Psychology Science Quarterly, Volume 50, 2008 (3), pp. 345-362

Using the LLTM to evaluate an item-generating system for reading comprehension

PHILIPP SONNLEITNER[1]

Abstract

Due to inconclusive findings concerning the components responsible for the difficulty of reading comprehension items, this paper attempts to set up an item-generating system using hypothesis-driven modeling of item complexity applying Fischer's (1973) linear logistic test model (LLTM) to a German reading comprehension test. This approach guarantees an evaluation of the postulated item-generating system; moreover construct validity of the administered test is investigated. Previous findings in this field are considered; additionally, some text features are introduced to this debate and their impact on item difficulty is discussed. Results once more show a strong influence of formal components (e.g. the number of presented response options in a multiple-choice-format), but also indicate how this effect can be minimized.

Key words: Rasch model, LLTM, item-generating rules, reading comprehension

[1] Philipp Sonnleitner, MSc., Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, University of Vienna, Faculty of Psychology, Liebiggasse 5, A-1010 Vienna, Austria, Europe; email: philipp.sonnleitner@univie.ac.at

Introduction

In recent years the need to test reading comprehension has risen. Many large-scale-assessments like PISA[2], PIRLS[3], IALS[4], ALL[5], or the Austrian Educational Standards test (Kubinger et al., 2006) aim to measure how well people comprehend texts. Nevertheless, theory-based item construction or the use of systematical item-construction rules for this purpose is rare. The statement of Embretson and Gorin (2001, p. 343) is still valid: "… item design is viewed primarily as an art. Item specifications are often vague." In spite of many studies addressing this issue and calling for "blueprints" for tests of reading comprehension and other verbal tasks (Jafarpur, 2003; Eckes, 2004a, 2004b, 2005), the suggested use of Item Response Theory (IRT) in order to handle this problem in a technically optimal way is limited. It has mainly been used to analyze items' quality only after test application, a strategy which Hornke and Habon (1984), for instance, call "a mere justification strategy."

To base item development on cognitive theory would not only guarantee construct validity (cf. Hornke & Habon, 1984, 1986; Hornke & Rettig, 1989); a completely operationalized and evaluated system of construction rules would also have additional benefits, most importantly the ability to predict item parameters for analogous but newly developed items (Embretson & Gorin, 2001). This automatized construction of items with a-priori known item difficulties would offer new possibilities for adaptive testing (Embretson, 1999; Fischer & Pendl, 1980).

Identifying the components of an item-generating system

On the basis of experimental findings to the preferred trait-representing cognitive model, it is necessary to define possible components of an item-generating system. According to recent studies (Bejar, 1993; Irvine, 2002; Gorin, 2005), the characteristics of an item can be classified as "radicals" or "incidentals". Whereas incidentals are components which should not affect item difficulty, radicals are substantial elements of an item responsible for its difficulty. Identifying the radicals of a test would therefore help clarify its construct validity, disclosing what is really measured by the test.

Most of the studies that investigate components of reading comprehension items and their effects on item difficulty only use a correlation approach. But as Gorin (2005) argues, a high correlation between item components and item difficulty is no proof of causality. Furthermore, correlation methods are strongly affected by the variance within the sample. For example, a pool of very difficult items would lead to low correlations with difficulty-causing components, whereas a pool with a wide range of difficulties would result in high correlations. Considering a pool of items with nearly equal difficulties, the components in question simply would not, using correlation, explain the small differences in item difficulties. In contrast, the IRT framework enables researchers to model cognitive complexity and estimate item difficulties in a non-correlative manner.

---

[2] Programme for International Student Assessment

[3] Progress in International Reading Literacy Study

[4] International Adult Literacy Study

[5] Adult Literacy and Life skills Survey

The linear logistic test model (LLTM)

As a special case of the Rasch model (or 1-PL model), the LLTM (Fischer, 1973) postulates that the item parameter σ_i can be described as the sum of weighted elementary operations:

$$\sigma_i = \sum_j^p q_{ij}\eta_j$$

σ_i … item difficulty parameter of item *i*
η_j … estimated difficulties of the cognitive operations
q_{ij} … weight of the cognitive operation η_j in item *i*

This structure can be used to model the cognitive processes involved in solving/ processing any item. After defining and modeling these elementary operations with appropriate weights, the difficulty of every single cognitive operation involved is estimated by means of conditional maximum likelihood method.

