntactic, (d) semantic, and (e) pragmatic levels.

Furthermore, two basic processing types are differentiated with regard to both reading and listening: bottom-up vs. top-down processing. Top-down or concept-driven processes at the semantic and pragmatic levels are the focus of Goodman's influential model of reading as a "psycholinguistic guessing game" (cf. e.g. Goodman 1967). In accordance with this view, reading consists primarily of constructing and testing conceptual hypotheses. However, as Lutjeharms (1994:37) correctly observes, reading based on top-down hypothesis testing would be "a very involved way of acquiring information compared to skilled, largely automatic, parallel processing on the lower levels". These latter processes are also referred to as "data-driven" processes and are the focus of bottom-up models of reading.

Nowadays, the view that reading and listening comprehension involves the interaction of both top-down and bottom-up processing has gained general acceptance. Thus, an individual's comprehension ultimately results from a complex interaction of processing factors on all levels.

However, both in first- and foreign-language reading pedagogy, most researchers and practitioners still work under the assumption that reading is primarily a top-down process. Recent cognitive reading comprehension research, by contrast, tends to emphasise the
significance of data-driven, automated processing on the lower levels (see Ehlers 1998, 1999 and Grotjahn 1995, 2000a for references).

Furthermore, there is increasing evidence that reading and listening comprehension processes are influenced to a considerable extent by comprehender-specific interindividual differences. Among others, the following factors are relevant: thematic and cultural background knowledge, interest and motivation, familiarity with the text type, (working) memory capacity, reading and comprehension ability in the native language, and general foreign-language proficiency (cf. the detailed discussion in Ehlers 1998; Urquhart/Weir 1998; Grotjahn 2000a).

With regard to the determination of task-difficulty factors, the implication of the above observations is that we must take account of:
a) the interaction between the task and the reader or listener, and
b) the dependency of the comprehension processes on interindividual variation in recipient characteristics which are not L2 specific.

3. Approaches to the measurement and prediction of task difficulty

3.1. Description of the problem

Due to the interaction of text characteristics with features of the items, the test takers and the instructions to the test takers, there are severe limitations on the possibility of making a prediction of text difficulty solely on the basis of features of the text. Thus, for example, items of various degrees of difficulty can be based on one and the same text. Or, the same score on two comprehension tasks can in one case be the result of easy items together with a difficult text and in the other case of difficult items together with an easy text. Moreover, the instruction to search for specific information in a difficult text which is largely incomprehensible to the test taker may nonetheless yield an empirically easy test task. Finally, a text from a field with which the test taker is

familiar is potentially easier than a text for which relevant background knowledge cannot be activated.

It has already been pointed out that it is possible to distinguish between text-centred and addressee-centred approaches to the determination of task difficulty. In text-centred approaches, factors of text difficulty are determined a priori without presenting the tasks to the test takers. By contrast, addressee-centred approaches attempt to identify the determinants of task difficulty on the basis of an empirical analysis of the test takers’ responses to the task.

3.2. Text-centred approaches

3.2.1. Readability research

In readability research, readability and thus also text difficulty are conceived as an inherent feature of the text. Individual reader characteristics are ignored. The goal is to predict text difficulty without reference to test takers, using a mathematical formula based on easily determined text features. The predictors most often used are word length (generally measured in terms of the number of letters per word), sentence length (generally measured in terms of the number of words per sentence), currency of the words used (determined on the basis of frequency lists) and the type-token ratio as a measure of lexical redundancy or density. In validating the prediction formula, among others the following criteria of (actual) difficulty/readability of the texts are used: a) expert ratings, b) performance on cloze tests, and c) performance on standardised reading tests.

In addition to simple predictors in common use such as word or sentence length, there are also theoretically more sophisticated predictors such as measures of syntactic complexity in terms of generative transformational grammar. Up to now, the validity of several hundred predictors has been investigated, and several hundred readability formulae have been published. I cannot discuss the
numerous possible criticisms of the readability approach, but shall restrict myself to a few illustrative examples.

Although as a rule text difficulty correlates with, for example, sentence length, this correlation must not be interpreted in terms of causality. Sentence length correlates for its part with the complexity of syntactic structure, and this correlates with the occurrence of conjunctions such as "because", "since" and "or". Hence, sentence length is not really the cause of difficulty. This implies that features on the word and sentence levels can only be valid predictors of comprehension difficulty if they actually correlate with features such as syntactic complexity or conceptual richness. However, Anderson/Davison (1988) have shown that under certain circumstances conjunctions such as "because", "since" and "or" can even make a text more easily readable. There are thus several reasons why inferring comprehension difficulty from sentence length is problematical. There is a similar problem when word length is used as a predictor. In German, for example, long words are often compounds. Transparent compounds are often far easier to understand than short, high-frequency words which can occur in numerous collocations (see also Davies/Irvine 1996:170 ff.).

