Computer-based assessment of reading skills[1]

Tobias Richter & Johannes Naumann[2]

In this paper, we present a recently developed and computer-based instrument for the detailed assessment of reading skills. The theory underlying its construction is van Dijk and Kintsch's (1983) *strategy model* of text comprehension. The target group of the instrument are adults with a presumably high level of reading ability, for instance university students. Therefore, we do not intend to assess difficulties in reading or achievement in learning to read. Apart from that, the subtests refer to basic cognitive processes of reading but not to metacognitive strategies or standards of comprehension. The instrument is designed for research purposes; we are planning to use it for the measurement of covariates in experiments on text comprehension (see Christmann, Groeben, Flender, Naumann & Richter, 1999).

We begin with a short discussion of some advantages and disadvantages of presently available methods for the assessment of reading skills. Next, we will argue that it is reasonable to ground the diagnostic efforts on a hierarchically structured model of component processes of reading like van Dijk and Kintsch's strategy model. Following a brief sketch of the strategy model, we will explain the structure of the instrument we have recently developed and illustrate its subtests by item examples. Finally, we shall report some empirical results concerning the psychometric properties of the instrument, that is its reliability and validity.

Why a new instrument for the assessment of reading skills?

Standardized reading comprehension tests and cognitive-psychological measures of reading ability

There are two major directions in the diagnosis of reading abilities: Standardized tests of reading comprehension on the one hand, and the differentiation of good and poor readers by measurements developed in cognitive psychology on the other hand. The historically older standardized tests of reading comprehension – the Nelson Denny Reading Test, the Verbal Scholastic Aptitude Test, to name just two examples – partly have a remarkable predictive validity, but at the price of a rather unclarified construct validity. From the perspective of the psychology of text processing, the tasks commonly used in such tests (for example multiple-choice items following an expository text) involve a lot of different processes which are

[1] Paper presented at the Computers in Psychology Conference (CiP 2000), March 29-31, 2000, York.

[2] University of Cologne, Department of Psychology. The research reported here has been conducted as part of a research project supported by the German Research Association (*Deutsche Forschungsgemeinschaft* – DFG; grant awarded to Dr. Ursula Christmann and Prof. Dr. Norbert Groeben). We would like to thank the DFG for this support. We would also like to thank Nadine van Holt and Katharina Gauss for their help in collecting the data.

confounded in the measure; therefore, the classical reading comprehension tests do not tell us precisely in which cognitive abilities good and poor readers differ from each other (e.g., Daneman, 1982). This has been the starting point for cognitive-psychological research on reading abilities which has its origins in the early seventies. Like the research on cognitive correlates of intelligence, the research on reading abilities aims at reducing interindividual differences in reading ability to theoretical constructs of cognitive psychology (see Perfetti, 1985, for a review). The first efforts resulted in several rather artificial or restricted measurements which – as a rule – were only moderately correlated with classical reading comprehension scores. Examples are *letter span*- or *digit span*-tasks as measurements of the capacity of short term memory (e.g., Guyer & Friedman, 1975; Perfetti & Goldman, 1976), or measurements of the speed of lexical access. In contrast to this, Daneman and Carpenter (1982) proposed a double task which proved to be empirically successful, the so called *reading span*. In the reading span-task, subjects have to read and comprehend a sequence of unrelated sentences, and are instructed to keep the last word of each sentence in memory. The reading span can be interpreted as a dynamic measure of working memory capacity (Just & Carpenter, 1992). It usually shows substantial correlations with classical reading comprehension scores, and even high correlations with naturalistic reading tasks for which recently encountered information must be related to information given much earlier in the text (for example, the resolution of pronominal reference).