Under the assumption that the Rasch model fits the data, the fit of the specified cognitive model can be tested in two ways: In a way similar to testing the fit of the Rasch model, a graphical model check can be used to compare the estimated item parameters of the Rasch model to the parameters reproduced by the LLTM parameter estimates. Items reducing the respective correlation deliver valuable information and can be used to modify or reformulate the structure of elementary operations, or to help to identify problems in item construction. Secondly, model fit is tested using a Likelihood-Ratio test (cf. Fischer, 1973; see also Kubinger, 2008).

Previous rule-based item construction for measuring reading comprehension

Mispelkamp (1985) first used the LLTM to test an item-generating system for reading comprehension. He formulated 4 radicals for constructing and describing the cognitive complexity of 9 items on basis of Kintsch's (1974, 1998) Construction-Integration (CI) model, which is one of the best-known cognitive models for reading comprehension (Rupp, Ferne, & Choi, 2006; Richter & van Holt, 2005). The main idea of the CI model is that every text is structured through and consists of a network of so-called propositions, which can be seen as "idea units" (Kintsch, 1998). During reading, these propositions are processed and finally integrated in a so-called mental model of the text, which is constructed by the reader. In Mispelkamp's opinion, the final results showed a satisfying explanation of the item difficulties which supports previous findings of the CI-model: When an inclusion of several propositions was needed for solving the item, the difficulty actually increased considerably. On the other hand, the length of the related chunk (indicating the demand on working memory) and the reference to a main topic of the text had no or only slight influence on item difficulty. As expected, the more often a proposition was processed (and therefore the better it was integrated in the mental model of the text), the easier the related item was. Interestingly, these findings were consistent regardless of differing response formats (multiple-choice vs. open-response format).

Embretson and Wetzel (1987) also used the LLTM for investigating two categories of processes involved in solving a reading comprehension item and influencing its difficulty: 1) the construction of an appropriate mental representation of the text, which consists of encoding and comprehending the words and building a network between them, and 2) the selection of the right answer achieved by encoding and connecting the options, finding the necessary information in the mental model, and verifying (or rather falsifying) the answer options. It turned out that in addition to characteristics of the text like density of adjectives or adverbs (the number of adjectives or adverbs in relation to other words used in the text), the arrangement of the response-options had a high impact on item difficulty.

Recently, Gorin (2005) investigated the feasibility of algorithmic variation of the difficulty of reading comprehension items by experimentally varying four radicals. Surprisingly, results showed that there were no effects of propositional density, information order, or response-alternative changes on the estimated item difficulty parameters. However, these findings contradict the previous findings mentioned above as well as most of the cognitive literature concerning reading comprehension.

Due to these inconclusive findings, the given paper defines another item-generating system in order to shed light on the debate by additionally investigating a central but hitherto neglected component of reading comprehension – the role of inferences in processing reading comprehension items.

Defining the item-generating system

The above-mentioned classification of radicals into two groups (constructing a mental representation of the text and processing the item) by Embretson and Wetzel (1987) is supported by Bachman and Palmer (1996), who also list the type of input and the type of response-format as methodical influences on the measurement of reading comprehension. For this reason, the item-generating system developed in the following study adheres to this categorization system. It differentiates between radicals based on input (more precisely the text needed for solving the item) and radicals based on response format.

Input-related radicals

- *Propositional Complexity*: The strong influence of propositional density in a text or the number of processed propositions on solving a reading comprehension item is beyond controversy. In contrast to Gorin (2005), there are many studies (Kintsch & Keenan, 1973; Mispelkamp, 1985; Barshy & Healy, 2002) showing clearly that item difficulty increases with increasing propositional density. Consequently, a parameter describing whether the relevant text was limited to a small passage or extended across the whole text was hypothesized in this study.
- *Type of Inference*: According to Kintsch (1998), the interdependency of text, reader and the reading context is most apparent when it is necessary to read "between the lines" and elaborate on implicit aspects of the text by inferring from known information. Even though McKoon and Ratcliff (1992) postulate a "minimalist hypothesis" stating that inferences are only made when comprehension of the text is interrupted, Zwaan, Graesser

