

Screening reading comprehension in adults: Development and initial evaluation of a reading comprehension measure

*René T. Proyer¹, Michaela M. Wagner-Menghin² &
Gyöngyi Grafinger³*

Abstract

Reading comprehension in adults is a rather neglected variable in the practice of psychological assessment. We propose a new screening instrument for adult reading comprehension based on a pragmatic definition of reading comprehension as the textual understanding of the text read. Using data from a calibration sample ($n = 266$) and a replication sample ($n = 148$) for cross-validation, we tested the model fit for the 1-PL model (Rasch-model; graphic model test, Anderson's Conditional Likelihood-Ratio test). Model fit was established and verified in the replication sample after the stepwise exclusion of three (out of 16) items. Correlations with a test for memory and the external criterion reading proficiency were in the expected direction. The comparison of a sub-group of putatively highly skilled readers ($n = 59$; University students and lecturers) and putatively low skilled readers ($n = 122$; participants undergoing psychological assessment for having their driving license reinstated after a ban) showed that a percent rank < 10 in the measure might indicate insufficient reading skills for practical purposes. Pending further research, the instrument seems to be a useful instrument for the screening of reading comprehension skills in adults.

Keywords: Computer Aided Testing, Item-Response Theory, Reading Comprehension, Test development

¹ *Correspondence concerning this article should be addressed to:* Dr. René Proyer, Department of Psychology, University of Zurich, Binzmühlestrasse 14/7, 8050 Zurich, Switzerland; email: r.proyer@psychologie.uzh.ch

² University of Vienna, Austria, Medical University Vienna, Austria

³ University of Vienna, Austria

Reading comprehension is a central prerequisite for many communication processes in the everyday life of adolescents and adults. One needs to read and comprehend written information in official forms, contracts at work, leaflets informing important decisions (e.g., when voting or when buying something) as well as health and care related information; see e.g., Doak, Doak, & Root, 1985). Reading comprehension is also an important factor in the process of psychological testing because often assessment instructions, questionnaire-items and test-tasks are presented in writing and have to be read and comprehended to perform. However, given the results of international studies on reading comprehension (e.g., Schwantner, Toferer, & Schreiner, 2013) it cannot be taken for granted that all test takers fully comprehend the content and meaning of each questionnaire item, of verbally given instructions or the verbal materials in ability tests. Furthermore, not all test takers ask the instructor for further explanations in case of comprehension problems.

This is a problem, when an individual's reading comprehension level impacts test results on constructs being associated with reading comprehension (e.g. memory, grammar or vocabulary knowledge) and on constructs not being associated with reading comprehension (e.g., extraversion, attention). Persons low in reading comprehension may have difficulties to follow detailed instructions or to understand an items' meaning, thus they may give responses arbitrarily with negative impact on their test scores and even serious implications for the test taker (e.g., in traffic psychology when evaluating adults with a record for risky driving).

In practice, the basic *skill* of reading comprehension is frequently not explicitly assessed objectively in adults (see also Baghaei, & Grotjahn, 2014; Messick, 1989; Vellutino, Scanlon, & Tanzman, 1998). Test takers are either assumed to have a sufficient level of reading comprehension, or it is assessed unsystematically by observing the test-taker's behavior during the completion of questionnaires and tests.

From an assessment practitioner's perspective we would benefit from considering reading comprehension by assessing it objectively prior to psychological testing. However, we experienced difficulties in finding appropriate instruments on reading comprehension in adults fitting in tight time schedules of a routine psychological assessment. The currently available tests on reading skills in adults typically use a compound-model of reading and comprehension to give a detailed picture on a set of reading related variables. And although they include reading comprehension measures they often require too much testing time. For example, Richter and van Holt's (2005) instrument requires a total testing time of more than 30 minutes to give seven different indicators of reading ability, based on the microstructural and macrostructural processes of reading as proposed in the Dijk and Kintsch model (1983; for other measures in adults see e.g., Jastak & Wilkinson, 1984; Jones, Long, & Finlay, 2006; Leslie & Caldwell, 2001; Woodcock, 1998). By way of this study, we provide a suggestion for a short screening measure of reading comprehension that can be used to assess reading comprehension in adults in routine practice.