In spite of numerous criticisms and fundamental deficiencies, the readability approach has in my opinion several positive aspects: (1) The empirical methodology is objective and reliable. (2) It is possible to evaluate the efficiency of the model empirically. (3) The procedure is easy to apply. Thus, despite its weaknesses, in view of the ease of application in particular, consideration should be given to the readability approach for selecting reading comprehension texts.

For example, the following procedure is conceivable: In a first step, experts make a preliminary selection of a relatively large number of texts taking criteria such as cultural and content neutrality into consideration. In a second step, the texts are computer analysed for readability on the basis of simple criteria. In addition to word length,

3.2.2. Expert ratings

In German-speaking countries, the best known approach to measuring text difficulty or comprehensibility on the basis of expert ratings is the Hamburg model developed by Langer/Schulz von Thun/Tausch (1974). The Hamburg researchers presented texts for rating on a large number of features on a seven-level bipolar scale. They then derived four dimensions of comprehensibility by means of factor analysis. The dimensions are as follows: (1) simplicity vs. complexity, (2) structuredness vs. unstructuredness, (3) brevity and conciseness vs. prolixity, and (4) additional stimulation vs. lack of additional stimulation. The dimension "simplicity vs. complexity", stands for features used in traditional readability research. The dimensions "brevity and conciseness vs. prolixity" and "additional stimulation vs. lack of additional stimulation" refer to macrostructural features which are generally not considered in traditional readability formulae because they are difficult to operationalise. Finally, the dimension "additional stimulation" stands for features such as "evoke", "interesting", "varied" and "exciting"; it thus measures motivational aspects. The comprehensibility of specific texts is measured in practice by having experts rate the text on the four dimensions using a five-level bipolar
scale (see also Groeben 1982:188 ff.; Christmann/Groeben 1999:178 ff.).

According to Sichelschmidt (1988) and Groeben (1982), the Hamburg approach has the following positive features:

a) it is easy to use in practice;
b) in addition to the readability indices in general use, it also takes account of macrostructural aspects;
c) it attempts to take motivational factors into account.

Among others, it has the following shortcomings:

a) the procedure was developed primarily by induction, and does not have a sufficient theoretical foundation;
b) the objectivity and validity of the rating is doubtful;
c) recipient characteristics are not sufficiently taken into consideration.
d) a rater training is required to be able to use it;\(^3\)
e) the experts’ rating is based on their own text reception, and thus leaves specific aspects of text reception on the part of the intended recipients out of account.

This last problem will probably be exacerbated if texts and items are rated by native speakers as Fortus/Coriat/Fund (1998:64) point out:

„Factors that affect difficulty in the eyes of a judge may not be those factors that affect difficulty for the examinee, particularly if the judges are native speakers of the language being tested and the examinees are not."

In spite of these criticisms, at least as far as methodology is concerned, the procedure described is basically appropriate.\(^4\) Expert rating can be used to take account of important features of the text and

\(^3\) At any rate, in case of assessment of text difficulty by human raters, special training is indicated. Thus, Fulcher’s (1997) research has shown that experts can differ to a huge extent in their judgements.

\(^4\) In case of listening comprehension, specific features such as the acoustic quality of the presentation must, of course, taken into consideration.

of the recipient – albeit with the restrictions mentioned above. Moreover, it is a relatively simple and economical procedure.

It is also conceivable that expert ratings and readability formulae be used conjointly to determine text difficulty. This view is also advocated, for example, by Fulcher (1997), who attaches more importance to expert judgements however.

\[3.2.3.\ ] A priori task analysis: within the framework of task-based language assessment

\[A priori\] determination of task difficulty within the framework of task-based language assessment is possible both on the basis of expert ratings and objectively determinable task characteristics. The most farreaching attempt to predict task difficulty using task-based language assessment is to my knowledge that of Norris/Brown/Hudson/Yoshioka (1998) (with regard to task-based language assessment see also Robinson/Ross 1996 and Skehan 1998, Chap. 7). Norris et al. (1998) describe in detail numerous authentic tasks including necessary realm. A (tentative) difficulty index is assigned to the tasks according to the following three dimensions:

\[a)\ ] „code complexity“ with the features „range“, „number of input sources“, „delivery of input“;
\[b)\ ] „cognitive complexity“ with the features „amount of information to be processed“, „organization of input“, „availability of input“;
\[c)\ ] „communicative demand“ with the features „mode“, „channel“, „response level“ (see p. 77).