and Magliano (1995) show that readers draw causal or temporal inferences all the time. Perfetti, Landi and Oakhill (2005) proved the influence of cognitive demand on the probability of making an inference. According to their results, an inference is more likely to be made if its cognitive demand is low. Therefore, the consideration of inferences in an item-construction system is not only justified by their central role in text comprehension, there is also strong evidence that the type of inference involved in solving an item has a considerable influence on cognitive demand. For this reason, two alternative types of classification are investigated in terms of their ability to explain different item difficulties: First, Kinsch's classification (see, for example, Kintsch & Rawson, 2005) divides inferences into four groups, depending on whether they are made automatically or intentionally and whether they are based on information within the text or on knowledge. The second investigated classification system is based on results of Graesser, Singer and Trabasso (1994), who identified several different types of inferences occurring in narrative texts. Five of these inferences are made automatically during reading; two more are made automatically if the reader is primed with a topic or task related to the text. According to Perfetti et al. (2005), it is hypothesized that items based on automatically drawn inferences produce less cognitive demand and are more likely to be solved.

- *Coherence of the text:* To guarantee a text's coherence, words serving as referents connecting new information with information that has already been mentioned are needed. Freedle and Kostin (1999) showed a significant influence of the number of referents on item difficulty. This supports the findings from Just and Carpenter (1980) that the frequency with which a text refers to a proposition dramatically affects how well that proposition is integrated into the mental model of the text. Therefore, it is hypothesized that an item relating to a highly coherent text should be easier to solve.
- *Text*: The items investigated in this study are related to two different texts. As each text requires the construction of its own situation model (Kintsch, 1998) and therefore depends on content-specific knowledge or a certain motivation to deal with the topic of the text, one parameter is introduced to model the text-specific characteristic.

Response-related radicals

Solving an item also consists of processing the item itself and verifying each response option; this process should be considered in any item-construction system. The response-related radicals postulated for the item-generating system are as follows:

- *Number of response options:* According to Embretson and Wetzel (1987) as well as Gorin and Embretson (2006), every response option is processed and finally verified or falsified. This means that every single response option can be seen as a cognitive operation. A parameter representing the number of response options per item should model the underlying cognitive demand in an adequate way.
- *Number of correct response options:* Gittler (1984), investigating spatial tasks, reported a higher difficulty for verifying a response option than for falsifying and rejecting it. Additionally, the current study hypothesizes that the perceived probability that there are still correct response options left decreases with an increasing number of already chosen options (provided that the number of correct choices is unknown). According to this hypothesis, a large number of correct response options probably increases item difficulty.

Although the specified response-related radicals should adequately model the solving process, additional effects should also be considered, given that there is strong evidence of interdependency between response format and reading comprehension items (for example, Rupp et al., 2006; Kobayashi, 2002). Analyzing the administered items in the face of these results, the definition of two more parameters seems unavoidable:

- *Temporal dependency of response options:* Analyses of the response options of the administered test showed that items asking about the actions of a text's protagonist are more difficult to solve if they ask for more than one activity. It seems likely that only one activity within the mental model is remembered during item processing and used to answer the question.
- *Ambiguity:* Language is often ambiguous and, depending on a reader's former experience or the textual context, words can be interpreted in different ways (Just & Carpenter, 1980). For this reason, the interpretations of the test-developer need not necessarily be the same as that of the test-taker. Therefore, items containing ambiguous words are hypothesized to have a higher difficulty.

In the present paper, the stated item-generating system is tested by applying the LLTM to an already existing German reading comprehension test. This approach offers several advantages: On the one hand, the radicals postulated above can be evaluated using a reading comprehension test that fits the Rasch model and therefore unambiguously measures the intended trait. Hence, concurrent hypotheses about the stated radicals can easily be tested by adjusting the structure matrix of the LLTM. On the other hand, the evaluation also delivers useful information about construct validity of the administered test by clarifying the underlying cognitive processes.