Arguments for grounding the assessment of reading skills on a hierarchical model of text comprehension

Obviously, the reading span is a valid measure of working memory capacity and captures a quantity which plays a central role in several models of text processing (most evidently, in Kintsch and van Dijk's, 1978, model of cyclical processing). Nevertheless, for a comprehensive assessment of reading ability a single measure like the reading span is not sufficient. Not only are there differences between good and poor readers in properties which cannot be put down to working memory capacity. There is an aspect which is even more important: In many cases, deficits in certain component processes of reading can be compensated for by other component processes. Meanwhile, a lot of empirical findings corroborate this assumption on different levels of text comprehension (see Kintsch, 1998, pp. 282-289, for an overview). For instance, there are interindividual differences in the speed of lexical access as well as in the speed of integrating word meaning and sentence context; but slow decoding of single words does not necessarily lead to slower reading times if the words are presented in an adequate sentence context (Stanovich, 1980). For the diagnosis of reading skills, the distinction between microstructural and macrostructural processes seems especially relevant. Deficits in processes at the microstructural level, such as establishing local coherence, *can* but *need not* impair processes at the macrostructural level, such as comprehending the gist of a text or constructing an adequate situation model of the text content (e.g., Graesser, Hoffman & Clark, 1980; Schneider, Körkel & Weinert, 1989). In comprehending narratives, for example, readers can use their knowledge about typical causal relations in order to compensate for low verbal abilities (Bisanz, Das, Varnhagen & Henderson, 1992).

A sketch of the strategy model

All in all, this has lead us to the conclusion to base the construction of our own instrument on a hierarchically structured model of text comprehension, namely the strategy model (van Dijk & Kintsch, 1983). The processes distinguished by this model belong to one of two major categories, processes *low* in hierarchy (or microstructural processes) and processes *high* in hierarchy (or macrostructural processes). *Propositional strategies* and *local coherence strategies* both represent types of lower processes. By the use of propositional strategies, readers are assumed to build up a propositional representation of the text. For this purpose, they have to decode the meaning of words, to disambiguate word meanings by the use of sentence context, and to combine word meanings to elementary propositions. *Local coherence strategies* associate elementary propositions to a propositional representation of the text base, by simple mechanisms such as argument overlap or coreferentiality of expressions. While these processes mainly rely on basic linguistic knowledge, the higher strategies make a stronger use of world knowledge. *Macropropositional strategies* generate so called macropropositions, that is representations of the main statements or topic sentences of a text, by deletion, generalization, or construction. The efficiency of macropropositional strategies heavily depends on the availability of domain specific knowledge. In contrast to this, *schematic strategies* make use of so called super structures, that is knowledge about typical structural properties of certain kinds of texts (for example, letters, expository texts, or fairy tales). In addition, *rhetorical strategies* in general refer to pragmatically relevant knowledge, such as knowledge about the author, his or her intentions, the discourse context and so on.

Description of the instrument

The seven subtests of our computer-based instrument for the assessment of reading abilities are each related to one or two types of strategies assumed by the van Dijk and Kintsch-model (with the exception of schematic strategies). Thus, we intend to assess interindividual differences in component processes of reading in a detailed way. Each subtest consists of 10 to 15 items. Subjects respond by one of two response keys or – in the case of one subtest – by clicking the mouse. For all items, accuracy and latency of responses are collected. For three subtests, there are also short texts to be read, and the reading times for these texts are recorded. Most of the subtests include practice trials with feedback for the accuracy of responses. It takes approximately half an hour to go through the whole test.

Let us now turn to the subtests themselves. In order to assess interindividual differences in propositional strategies and local coherence strategies, the instrument includes sentence verification tasks (1), judgements of the meaningfulness of sentences (2), a vocabulary test (3), and judgements of the meaningfulness of sequences of sentences (4). For the assessment of higher processes, firstly there is a test similar to classical reading comprehension tests (5). Following two short texts, topic statements (at macrostructure level) are presented; subjects are instructed to judge whether the statements were contained in the respective text. Furthermore, the instrument includes a subtest related mainly to rhetorical strategies (6). In this subtest, we ask subjects to classify statements of a text as either statements of facts or opinions of the author. The last subtest refers to rhetorical strategies as well as macrostrategies; again after reading a short test, subjects are asked to decide if items represent implications of the text or if they don't (7).