Development of the reading comprehension measure: Theoretical background

The inventory was developed based on a simple definition of the construct: *Reading comprehension is the textual understanding of the text read* (e.g., Grisseemann, 1986) similar to the subtest *text understanding* (“Textverstehen”) by Richter and van Holt (2005). We acknowledge that this is a pragmatic approach to test development and that a broad range of literature exists that provides a much more fine-grained understanding of what reading comprehension is and how it develops (e.g. Frith, 1985; Verhoeven, & van Leeuwe, 2008). However, such detailed models inform the development of our stimulus material and the following components of the reading comprehension were considered: (a) *Word understanding* (i.e., the integration of single words to a complete text is, in general, an automatic process in adult readers. No voluntary controlled attention is needed; see also Katz, Branacazio, Irwin, Katz, Magnuson, & Whalen, 2011); (b) *vocabulary* (for the recognition of words and for the understanding of the whole text the actual size of vocabulary is of high relevance; it can be seen as a precondition that facilitates further processes in the understanding of a text; of course, it does not exist independently from the educational level; see e.g., Ouellette, 2006); (c) *sentence understanding*. The interpretation of a sentence starts with the first word (see e.g., Just, Carpenter, & Woolley, 1982) and structural assumptions on the relevance of the meaning of the read text are taken and modified or approved according to the gain of information from the following words (Rickheit, & Strohner, 1993). Thus, during the reading process assumptions on the most likely sentence structure will be developed, accepted, or if necessary revised (cf. Langenmayr, 1997); (d) *understanding of the text*; and (e) *memory*. Text understanding should not be understood as an all-or-nothing-process. Several authors (e.g., Langer, Schulz von Thun, & Tausch, 1981) have developed criteria for the comprehensibility of a text. Factors like simple verbal formulation, good structure, shortness, and conciseness of the verbal effort in relation to the aim of information, or stylistic characteristics play an important role. Furthermore, influences due to memory processes have to be taken into account. Word recognition is impossible without memory and there is a great deal of literature on the relation between working memory capacity and reading comprehension (e.g., Haarmann, Davelaar, & Usher, 2003; Hannon, 2012; Mellard, Anthony, & Woods, 2012; Norman, Kemper, & Kynette, 1992; Waters, 1996).

It was expected that each of these components help to explain a total score for reading comprehension without fully overlapping. For example, memory is necessary for being able to answer questions on a text which has been read before answering questions. Nevertheless, it was expected that correlations with a memory test will be far from indicating redundancy.

Development of the test material. The practical implication of our theoretical assumptions is that the items of a new instrument have to deal with the following aspects: (a) *Content* (to assess the textual understanding); (b) *conclusions* (independent conclusions have to be drawn, which are prerequisites for the textual understanding); (c) *meta-cognitions* (there are inconsistencies in the story which have to be considered for the understanding of the content; an example from a previous study would be a text where a person at the beginning of the story likes his nickname, and dislikes it in later parts of the

story without any obvious reason); (d) *mental models* (concepts have to be understood and reproduced); and (e) *conjunctions* (contents have to be understood, set in the right context, and applied).

An unpublished short story ("*Traffic light*", 695 words) from a professional author was selected and adapted to meet the criteria described above. The story was carefully chosen to also meet other criteria like general comprehensibility or structural issues. At its end an unexpected twist has to be integrated into the full context and, thus, the reader has to deal deliberately with conjunctions. This strategy was used because a special focus had to be set on the textual understanding that is different from the activated script in the text.

Development of the items. Sixteen questions and answers as well as suitable distractors for the multiple-choice format were developed covering the five components we described earlier. The distractors were carefully chosen so that each of them may be plausible and equally attractive for the test takers. For each item seven answer possibilities were developed and two further options were added; i.e., "None of the answers is correct" and "I do not know the answer." The latter should give the test takers the possibility of not having to guess but to answer honestly, if actually not knowing the answer. In each case at least one answer possibility is correct, in some cases more than one. A question is answered correctly *only* if all answer possibilities are selected which had to be selected and those, which had not to be selected actually were not selected; hence, each item is scored dichotomously (correct/incorrect).