Method

The test used for the analysis, the LEVE-E (Leseverständnistest für Erwachsene [Reading Comprehension Test for Adults]) by Proyer, Wagner-Menghin and Grafinger (2006) is a computer-based multiple-choice test consisting of two texts of equal length on two different topics. The first text describes a chaotic scene at a Greek crossroad caused by a newly installed traffic light. In the second text, a man hesitates to walk into a bank to trade a golden ducat. According to Grafinger (2002), the selection of the texts was made according to the following criteria: type of text (it should be a narrative), text length, organization of the text, acceptance of the text by the target group (adults), and finally the content of the text (it should allow a broad variety of questions). The items were constructed to measure word comprehension, vocabulary, sentence comprehension, text comprehension, and retentiveness.

The data used for the analysis were collected in four different settings during the development and calibration of LEVE-E[6] (see Grafinger, 2002, for further details). The sample consists of 301 examinees (university students of different academic disciplines and clients taking a psychological driver's examination), each of whom answered the items of at least one text.

[6] At this point I would like to thank René Proyer, Ph.D. and Gyüngü Grafinger, MSc. for providing the data set used for the calculations

As mentioned before, the LLTM models the cognitive complexity of each item by decomposing it into the underlying cognitive processes or elementary operations. The aforementioned radicals can be seen as elementary operations and therefore, using certain weights, every item can be modeled by them. The result is a complete structure matrix describing the cognitive complexity of each item. To illustrate this, a part of the structure matrix of Model 2 is given in Table 1.

Table 1:
Part of the structure matrix of Model 2

item	propositional complexity high=1, low=0	inference of causality needed, yes=1, no=0	inference of the emotional reaction of a character needed, yes=1, no=0	inference of a subordinate noun category needed, yes=1, no=0	inference of a used instrument needed, yes=1, no=0	…
1	1	0	0	0	0	…
2	0	1	0	0	0	…
3	1	1	0	1	0	…
4	1	1	1	0	0	…
5	1	1	0	0	1	…
…	…	…	…	…	…	…

The cognitive complexity of the items was modeled by the author and a test-construction expert in an iterative way. With the exception of Item 21, which was difficult to categorize using the inference classification of Graesser et al. (1994), all items could unambiguously be classified according to their radicals.

To give an example, the categorization of Item 7 by using the inference categorization of Graesser et al. is described in detail. First, the relevant text passage, needed for solving Item 7 was identified to apply the input-related radicals:

"*The iceman ran out of ice cream and cones. The lack of ice cream took some children the interest out of the game."*

Because the relevant text is limited to a small passage consisting of only two sentences following each other, the propositional complexity was modeled by "0". The two sentences are not only directly one after another, there is also the referent "*the lack of ice cream*" in the second sentence which strengthens the connection by referring to the sentence before. Therefore, the high coherence of this small passage is modeled by (1+1=) "2". Finally, in connection with the corresponding question "Why did the children stop playing?" an inference of causality is needed, connecting the loss of interest with the end of the game. This is characterized in the according radical by the weight "1". No other inferences are needed to solve

the item and therefore all other input-related radicals were weighted with "0". As Item 7 belongs to text number 1, the "text parameter" is weighted with"0".

Concerning the response-related radicals, Item 7 is characterized as follows. It provides 6 response options, of which 1 response option, number 3 ("*Because there was no more ice cream.*") is counted as correct. As described above, the cognitive demand of processing the (correct) response options can be described by their number; hence, the item complexity was modeled by the weights "6" in the parameter "number of response options" and "1" in the parameter "number of correct response options". As the response options are not temporally dependent on each other and do not contain any ambiguous words – and neither does the relevant text passage – both parameters ("temporal dependency", "ambiguity") were weighted with "0".

By combining the aforementioned radicals, different models for developing the item-generating system and proving the construct validity of the analyzed test were set up. While in the first step only input-related radicals were considered, the second step extended the analysis by including response-related radicals.

Step 1: Proving construct validity on the basis of input-related radicals

Unambiguous confirmation of construct validity would be given if the item parameters estimated with the Rasch model could be explained by using only input-related radicals as a model of the cognitive complexity of the items. In the light of the findings mentioned above (for example, Rupp et al., 2006), this seems very unlikely. Comparing the two categorizations of inferences should, however, help indicate which model is more useful for setting up an item-generating system. Table 2 shows the radicals used for the two compared item-generating systems.