Subtests with item examples

In this section, we briefly describe the seven subtests of the instrument and give an item example for each of them (the examples are translated from German).

(1) The 15 sentence verification tasks of the first subtest contain conceptual statements of commonly known concepts and vary with respect to syntactic and semantic complexity. We assume that this task requires different processes which belong to the category of propositional strategies.

Item example:
"Strawberries are a red and sweet kind of vegetables." (response categories: *true* / *false*)

(2) In the second subtest, subjects decide for 15 sentences if they represent meaningful statements or if they do not. Again, we assume that propositional strategies and local coherence strategies like decoding of word meanings or establishing meaning relations between individual expressions are necessary for this task.

Item example:
"When you eat rotten food the skin often changes its colour in a full-sounding way." (response categories: *meaningful* / *meaningless*)

(3) The vocabulary test is designed primarily to assess the availability of the meaning of relatively rare and difficult words (15 items). As it contains explanations of word meanings in the form of statements, additional propositional strategies should also be required.

Item example:
"The expression 'determinant' means 'causal factor'." (response categories: *true* / *false*)

(4) In the fourth subtest, subjects are asked to decide for 15 pairs of sentences whether they stand in a meaningful relation to each other or whether they do not. After reading the first sentence, subjects press a key and the second sentence appears. Clearly, local coherence strategies are central for this task.

Item example:
"The psychologist Dietrich Dörner is a respectable scientist in the domain of problem solving." (1st sentence)
"Therefore, it has been awarded the famous Leibniz-Preis." (2nd sentence)
(response categories: *meaningful* / *meaningless*)

(5) For the reading comprehension test, two texts about a sociological and a technical topic (social classes and ship building) have to be read; after reading each text, 7 items have to be judged if they were contained in the text. These items represent statements on macrostructure level.

Item example:
"Profits gained at the stock market cannot be regarded as income." (response categories: *contained in the text* / *not contained in the text*)

(6) The 10 items of the subtest in which facts and opinions have to be distinguished, are taken from a text about personality dispositions and genetic influences on social affairs. Respondents judge the items by clicking the mouse while the text is still on the screen.

Item example:
"Heredity and environment are both relevant for the development of the phenotype." (responses on a 6-point Likert-type scale, with the end points labeled *is a fact* / *is an opinion*)

(7) The 10 items of the implications-subtest refer to the same text than the latter subtest. Before the items are presented, the text disappears from the screen.

Item example:
"Large parts of the societal order are determined by biology." (response categories: *implication* / *no implication*)

Scoring of the responses

The three variables collected – accuracy, response latency, and reading times – are combined into integrated test scores. We assume that the different component processes of reading can be regarded as efficient if the processing leads to a correct result and does not take too long. Therefore, in most of the subtests an item is scored as "solved" if the task is accomplished correctly and at the same time within the lower two quartiles of the item-specific distribution of response latencies; these "norms" were obtained by employing a different sample of the target population (namely, university students). For the three subtests which require the reading of texts, the individual reading times are also included. In these cases, the median of the product of the decision latencies and the reading times is taken as the cut-off value.

Reliability and validity of the instrument

The present version of the instrument already is the outcome of scale revisions based on a construction sample of 50 subjects. Furthermore, we have tested the optimized version employing a sample of 102 university students. For about half of this sample, there are also questionnaire data available which comprise self-ratings and information concerning the subjects' actual reading behavior.

Reliabilities

As can be seen from Table 1, the items of most of the subtests are dealt with with a high degree of accuracy; only in the vocabulary test, the text comprehension test, and the text implications tests the mean subjects make more mistakes. On the other hand, the decision latencies show a rather high variability. The internal consistencies estimated using the integrated test scores are generally satisfactory, with the exception of the vocabulary test that shows at the most an acceptable reliability.