Note in particular that the authors regard the cognitive processing characteristics of the tasks as a central factor of difficulty.
3.3. Addressee-centred approaches to determining task difficulty

3.3.1. Some remarks on the state of research

The crucial feature of addressee-centred approaches to determining task difficulty is that while text-centred approaches can be used in the choice of texts and the construction of tasks, factors of task difficulty are ultimately determined on the basis of empirical analysis of the functioning of the items with a sample of test takers. Accordingly, addressee-centred approaches can also be characterised as item-centred.

There is a number of different approaches to addressee-centred determination of task difficulty. Here I shall restrict myself to those approaches which attribute difficulty to a complex interaction of item-specific, text-specific and examinee-specific features.

A relatively early and theoretically sophisticated approach to determining the difficulty of reading comprehension tasks with opendended responses was developed by Pollitt/Hutchinson/Entwistle/De Luca (see Pollitt/Hutchinson 1986). These authors proposed an explicit model of the process of answering questions on text passages, predicting the difficulty of each question with the help of a regression formula using task-specific and textual characteristics. Pollitt/Hutchinson (1986:55 f.) identified among others the following features of difficult questions:

1. the wording of the question is potentially ambiguous;
2. the relevant text contains content words – in particular ones denoting attitudes or emotions – which tend to dominate functional details;
3. the relevant information is spread throughout the text;
4. the question consists of two or more parts which are related to each other in a complex way;
5. the answer requires more than simply quoting from, or minimally transforming, the text;
6. an incorrect previous answer may lead to an incorrect subsequent answer;
7. the sentence structure of the text is complex.

A model for the prediction of the difficulty of the multiple-choice tasks (M-C tasks) of a standardised reading-comprehension test for English as a second language was developed on the basis of theories taken from cognitive psychology by Perkins/Brutten (1993). After empirically testing the model, the authors came to the following conclusion:

"... difficult items have (1) texts that have a high proportion of content words, (2) many different predicates densely nested, and (3) cognitive demands that require higher-order processing." (p. 217)

As do most authors, Perkins and Brutten use a regression model for prediction. The above-named characteristics explained 72% of variance.

3.3.2. Buck/Tatsuoka/Kostin's (1997) rule-space analyses

In an innovative study Buck/Tatsuoka/Kostin (1997) analysed the 40-item multiple-choice reading comprehension part of the Test of English for International Communication (TOEIC) on the basis of a random sample of 5,000 Japanese learners of English. The TOEIC M-C reading comprehension part consists of a number of short texts of 2 to 4 items each. In addition, there are two further reading comprehension sections: in one, the test taker has to recognise errors in sentences, in the other, fill in a blank within a sentence. In sum, the reading comprehension sections comprise 100 items. On the basis of a comprehensive review of the pertinent literature, the authors established an extensive list of theoretically and empirically founded determinants of task difficulty. It contains three types of variables: (1) characteristics which relate exclusively to the text, as for example vocabulary and syntax of the text; (2) characteristics which refer only to the M-C items such as length of options; and (3) characteristics which relate to both text and item, as for example the extent of lexical overlap between text and items (cf. also Freidel/Kostin 1999). Since the attributes used were of a surface-level type, it was relatively easy to decide whether or not a characteristic applied to a task.
The authors then coded the 40 M-C items dichotomously on the item characteristics, and made inferences about the cognitive abilities needed to perform the tasks. They thus also took cognitive features of the test takers explicitly into account. On the basis of several empirical item analyses they then established a list of 24 task characteristics, eight of which related to two- or three-way interactions among attributes.

Buck/Tatsuoka/Kostin (1997) did not use the usual regression approach to empirically determine the importance of specific attributes, but rather carried out a so-called rule-space analysis involving an innovative statistical pattern-recognition and classification technique. With the help of rule-space analysis 91% of the test takers could be given a score on each task attribute. Regression of the total test score onto the 24 task attributes resulted in an adjusted R-square of .97, indicating that the attributes explained 97% of the variance for the 91% of the test-takers successfully classified. This is a surprisingly strong relationship between task attributes and total test score.

In Table 1 a selection from the list of the 24 task attributes is presented together with some important statistics.