A pre-study with a student sample ($n = 40$ between 20 and 65 years) was used to evaluate the story and the items (including all answer possibilities). Feedback from the participants led to minor refinements in the presentation of the test material and in the formulation of specific items.

The text as well as the test-items were presented as a computerized power test. The task of the test taker is to read a story and, afterwards, to answer multiple-choice questions dealing with the text without having the opportunity to scroll back to the text.

Aims of the present study

The study had two main research aims:

(1) *Evaluation and cross validation of the newly developed test items' psychometric properties.* The main aim of the present study was to develop a short screening instrument for reading comprehension in adults. We do not aim for testing a component-model of reading and reading comprehension, but to derive a global estimation of the latent variable reading comprehension. Although it is not intended to provide scores for each of the components, we consider them as a framework to develop valid test content. The main aim is to develop a unidimensional global measure of reading comprehension.

(2) *Testing the construct validity.* One possibility to collect hints indicating construct validity is to compare scores from two groups who putatively differ from one another regarding the dimension of interest. Data from a group of potentially *highly skilled* persons in reading comprehension (i.e., graduate students and scientists working at the

University) was compared with a group of potentially average to low reading competencies (selected from a group of drivers undergoing psychological assessment for having their driving-license reinstated after a ban). In a direct comparison, higher scores in the measure in those presumably highly skilled will be interpreted as support for the validity of the new measure. Further, it was expected that participants indicating to read a lot in their free time will score higher than those describing themselves as “non-readers.”

Another possibility to establish construct validity is to theorize about the new measure's correlation to other constructs. Memory is considered to be an important aspect of reading comprehension. Hence, there should be a positive association between the reading comprehension score and memory, but this overlap should not indicate redundancy. Overall, we expected a positive correlation of moderate size.

Method

Samples

Sample for evaluating psychometric properties. The calibration sample, consisted of $n = 266$ German-speaking participants (135 male, 131 female) between 18 and 66 years ($M = 30.59$, $SD = 9.07$). Of those, 13 (2.8%) had less than nine years of education in school, 109 (23.9%) had a completed vocational training (that means 10 to 12 years of education, including school), 300 (65.6 %) had a school-leaving diploma qualifying them for university (that means 12 to 13 years of education in school), and 35 (7.7 %) held a university degree. This means that persons with higher educational level are over-represented in the sample. The sample consisted of the following sub-groups; (a) Students and assistants from the Institute for Machine Elements at the Technical University of Vienna, and psychology students from Vienna University; (b) Persons who needed a traffic psychology examination, and were assessed by the “*Angewandte Psychologie und Forschung GmbH*” (AAP; Austrian Applied Psychology Company) in various Austrian cities (Vienna, Leoben, Graz, and St. Veit); and (c) Persons that were assessed in a selection procedure for a University of Applied Sciences in Vienna.

Sample for cross validation. A sample of $n = 148$ psychology students and applicants for a University of Applied Sciences in Vienna (29 males and 119, females; aged 21-44 years; $M = 25.50$; $SD = 4.38$) was tested.

Samples for establishing construct validity. The traffic psychology sample (i.e., group (b); $n = 122$ German-speaking adults; 107 males and 15 females) was used for an analysis of two putatively extreme groups (with comparatively lower vs. higher reading skills). As a comparison, a high potential group consisted of a group of German-speaking graduate students and teaching and research associates at Vienna's Technical University (i.e., group (a); $n = 59$; 51 males and 8 females). This group was selected because they presumably have to have good reading comprehension and to match the distribution of males and females in both samples. *Discriminant validity* regarding *memory* was obtained by analyzing data of the participants that completed the memory test additionally to the reading comprehension screening and after excluding persons with higher-than-

average and lower-than-average scores in memory ($n = 124$, 111 males and 15 females; age 18-66, $M = 33.91$, $SD = 10.70$; $n = 91$ from the traffic psychology sample and $n = 41$ from the university sample remained in the sample) to avoid distortions because of outlying scores.