Table 1: Means and standard deviations for accuracy and latency of responses, and internal consistencies for the integrated test scores.

Subtest	*Items*	M_{accuracy}	SD_{accuracy}	M_{latency}	SD_{latency}	α
Sentence verification	15	14.15	0.84	2928	892	.87
Meaningfulness (sentences)	15	14.53	0.84	3914	1283	.89
Vocabulary	15	11.97	1.59	4899	1430	.71
Meaningfulness (sequences of sentences)	15	14.47	0.67	3277	1283	.84
Text comprehension (macropropositions)	14	10.35	1.41	6626	2124	.83
Distinction of fact vs. opinion	10	8.98	1.04	11230	6519	.89
Text implications	10	6.53	2.07	6331	2418	.83

Notes. α = Cronbach's Alpha for the integrated test scores. N = 102.

Scale intercorrelations

First pieces of evidence for the construct validity of the instrument (in terms of the strategy model) are revealed by the scale intercorrelations. As can be taken from Table 2 (submatrix on the upper left), the four subtests relating to lower strategies, namely sentence verification, meaningfulness judgements of sentences or sequences of sentences respectively, and the vocabulary test, are highly intercorrelated. The same holds for the three subtests relating to higher strategies, that is the text comprehension test, distinction of facts vs. opinions, and text implications (see the submatrix on the lower right of Table 2). The correlations between subtests from different categories reach significance in eight out of twelve cases, they are generally lower although in some cases substantial. This pattern is theoretically plausible because the subtests relating to either lower or higher strategies should assess similar processes. At the same time, the accomplishment of the subtests requiring higher strategies is certainly not independent from the efficiency of lower processes. A principal components analysis of the correlation matrix of the seven subtests clearly results in two factors (explaining 75 percent of variance) which combine the four subtests relating to lower processes on the one hand and the three subtests relating to higher processes on the other hand.

Table 1: Means and standard deviations for accuracy and latency of responses, and internal consistencies for the integrated test scores.

	1 SV	2 MF I	3 VO	4 MF II	5 TC	6 FO
1 Sentence verification (SV)						
2 Meaningfulness (sentences) (MF I)	.72**					
3 Vocabulary (VO)	.65**	.73**				
4 Meaningfulness (sequences of sentences) (MF II)	.63**	.62**	.58**			
5 Text comprehension (macropropositions) (TC)	.27**	.30**	.39**	.23*		
6 Distinction of fact vs. opinion (FO)	.25*	.17	.20*	.14	.52**	
7 Text implications (TI)	.17	.26*	.31**	.16	.46**	.75**

Notes. *N* = 102. * *p* = .05, ** *p* = .01 (2-tailed).

Correlations with actual reading behaviour and self-ratings

For validation purposes, we have also collected detailed information about the actual reading behaviour from about 50 subjects, for instance, how many texts (broken down into different categories) they have read for their studies or in their spare time. Unfortunately, we found no systematic relationships between these data and the test scores. We would attribute these zero-results rather to an inappropriate choice of the criteria than to the instrument to be validated; the number of texts that people usually read probably depends on a lot of factors including academic interests, study obligations or the fun someone sees or does not see in reading. The other way round, it is rather implausible that the basic reading abilities of adult readers should profit from the amount of reading activities within the last twelve months.

Table 3: Items of the self-rating scales related to "higher" and "lower" processes (4-point Likert-type scales)

Items related to "lower" processes:

1. Frequently I have to read a sentence several times in order to understand its meaning.
1. While reading a text, it rarely happens to me that I do not know what a particular word means.
2. Usually, I recognize immediately if two subsequent sentences in a text do not match grammatically.

Items related to "higher" processes:

1. Even if I only glance over a text I am able to grasp the main points.
2. It is generally easy for me to infer the views or the position of the author of a text I am reading.
3. I am able to read a text quickly and comprehend it.