Table 1
Statistical characteristics of some cognitive/linguistic task attributes

<table>
<thead>
<tr>
<th>Label</th>
<th>Attribute</th>
<th>Task Characteristic Coded</th>
<th>M</th>
<th>SD</th>
<th>r₁</th>
<th>r₂</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK6</td>
<td>The ability to use a word-matching strategy in selecting the correct option.</td>
<td>There is an above average amount of lexical overlap (exact words and lexically related words) between the</td>
<td>.99</td>
<td>.09</td>
<td>.28</td>
<td>.24</td>
<td>18</td>
</tr>
</tbody>
</table>

Determinants of the difficulty of comprehension tasks

<table>
<thead>
<tr>
<th>PI3</th>
<th>The ability to hold information in memory and use it to make an inference.</th>
<th>The item requires an inference based on information spread over two sentences.</th>
<th>.79</th>
<th>.28</th>
<th>.62</th>
<th>.52</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI4</td>
<td>The ability to hold information in memory and use it to make an inference.</td>
<td>The two items of information for the inference are scattered across the text.</td>
<td>.53</td>
<td>.35</td>
<td>.74</td>
<td>.67</td>
<td>3</td>
</tr>
<tr>
<td>PI1 b</td>
<td>The ability to identify the gist of a passage.</td>
<td>The item is a main idea item.</td>
<td>.83</td>
<td>.26</td>
<td>.55</td>
<td>.51</td>
<td>5</td>
</tr>
</tbody>
</table>

Interaction Attributes

<table>
<thead>
<tr>
<th>TC1 x PI4</th>
<th>The ability to hold information in memory and use it to make an inference when the text is laid out in dense continuous formatting.</th>
<th>.51</th>
<th>.28</th>
<th>.65</th>
<th>.60</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC1 x PI18</td>
<td>The ability to understand the gist when the paragraph or segment is longer and the text is laid out in dense continuous formatting.</td>
<td>.51</td>
<td>.35</td>
<td>.73</td>
<td>.66</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: Buck/Tatsuoka/Kostin (1997:441-443)

Column 1 of Table 1 shows the variable labels used and column 2 the ability and knowledge purportedly involved in item solving. Ability and knowledge has been inferred from the coded item characteristics in column 3. In column 4, M stands for the average performance of all test-takers on the corresponding characteristic and thus measures its difficulty for the sample of test takers investigated. The lower M is, the higher the difficulty is. In columns 5-7, SD denotes the standard deviation, and r₁ and r₂ are discrimination indices. r₁ corresponds to the correlation between item characteristic and total score in the multiple-choice reading comprehension part (part 7 of the TOEIC), and r₂, to the correlation between item characteristic and total score in all three parts of the reading comprehension subtest. Finally, the last column shows
the number of items for the solution of which the characteristic was
deed necessary.

Besides mean difficulty, discrimination is of particular interest. The
discrimination indices of the 24 task characteristics are relatively high
in most cases, which testifies to the predictive power of the
characteristics determined with the help of the rule-space methodology.
Furthermore, the fact that the values for \( r_2 \) are almost always only
slightly lower than those for \( r_1 \) is an indication of the robustness of the
estimates. By the same token this means that the task characteristics are
not restricted to the prediction of the difficulty of the M-C items, but
also extend to the item formats of the other two parts of the reading
comprehension subtest.

The task characteristics and cognitive attributes listed in Table 1 are
to be interpreted as followings: IK6 relates to the extent of lexical
overlap between the correct option of a M-C item and the information in
the text necessary to solve the item, and measures the ability to match
textual information in the item with lexical information in the text. It is
a characteristic typical of M-C items tapping the micro-level of text
processing. For the sample investigated, it is an extremely easy
characteristic, its probability of success being 99%. Moreover, its
correlation with the two total reading scores is low, in part as a result of
its high facility and low standard deviation. The fact that IK6 is
involved in the solution of 18 of the 40 items indicates a potential
weakness of the M-C reading comprehension part.

PI3 and PI4 apply to items the solution of which requires an inference
based on information in the text. They thus measure both the capacity of
the working memory and the reasoning ability of a reader. PI3 and PI4
are considerably more difficult than IK6 and also discriminate much
better. As is to be expected on theoretical grounds, PI4 is more difficult
than PI3 because in the case of PI4 the information necessary for the
solution of the item is spread out over the whole text. Since PI4 also
discriminates very well, it appears to be an important characteristic.

Solution of an item becomes even more difficult if PI4 co-occurs with
a text layout which increases processing difficulty (interaction attribute

TC1 x PI4). Similarly, the characteristic PI1b ("the ability to identify
the gist of a passage") becomes much more difficult if it interacts with
TC1 ("dense continuous formatting") and IK4 ("the paragraph or
segment is longer").