Instruments

The *reading comprehension* test for adults consists of one story (695 words) and 16 questions which have to be answered by selecting correct answers out of a set of nine answer possibilities. Participants are not allowed to scroll back to the text, but have to answer the questions based upon their understanding of the text. There is one total score for the measure, which is the sum of all correct answers (i.e., correctly selected answer possibilities and correctly not selected answer possibilities). Testing time (including instruction phase) is about 10 to 15 minutes depending on the individual working speed.

The subtest “*memorizing goods*” of the computerized *Intelligenz-Struktur-Analyse* (Intelligence-Structure-Analysis; ITB, & Gittler, 1998) was used for testing memory. The task of the test taker is to memorize goods, prices and places that have to be recognized after a distraction phase. All answers are given in a multiple-choice answer format. The ISA has already been used in several empirical studies (e. g., Neubauer, & Fink, 2003; Neubauer, Fink, & Schrausser, 2002; Proyer, 2006) and has shown its usefulness and validity for the assessment of different facets of intelligence there.

Sociodemographic data and information on self reported reading behavior were collected in a standardized sheet at the beginning of the testing. Information on reading behavior (“Do you read in your spare time?”, What do you read? – newspaper, fiction, non-fiction/technical books; “How many books do you read per year? 0-5, 6-10, 10-20, >20) was combined in a four-point scale representing different levels of reading proficiency; i.e., 1 = *non readers*: They do not read in their spare time, and do not indicate what they read; 2 = *less proficient readers*: They read newspaper in their spare time but do not indicate to read books; 3 = *proficient readers*: They read different types of printed material, in their spare time, and do read up to 10 book/year; and 4 = *highly proficient readers*: They indicate to read different types of printed material in their spare time and to read more than 10 books per year.

Procedure

All participants filled in a standardized sheet on demographics first and read the story and its accompanying multiple-choice questions afterwards. Finally, they completed the memorizing goods-subtest. Data were collected at the University of Vienna (Institute of Psychology and Institute of Machine Elements), from a selection process for a University of Applied Sciences conducted through the testing service at the Division for Assessment and Applied Psychometrics, and at the AAP. Participants were not paid for their services but received individual feedback directly after the testing session.

Analysis

Psychometric properties were tested based on the *One-Parameter Logistic Model* (1-PL, Rasch-model) using the *Andersen-Conditional Likelihood-Ratio test* (ACLR; Andersen, 1973). Subsequently *z*-values (Fischer & Scheiblechner, 1970) were used to identify non-fitting items. We split the data at the median for the internal criterion “high vs. low” raw score and also considered “sex”, “age”, “educational level”, and “working time” (the time needed to answer all items) as external criteria. Item parameter estimates and test-statistics were computed with *LpcmWin 1.0* (Fischer & Ponocny-Seliger, 1998), using *Conditional-Maximum-Likelihood* (CML) estimators. For the cross-validation, the calibration sample’s item parameters will be compared with the independently collected cross-validation sample’s item parameters, also using the Andersen-Conditional Likelihood-Ratio test and *z*-values.

Construct validity was tested by calculating correlation coefficients between the new reading comprehension measure and the memory measure. The performance of the extreme groups (traffic psychology sample vs. university sample; non-readers vs. reader) was compared using an analysis of variance.

Results

Psychometric Properties: Testing the Rasch model. The Andersen-Conditional Likelihood-Ratio test for the criterion high vs. low tendency score was significant at the 1% level indicating that the overall fit between the within score group estimates and the overall estimates of item difficulties was not given. Hence, there were some items in the initial set, which functioned differently for the low and the high scoring group.

The *z*-value statistic for the individual fit of each item indicated that item 2 and 10 did not fit at the 1% level. The stepwise exclusion of the items 2, 6 and 10 showed that for the rest of the (13) items the overall fit regarding all criteria was given (Andersen-Conditional Likelihood-Ratio test: $\chi^2 = 14.99$; $df = 12$; $\chi^2 (\alpha = 1\%) = 26.25$; *n.s.* for the criterion raw score). Results for the other criteria are shown in Table 1.

The table shows that the stepwise exclusion of three items led to a satisfying model fit for all criteria. The exclusion was not only based on the reported coefficients, but was also supported content-wise. The re-inspection of the items showed that in some cases it was not clear in which part of the story the answer was “hidden.” In other cases the answer possibilities were too similar, and in one case the question asked for a listing of answers, which may be too closely related to memory rather than reading comprehension only.