Table 4: Correlations of the subtests with the self-ratings related to "higher" and "lower" processes

Subtest	Self-ratings "lower" processes	Self-ratings "higher" processes
Sentence verification	.23	.10
Meaningfulness (sentences)	.24*	.00
Vocabulary	.34**	.13
Meaningfulness (sequences of sentences)	.29*	.02
Text comprehension (macropropositions)	.29*	.34**
Distinction of fact vs. opinion	.22	.25*
Text implications	-.02	-.07
Factor scores lower processes	.29*	.03
Factor scores higher processes	.25*	.34**

Notes. *N* = 52. * *p* = .05, ** *p* = .01 (2-tailed).

There are, however, some indications of the validity of the instrument which can be inferred from the relationships of the test scores with self-rating data. We applied two short scales with three items each; the subjects filled in the questionnaire approximately one week *before* the reading tests were conducted. One of these scales is meant to assess self-ratings of one's own efficiency in the application of lower reading strategies, complementarily, the other scale should assess self-ratings of the efficiency of higher strategies (see Table 3). As can be taken from Table 4, three out of four subtests relating to lower strategies show correlations with the self-ratings for lower processes but there are no correlations with the complementary scale. Correspondingly, most of the subtests relating to higher strategies (with the exception of the implications-subtest) are correlated with their respective self-rating scale; the text comprehension test is also correlated with the self-ratings for lower processes. A similar picture results from the correlations with the factors that combine either the subtests relating to higher processes or the subtests relating to lower processes.

Conclusions

The indications of validity obtained so far are encouraging but, of course, by no means sufficient. However, we are planning further validation studies. At the moment, we are using the instrument for the assessment of covariates in an experiment concerned with learning from linear text compared to learning with hypertext. In additional validation studies we are going to relate the scores obtained with the recently developed instrument to classical reading comprehension tests with texts of different topical domains and measures of domain specific knowledge; if the instrument could provide an increment in explaining the variance of the classical reading comprehension scores (across domains) this could be regarded as evidence for its criterion validity. In addition, information about the construct validity of the instrument could be gained by exploring its relationships to cognitive-psychological measures (such as the reading span). Finally, in order to examine discriminant validity, we are planning to test if the scores obtained with our instrument can be separated from measures of general cognitive ability, such as language-independent measures of intelligence.

References

Bisanz, G. L., Das, J. P., Varnhagen, C. K. & Henderson, H. R. (1992). Structural components of reading time and recall for sentences in narratives: Exploring changes with age and ability. *Journal of Educational Psychology, 84*, 103-114.

Christmann, U., Groeben, N., Flender, J., Naumann, J. & Richter, T. (1992). Verarbeitungsstrategien von traditionellen (linearen) Buchtexten und zukünftigen (nicht-linearen) Hypertexten. In N. Groeben (Ed.), *Lesesozialisation in der Mediengesellschaft* (pp. 175-189). Tübingen: Niemeyer. [Processing strategies of traditional (linear) book texts and future (non-linear) hypertexts]

Daneman, M. (1982). The measurement of reading comprehension: How not to trade construct validity for predictive power. *Intelligence, 6*, 331-345.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450-466.

Graesser, A. C., Hoffman, N. L. & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior, 19*, 135-151.

Guyer, B. L. & Friedman, M. P. (1975). Hemispheric processing and cognitive styles in learning-disabled and normal children. *Child Development, 46,* 658-668.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A. & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior, 14*, 33-42.

Schneider, Körkel & Weinert (1989). Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology, 81,* 306-312.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16,* 32-71.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Correspondence:

Tobias Richter
University of Cologne
Department of Psychology
Chair II: General and Cultural Psychology
Herbert-Lewin-Str. 2
D-50931 Koeln, Germany

Phone: +49-(0)221-4703848
Fax: +49-(0)221-4705002
E-mail: tobias.richter@uni-koeln.de