Table 2:
Radicals used in the compared item-generating Models 1 and 2

compared models	**Model 1: Inferences classified sensu Kintsch (7 resulting basic parameters)**	**Model 2: Inferences classified sensu Graesser et al. (1994) (8 resulting basic parameters)**
radicals for modeling cognitive complexity	propositional complexity	
	degree of coherence	
	text	
	automatic, knowledge-based inference	
	automatic, text-based inference	
	controlled, knowledge-based inference	
	controlled, text-based inference	
		inference of causality
		inference of the emotional reaction of a character
		inference of a subordinate noun category
		inference of a used instrument
		inference of general conditions

Results

The data set was analyzed using the software *LPCM-Win* 1.0 (Fischer & Ponocny-Seliger, 1998). The level of significance was set to $\alpha = .01$.

Before applying the LLTM, the assumption of a Rasch model-fitting item pool had to be tested. This was done using state-of-the-art techniques (cf. Kubinger, 2005), specifically Andersen's Likelihood-Ratio test (LRT). The criteria for partitioning the examinee sample were score (high vs. low) and sex. Table 3 shows the results: the Rasch model holds.

For each of the two models compared, a structure matrix was set up and tested by applying the LLTM. Applying the indicated LRT resulted in significance for both models (Table 4).

Table 3:
Results of Andersen's Likelihood-Ratio test

partition criteria	(*df*=*k*-1) *k*…number of estimated parameters	$\chi^2_{\alpha=0.01}$	χ^2_{emp}
score	21	38.9321	22.3720
sex	21	38.9321	20.9920

Table 4:
Results of the LRT comparing the data's likelihoods of Model 1 and 2 each to the data's likelihood given the Rasch model

model	**number of parameters**	(*df*=*k*-1-*p*) *k*…number of estimated parameters *p*..number of elementary operations	$\chi^2_{\alpha=0.01}$	χ^2_{emp}
1	7	14	29.1413	1087.7793
2	8	13	27.6883	568.2369

Thus, the 22 item parameters of the Rasch model cannot be explained by only 7 or 8 hypothesized basic parameters. Therefore, both models must be rejected. This means that neither Model 1, modeling the inferences of the items according to Kintsch, nor Model 2, using the categorization of Graesser et al., describes the underlying cognitive processes in an adequate way.

Step 2: Testing the item-generating system

Hence, the complete formulated item-generating system should be evaluated by adding the response-related radicals. Although previous results suggest that the categorization of Graesser et al. should be used for modeling the complexity of inferences, it is important to investigate whether the LLTM estimates for the two models taking response-related radicals into account are able to match the item parameter estimations of the Rasch model (i.e. do not differ significantly from these). The two models including the response-related radicals are illustrated in Table 5.

Table 5:
Radicals used in the compared Models 3 and 4

compared models	**Model 3: Inferences classified sensu Kintsch (11 resulting basic parameters)**	**Model 4: inferences classified sensu Graesser et al. (1994) (12 resulting basic parameters)**
input-related radicals for modeling cognitive complexity	propositional complexity	
	degree of coherence	
	text	
	automatic, knowledge-based inference	
	automatic, text-based inference	
	controlled, knowledge-based inference	
	controlled, text-based inference	
		inference of causality
		inference of the emotional reaction of a character
		inference of a subordinate noun category
		inference of an used instrument
		inference of general conditions
response-related radicals	number of response options	
	number of correct response options	
	temporal dependency of response options	
	ambiguity	

Results

Analogously to the previous investigation, two LRTs were carried out; their results are shown in Table 6. Again the data's likelihood based on the 22 Rasch model item parameters was compared to the respective likelihoods based on 11 and 12 basic parameters.

Table 6:
Results of the LRT comparing the data's likelihoods of Model 3 and 4 each to the data's likelihood given the Rasch model

model	**number of parameters**	(df=k-1-p) k…number of estimated parameters p..number of elementary operations	$\chi^2_{\alpha=0.01}$	χ^2_{emp}
3	11	10	23.2093	268.0427
4	12	9	21.6660	37.1296

Again, both models must be rejected.