Psychometric properties: Cross-validation. The item parameters found in the calibration sample were compared with those computed from the replication sample (see Table 2).

The table shows that there were no differential functioning of items between the two data sets regarding the Andersen-Conditional Likelihood-Ratio test or the *z*-statistics. Thus, sufficient consistency of the item parameter estimations can be assumed.

Table 1:
Item fit after exclusion of three items for internal and external criteria

Criteria	χ^2 at $df = 12$	χ_0^2 at $\alpha = 1\%$
internal	14.99	26.25
age	13.29	26.25
educational level	16.62	26.25
sex	14.09	26.25
working time	17.60	26.25

Note. The internal criterion was the raw score (high vs. low, split by median).

Table 2:
Testing the Rasch Model: Cross validation of item parameter

ID	calibration sample $n = 266$		cross validation $n = 111$		
	ip	se	ip	se	z-value
1	-0.89	0.14	-0.98	0.22	0.373
2	--	--	--	--	
3	1.88	0.14	1.82	0.22	0.221
4	-0.95	0.13	-1.19	0.13	1.347
5	-0.31	0.13	-0.57	0.18	1.174
6	--	--	--	--	
7	-1.81	0.17	-1.93	0.22	0.451
8	-1.02	0.09	-0.84	0.13	-1.166
9	2.30	0.13	2.42	0.30	-0.369
10	--	--	--	--	
11	2.70	0.19	2.22	0.27	1.441
12	-1.53	0.15	-1.37	0.23	-0.595
13	-1.07	0.08	-1.03	0.14	-0.197
14	-1.11	0.09	-0.89	0.14	-1.353
15	-0.56	0.08	-0.44	0.20	-0.551
16	2.36	0.13	2.78	0.35	-1.112

Andersen-Conditional Likelihood-Ratio test $\chi^2 = 6.67$, $df = 12$

$\chi^2(5\%) = 21.01$; $\chi^2(1\%) = 26.30$ (Wilson-Hilferty approximation)

Notes. ID = item id; ip = item parameter; se = standard error of measurement.

Construct validity. We tested the overlap between reading comprehension and memory by correlating the reading comprehension measure with the ISA subtest “memorizing goods” (using the data from the calibration sample). The correlation coefficient was $r = .37$ ($p < .05$) and the squared correlation coefficient indicated that there was about 14% shared variance between reading comprehension as measured with the reading comprehension measure and memory. The overall working time in the reading comprehension measure was negatively correlated with reading comprehension ($r = -.14$, $p < .05$) but practically negligible; i.e., about 2% overlapping variance.

For the comparison of the two extreme groups (*highly* skilled readers [University sample] vs. *low* skilled readers [traffic psychology sample]) an analysis of covariance was computed between the two groups using age and memory as covariates for controlling for their potential impact. The two groups differed in their reading comprehension, $F = 5.786$, $df = 3$, $p = .001$. The means indicated that those in the University sample ($M = -0.16$) outperformed participants from the traffic psychology sample ($M = -0.99$; $d = 0.38$). Further, *memory* was found to have a significant impact ($F = 6.030$, $df = 1$, $p = .016$) while age did not. Hence, the reading comprehension measure allows for the differentiation of groups of people with presumably higher reading comprehension from those with lower reading comprehension.

The distribution of participants from both samples (university sample = presumably high skilled and traffic psychology sample = presumably low skilled) based on the measure’s preliminary norm tables (derived from the data collected for this study) are given in Figure 1. Based on the findings from the analysis of covariance, we excluded participants with higher-than-average and lower-than-average scores in memory from this comparison to control for memory effects.