Buck/Tatsuoka/Kostin/Phelps (1997) and Buck/Tatsuoka (1998) also
carried out rule-space analyses with listening comprehension tests.
Overall, the rule-space approach has several positive features:

1. determinants of task difficulty are justified both on theoretical and
   empirical grounds;
2. the procedure yields precise information as to the relative
   importance of specific characteristics including interaction
   attributes;
3. the information provided by rule-space is relevant with regard to
   both the selection of texts as well as the construction and revision of
   items;
4. the information can be used for reporting the test results to each test
taker in the form of a detailed profile and for making didactic
decisions in the classroom.

At least one problem should be mentioned. Up to now, the task
characteristics and statistical properties established with the help of the
rule-space methodology have been based exclusively on tests of English
as a foreign language. It cannot, however, be ruled out that, for
example, with German texts and items rule-space analyses may lead to
different results due to the high frequency of inflected forms in German.
Moreover, rule-space methodology has so far not been sufficiently
cross-validated.

3.3.3 Central determinants of task difficulty

On the basis of the relevant literature, I shall now attempt to establish a
list of some central determinants of difficulty. Because of the affinity of
reading and listening comprehension, I take research in both reading
and listening into account while at the same time ignoring modality-
specific variables such as rate of speech or page layout. The proposed list has primarily a heuristic and sensitising function, namely to draw attention to a number of aspects to which one should pay attention in constructing test tasks and in interpreting test results. Of course, the variables listed require further validation.

It should be noted that the compilation of the list involved a number of problems: the investigations on which the list is based differed, for example, with regard to the variables studied, the test takers and the tasks (e.g. reading, listening, M-C items, constructed responses). Moreover, the results were not very consistent (e.g. with regard to statistical significance).

The tentative list of variables is presented in Table 2. In accordance with Freede/Kostin (1999) the following three types of predictors of difficulty were distinguished:

1. **Item variables**: Pure item variables relate only to items and are coded without taking the test into account. An instance is: number of words in the item stem. Item variables also include 'flaws' in the construction of items such as the use of double negation.

2. **Text variables**: These variables refer only to the content and structure of the texts without taking the items into account. An example is the type-token ratio in a text.

3. **Text-by-item overlap variables**: These relate both to texts and items. An instance is: the information necessary for the solution of an item is located at the beginning of the text.

In most cases the description of the variables in Table 2 corresponds to that found in the literature. Operationalisations of most of the variables listed are described in Freede/Kostin (1999).

In compiling the list the following criteria were used: (1) The variable has proved to be a significant predictor of item difficulty and/or the overall test score; (2) the effect of the variables can be theoretically interpreted on the basis of a cognitive model of task processing; (3) the variable is operationalised in such a way that it can be relatively easily coded.

For the variables listed in Table 2, "yes" indicates a (relatively) high level of difficulty, "no" a relatively low level. In the case of quantitative variables (e.g. number of content words), a high value indicates a high degree of difficulty. The expression "relevant information" is a short form for "text or part of text in which the information required for the solution of the item is contained". As far as pure text variables are concerned, it should be noted that they can also function as text-by-item overlap variables if they pertain to that part of the text which is necessary for the solution of the item.

**Table 2: Central factors of difficulty of comprehension tasks**

<table>
<thead>
<tr>
<th>Item Variables</th>
<th>1-1</th>
<th>1-2</th>
<th>1-3</th>
<th>1-4</th>
<th>1-5</th>
<th>1-6</th>
<th>1-7</th>
<th>1-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguity of item wording</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity of item structure (the item consists of several intricately related parts)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of words in distractors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of difficult and/or unfamiliar (rare) words in distractors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of negations in distractors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of words in stem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of negations in stem (no, not, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of intertextual and extratextual referentials in correct option (he, this, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7 It should be kept in mind that the use of regression analysis to predict item difficulty is methodologically problematical.
4. Conclusion

It should have become clear that on the basis of a readability approach a text-centred a priori determination of the difficulty of comprehension tasks is possible only to a limited extent. Without additional addressee-centred analysis, prediction of task difficulty by raters is also not dependable. However, Fortus/Coraia/Fund (1998) have presented evidence that the a priori determination of the difficulty of reading tasks can be significantly improved by training the raters on the basis of the results of detailed text- and addressee-centred analyses. Moreover, these authors convincingly argue that item construction can profit from such analyses. Using a hierarchical linear modelling approach, Jensen/Hansen/Greer/Akeley (1997) also advocate this view on the basis of a cognitive linguistic analysis of a listening comprehension test. This supports my opening claim that the identification of the determinants of the difficulty of comprehension tasks is an issue of great theoretical and practical relevance.

Bibliography


Determinants of the difficulty of comprehension tasks


Tests and Translation
Papers in memory of Christine Klein-Braley

AKS-Verlag Bochum
2001