Although item difficulties of the LEVE-E could not be explained by the postulated input-response-related radicals, it is quite evident that Model 4, based on the classification of inferences by Graesser et al. (1994), shows the comparatively better fit – though attention should be paid to the fact that Model 3 postulates fewer basic parameters and is therefore more restrictive than Model 4. The correlation between estimated Rasch model item parameters and Model 4 LLTM-based item parameter estimations amounts to .9872, which is nevertheless of practical value (cf. Tab. 7). A graphical model check of both these item parameter estimations also discloses a very high accordance (cf. Fig. 1).

As mentioned above, during parameterization of the items, a problem occurred with Item 21, which was difficult to classify according to the category system of Graesser et al. (1994). Interestingly, the graphical model check (see Fig. 1) shows that Item 21 (colored black) is further from the 45° line than any other item. Therefore, another analysis was carried out "a-posteriori," excluding Item 21 and modifying the structure matrix of Model 4 into Model 4 *post* (cf. Tab. 8).

Table 7:
Comparison of $\beta i^{(LLTM)}$ to $\beta i^{(Rasch\ model)}$ of Model 4
(inferences classified sensu Graesser et al. (1994))

item	$\beta_i^{(LLTM)}$	$\beta_i^{(Rasch\ model)}$
1	-0,8959	-1,0189
2	1,7994	1,6754
3	-0,9497	-1,0842
4	-0,4269	-0,4457
5	-2,1889	-1,9419
6	-1,2879	-1,1513
7	1,8622	2,0871
8	2,7889	2,4874
9	-1,6479	-1,6590
10	-1,0941	-1,1970
11	-1,1879	-1,2437
12	1,7994	2,1537
13	0,4523	0,3301
14	-0,8526	-0,5181
15	-1,0282	-1,3024
16	-0,9720	-0,8248
17	0,1570	-0,0278
18	1,0093	1,3279
19	1,8912	2,1454
20	-1,2832	-1,5500
21	1,8912	1,4277
22	0,1643	0,3301

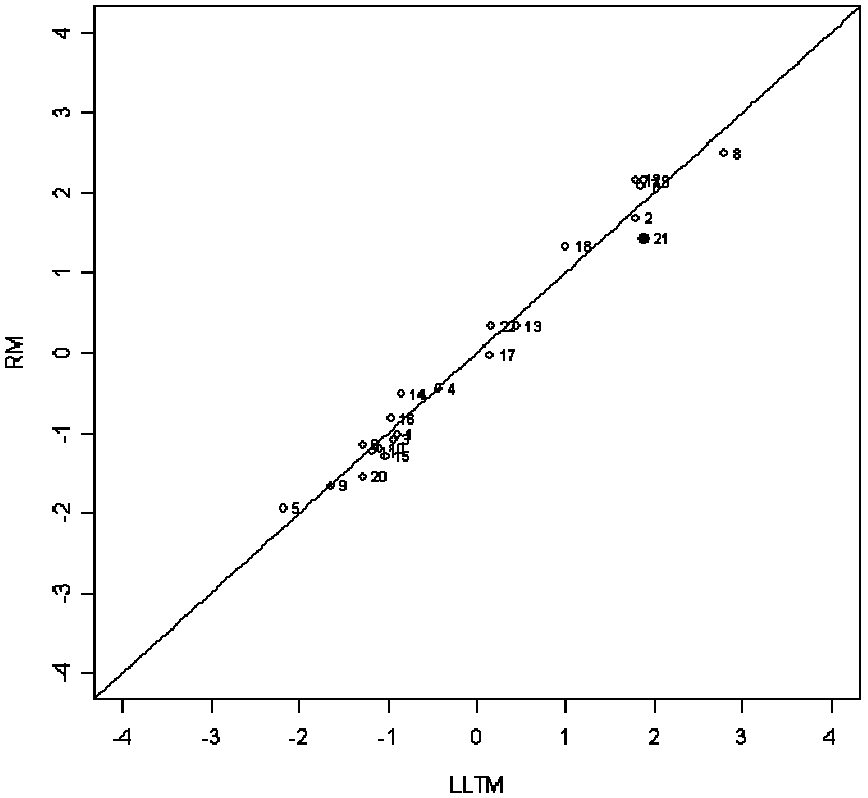

Figure 1:
Graphical model check of Model 4 (Item 21 with a very large distance to the 45° line is specially marked)

Results

The result of the LRT testing the fit of Model 4 *post* is shown in Table 9. According to the low χ^2-value, the data's likelihood given Model 4 *post* does not significantly differ from the data's likelihood under the assumption of the saturated model, which means that the postulated 12 basic parameters (i.e. radicals) of Model 4 *post* are able to explain the observed item parameters.