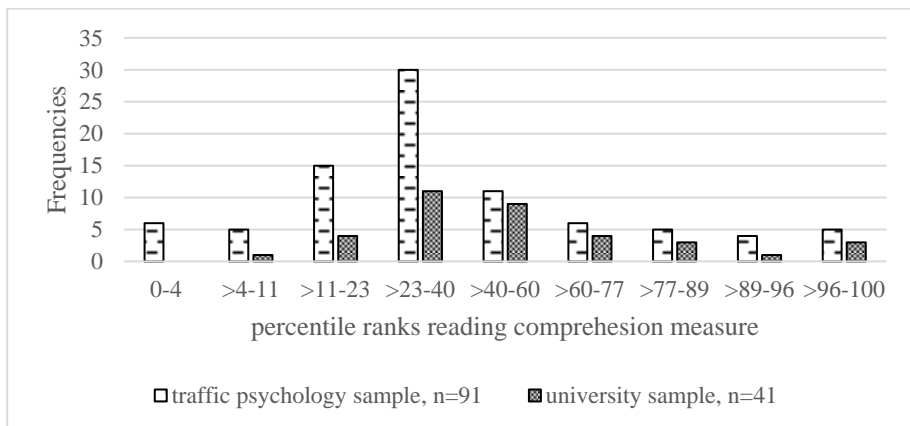


Figure 1:

Distribution of the traffic psychology sample and the university sample by means of the reading comprehension measure’s preliminary norm tables; controlled for memory

In the university sample, all participants – except one outlier – had results above a percent range of 13.8. Out of the traffic psychology group, 80 participants had a score above a percent range of 13.8. Eleven participants of this sample had a lower score. For screening purposes, it can be assumed that graduate students had a sufficient reading comprehension. For practical purposes, a score *higher* than a percent range of 10 as a stricter cut-off score indicates sufficient understanding of a text so that potentially inconsistent findings from an assessment procedure (e.g., when completing a questionnaire) could not be traced back to problems in reading comprehension.

Discussion

Following a pragmatic definition, we propose a new computerized German-language screening instrument for reading comprehension. The test taker reads a story and subsequently answers multiple-choice questions constructed to assess the understanding of the text read. This study provides initial information on the instruments' psychometric properties and its construct validity.

We found satisfactory fit for the 1-PL model (Rasch model) through the stepwise exclusion of three (out of sixteen) items. The exclusion was based on statistical, but also theoretical criteria. The model tests (for raw score; sex, age, educational level, and working time) in two independently collected samples supported the psychometric suitability of the new reading comprehension measure.

As expected, the reading comprehension measure correlated positively with a memory test; the two instruments demonstrated about 14% shared variance. Memory was seen as one of the components which were basic in the reading comprehension process. The result indicates that the test construction has worked satisfying in this relation since this reflects the impact that memory has on reading comprehension. The coefficient was in an expected range showing that reading comprehension and memory are correlated but not redundant, thus contributing to construct validity.

Statistically significant, but practically negligible correlations were found between the reading comprehension measure and educational level and working time. From a theoretical point of view a certain impact of the educational level was expected. The latter may also have an impact on a person's vocabulary. However, when constructing the measure we took care to avoid unusual words and also the psychometric analyses conducted here using educational level as one of the external criteria suggest that the impact of the educational level on the test results is only minor in nature. Our findings support the notion that the measure can differentiate between two extreme groups (putatively low vs. high skilled readers).

The screening instrument can be used for the assessment of reading comprehension whenever the examiner may have doubts or wants to make sure that the participant has the basic ability to understand the instruction and all item contents. Of course, the question as to whether specific training programs that evaluate and strengthen specific reading strategies can have a positive impact on reading comprehension will have to be addressed (cf. Magliano, Millis, RSAT Development Team, Levinstein, & Boonthum,

2011). The application of this screening instrument may lead to the decision to use non-verbal instruments or to use instruments that are comparatively easier to understand. A further field of practical use may be research in computer-based online assessment and learning (see e.g., Sonnleitner, Brunner, Greiff, Funke, Keller, Martin, Hazotte, Mayer, & Latour, 2012). A screening for reading comprehension may be helpful in such settings to determine whether prospective learners/test takers are eligible for this type of environment.

Future research will focus on the development of a parallel version. A second story and its items have already been constructed and will be tested in the same way as shown here. It needs to be mentioned that the cut-off score presented here is of a preliminary nature and should mainly be used for screening problems in reading comprehension.

Limitations. Further information is needed for supporting the construct validity of the new measure. Especially, data on the convergent validity (e.g. correlations with subtest of instruments focusing on assessing the single components of reading comprehension like the instrument of Richter & van Holt, 2005) will contribute to a better understanding of the new measure's validity. Also data from special groups such as people with reading disabilities will be necessary for further testing the instruments' potential in identifying participants with low and very low ability in reading comprehension. For future studies it would also be fruitful to have a parallel version of the test. We have addressed people whose mother tongue is German for this study, but cannot exclude that people with German as a secondary language are also in the sample. Additionally, due to certain restrictions, we do not have full information on the migration background for all participants. Although our data were highly stable across independently collected samples, it cannot be ruled out that a high number of participants that do not have German as a mother tongue might have had an impact of the findings. Of course, it needs to be mentioned that measurement instruments differ in their levels of difficulty and the way instructions are written. Therefore, the effects of low reading comprehension need to be discussed separately for each instrument. Additionally, with respect to criterion validity it would be interesting to test how the participants deal with the written material in each instrument (see e.g., Sonnleitner, 2008) and whether the reading comprehension test also can assess such differences (see e.g., Baghaei, & Carstensen, 2013). Finally, the question arises on whether other answer formats (e.g., open ended questions) would have advantages over the one used here (see e.g., Rauch, & Hartig, 2010).

Conclusion. The application of a new reading comprehension measure is suggested in cases where (for different reasons) doubts on the reading comprehension abilities of the test-taker arise. A cut-off value has been determined and if the score of the subject falls below this value the use of test material with a strong focus on verbal test material (e. g., personality questionnaires) cannot be advised. Otherwise, it is unclear whether the findings reflect the expression of the person with respect to the target characteristic or whether the result has been distorted by low reading skills.

Acknowledgments

Franz Felix Schedl wrote the story used as stimulus material. The inventory was developed in cooperation with the Dr. G. Schuhfried Company (Austria), and has been published there under the name of LEVE. The authors wish to thank the “Angewandte Psychologie und Forschung GmbH” (AAP; Austrian Applied Psychology Company) for their support in collecting the data.

References

- Anderson, E. (1973). A goodness of fit test for the Rasch-model. *Psychometrika*, *38*, 123-140. doi:10.1007/BF02291180
- Baghaei, P., & Carstensen, C. H. (2013). Fitting the Mixed Rasch Model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, *18*(5). Available online: <http://pareonline.net/getvn.asp?v=18&n=5>
- Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling*, *56*, 60-82.
- Doak, C., Doak, L., & Root, J. (1985). *Teaching patients with low literacy skills*. Philadelphia, PA: Lippincott.
- Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*, 59-77. doi:10.1007/BF02293919
- Fischer, G. H., & Scheiblechner, H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch [Algorithms and programs for the probabilistic test model of Rasch]. *Psychologische Beiträge*, *12*, 23-51.
- Fischer, G.,H. & Ponocny-Seliger, E. (1998). *Structural Rasch modeling: Handbook of the usage of LPCM-WIN 1.0*. Groningen, NL: ProGAMMA.
- Frith, U. (1985). Beneath the surface of surface dyslexia. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). London, UK: Routledge and Kegan Paul.
- Grabner, R. H., Fink, A., Stipacek, C., & Neubauer, A. C. (2004). Intelligence and working memory systems: evidence of neural efficiency in alpha band ERD. *Cognitive Brain Research*, *20*, 212-225. doi:10.1016/j.cogbrainres.2004.02.010
- Grissemann, H. (1986). *Pädagogische Psychologie des Lesens und Schreibens* [Paedagogic psychology of reading, and writing]. Bern, Switzerland: Huber.
- Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory and Language*, *48*, 320-345. doi:10.1016/S0749-596X(02)00506-5
- Hannon, B. (2012). Understanding the relative contributions of lower-level word processes, higher-level processes, and working memory to reading comprehension performance in proficient adult readers. *Reading Research Quarterly*, *47*, 125-152. doi:10.1002/RRQ.013