Table 10 shows each radical's contribution to item difficulty. With the exception of the input-related radical "degree of coherence", all other radicals verifiably influence item difficulty. Whereas negative values imply that a radical makes item processing more difficult, positive values describe an alleviative effect of the corresponding radical on item processing.

Table 8:
Structure matrix of Model 4 *post*

item	pc	ic	ier	inc	iui	igc	ch	nro	ncro	td	a	t
1	1	0	0	0	0	0	0	7	4	0	0	0
2	0	1	0	0	0	0	0	5	1	0	0	0
3	1	1	0	1	0	0	0	9	1	0	1	0
4	1	1	1	0	0	0	0	5	2	0	1	0
5	1	1	0	0	1	0	0	7	5	1	1	0
6	1	1	0	0	0	0	0	7	3	0	0	0
7	0	1	0	0	0	0	2	6	1	0	0	0
8	0	0	0	0	0	0	2	6	1	0	0	0
9	1	0	0	0	0	0	1	7	4	1	0	0
10	0	1	0	0	0	0	0	6	2	1	1	0
11	0	1	0	0	0	1	0	7	3	1	0	0
12	0	1	0	0	0	0	0	5	1	0	0	0
13	0	0	0	0	0	1	0	4	1	0	0	1
14	0	0	0	0	0	0	0	6	1	0	1	1
15	0	0	1	0	0	0	0	6	3	0	1	1
16	0	0	0	1	0	0	1	7	5	0	0	1
17	0	1	0	0	0	0	1	2	2	0	0	1
18	0	1	1	0	0	0	1	4	1	0	0	1
19	0	1	0	0	1	0	0	5	1	0	0	1
20	0	1	0	0	0	0	1	7	2	0	0	1
22	0	0	0	0	0	1	0	5	1	0	0	1

pc… propositional complexity
ic… inference of causality
ier… inference of the emotional reaction of a character
inc… inference of a subordinate noun category
iui… inference of an used instrument
igc… inference of general conditions
ch… degree of coherence
nro… number of response options
ncro… number of correct response options
td… temporal dependency
a… ambiguity
t… text

Table 9:
Results of the LRT comparing the data's likelihood given Model 4 *post* to the data's likelihood given the Rasch model

model	number of parameters	(df=k-1-p) k…number of estimated parameters p..number of elementary operations	$\chi^2_{\alpha=0.01}$	χ^2_{emp}
4 *post*	12	8	20.0902	18.2223

Table 10:
Estimated basic parameters of Model 4 *post*

radical	parameter	significance
propositional complexity	-1.4657	sign. (1%)
inference of causality	-0.9631	sign. (1%)
inference of the emotional reaction of a character	1.0399	sign. (1%)
inference of a subordinate noun category	-0.5157	sign. (1%)
inference of an used instrument	2.5668	sign. (1%)
inference of general conditions	0.9556	sign. (1%)
degree of coherence	0.1270	n.s.
number of response options	-0.2795	sign. (1%)
number of correct response options	-0.5746	sign. (1%)
temporal dependency	-0.8822	sign. (1%)
ambiguity	-1.2700	sign. (1%)
text	-2.1455	sign. (1%)

Discussion

As mentioned at the beginning of this paper, the formulation of an item-generating system based on the LLTM and consisting of radicals derived from cognitive psychology literature served two purposes: investigation of the construct validity of the LEVE-E and evaluation of the aforementioned item-generating system.