- ITB & Gittler, G. (1998). *Intelligenz-Struktur-Analyse (ISA)* [Intelligence-Structure-Analysis]. Mödling, Austria: Schuhfried.
- Jastak, S. & Wilkinson, G. S. (1984). *The Wide Range Achievement Test - Revised*. Wilmington, DE: Jastak Associates.
- Jones, F. W., Long, K., Finlay, W. M. L. (2006). Assessing the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disability Research*, *50*, 410-418. doi:10.1111/j.1365-2788.2006.00787.x
- Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228-238. doi:10.1037//0096-3445.111.2.228
- Katz, L., Branacazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2011). What lexical decision and naming tell us about reading. *Reading and Writing*, *25*, 1259-1282. doi:10.1007/s11145-011-9316-9
- Langenmayr, A. (1997). *Sprachpsychologie* [Language Psychology]. Göttingen, Germany: Hogrefe.
- Magliano, J. P., Millis, K. K., RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacognition and Learning*, *6*, 131-154. doi:10.1007/s11409-010-9064-2
- Mellard, D. F., Anthony, J. L., & Woods, K. L. (2012). Understanding oral reading fluency among adults with low literacy: dominance analysis of contributing component skills. *Reading and Writing*, *25*, 1345-1364. doi:10.1007/s11145-011-9322-y
- Langer, I., Schulz von Thun, F., & Tausch, R. (1981). *Sich verständlich ausdrücken* [Articulating Oneself Clearly] (2nd ed.). München, Germany: Reinhard.
- Leslie, L., & Cadwell, J. (2001). *Qualitative reading inventory*. New York, NY: Addison Wesley.
- Messick, S. (1989). Validity. In R. L. Lin (Ed.), *Educational measurement* (3rd ed., p.13-103). New York, NY: Macmillan.
- Neubauer, A. C., & Fink, A. (2003). Fluid intelligence and neural efficiency: Effects of task complexity and sex. *Personality and Individual Differences*, *35*, 811-827. doi:10.1016/S0191-8869(02)00285-4
- Neubauer, A. C., Fink, A., & Schrausser, D. G. (2002). Intelligence and neural efficiency: The influence of task content and sex on the brain-IQ relationship. *Intelligence*, *30*, 515-536. doi:10.1016/j.cogbrainres.2005.05.011
- Norman, S., Kemper, S., & Kynette, D. (1992). Adults' reading comprehension: Effects of syntactic complexity and working memory. *Journal of Gerontology*, *47*, 258-265. doi:10.1093/geronj/47.4.P258
- Oullette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*, 554-566. doi:10.1037/0022-0663.98.3.554
- Proyer, R. T. (2006). The relationship between vocational interests and intelligence: Do findings generalize across different assessment methods? *Psychology Science*, *48*, 463-476.

- Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52, 354-379.
- Richter, T., & van Holt, N. (2005). ELVES: Ein computergestütztes Diagnostikum zur Erfassung der Effizienz von Teilprozessen des Leseverstehens [ELVES: Computer-based assessment of the efficiency of component processes involved in reading comprehension]. *Diagnostica*, 51, 169-182. doi:10.1026/0012-1924.51.4.169
- Rickheit, G., & Strohner, H. (1993). *Grundlagen der kognitiven Sprachverarbeitung* [Basic principles of cognitive language processing]. Tübingen, Germany: Franke.
- Schwantner, U., Toferer, B., & Schreiner, C. (Eds.). (2013). *PISA 2012. Internationaler Vergleich von Schülerleistungen. Erste Ergebnisse. Mathematik. Lesen. Naturwissenschaften.* (PISA 2012. International Comparison of Students' Performance. First Results. Math. Reading. Science.) online https://www.bifie.at/system/files/buch/pdf/pisa12_erste_ergebnisse_2013-12-03.pdf. Graz, Austria: Leykam.
- Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, 50, 345-362.
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling*, 54, 54-72.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.
- Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1998). The case for early intervention in diagnosing specific reading disability. *Journal of School Psychology*, 36, 367-397. doi:10.1016/S0022-4405(98)00019-3
- Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22, 407-432. doi:10.1002/acp.1414
- Waters, G. S. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology: Section A*, 49, 51-79. doi:10.1080/713755607
- Woodcock, R.W. (1998). *Woodcock Reading Mastery Tests-Revised*. Circle Pines, MN: American Guidance Service.