If the item parameters of the LEVE-E could be explained on the basis of only input-related radicals, construct validity would be proven. For this reason, two different item-construction systems (Models 1 and 2) were postulated to model item complexities. These systems, however, finally were rejected because of significant LRTs. Due to the fact that the item-generating rules used to develop the LEVE-E are not accessible, the question arises whether the postulated models considered the actual underlying cognitive processes. If not proving construct validity would be doomed to failure. But when response-related radicals were considered, the LRT testing the fit of model 4 *post* was not significant. For this reason, construct validity can be assumed for the LEVE-E under the restriction that the used response format has a strong influence on item difficulty. In the light of former studies in the field of reading-comprehension testing (Rupp, et al., 2006; Kobayashi, 2002; Gorin & Embretson, 2006; Embretson & Wetzel, 1987 or Freedle & Kostin, 1999), this is not surprising.

However, investigation of the fit of Model 3 and 4 confirmed the strong influence of inferences on item difficulty. Modeling inferences after Graesser et al. (1994) led to a much better model fit. The question of which categorization is better suited for describing item difficulty can therefore be clearly answered. It must, however, be said that this categorization was developed especially on narrative texts and therefore allowed a more precise and differentiated description of the LEVE-E. Moreover, the models using the categorization of Kintsch (for example Kintsch & Rawson, 2005) were more restrictive, using fewer parameters to describe item complexity. Nevertheless, the findings lead to the conclusion that

Kintsch's categorization can be understood as only theoretical and not able to explain cognitive complexity – in contrast to the findings of Perfetti et al. (2005).

As mentioned above, Item 21 was excluded from the analysis applying model 4 *post*. Needless to say, eliminating tested items is a suboptimal strategy for evaluating an item-generating system that claims to explain the item complexity of a whole domain. For this reason, it is necessary to explain why this item can be seen as special case. The text that Item 21 refers to, seems to describe an upcoming robbery by activating associated contents. Given the findings of Rumelhart (1975), it is evident that in dependence on the type and topic of the text, such story-specific content is activated in readers' minds and therefore influences the ongoing construction of a mental model of the text. As a result, the "small piece of metal" in the protagonist's hands seems to be a gun, but finally turns out to be a golden ducat. The related question "What kind of weapon was he carrying?" enforces this first impression. According to Kintsch and Rawson (2005), a controlled and text-based inference is necessary to understand the text correctly and solve this item by answering "none." But using the categorization of Graesser et al. (1994), no inference-class could be found which accurately represents this issue. To do this, the introduction of a new basic parameter would have been necessary. This was rejected due to the relation of basic parameters to items (12:22) and to the fact that only one item was affected.

After excluding Item 21, Model 4 *post*, consisting of 12 basic parameters, explained item difficulties as well as the saturated model. Therefore an interpretation of the stated cognitive operations, given in Tab. 10, is finally possible. Based on former findings, it was hypothesized that the propositional complexity (pc) involved in solving an item strongly influences its difficulty. Although this complexity was roughly modeled by only 2 different values, the expected tendency could be confirmed by the value of -1.4657. Items referring to information scattered over the whole text and therefore consisting of more propositions are more difficult than items referring to single propositions. Results concerning inferences can be summed up as follows. Whereas inferences of causality (ic) as well as inferences of a subordinate noun category (inc) are obviously hard to process and hence increase item difficulty, the other investigated inferences (ier, iui, igc) seem more likely to be made. Especially inferences on a used instrument (iui) seem to be very easy or, following the argumentation of Perfetti et al. (2005), do not have a high cognitive demand and are therefore most likely to be made. This high contribution to the explanation of item difficulties can only be explained by a high relevance of inferences in reading comprehension. Surprisingly, coherence of the text (ch) has no significant impact on item difficulty, contradicting the findings of Freedle and Kostin (1999), Just and Carpenter (1980), and Kobayashi (2002), who proved effects of text coherence on difficulty of response formats. The significant contribution of the input-related radical "text" shows that items of Text 1 are generally easier and that Text 2 is more difficult to comprehend.

As expected, the response-related radicals influence item difficulty to a great extent. Difficulty increases according to the number of response options (nro) as well as the number of correct response options (ncro). The fact that some items produce a "temporal dependency of response options" by referring to several actions of a text's character also significantly increases item difficulty. Finally, in view of Just and Carpenter's findings, it was hypothesized that items containing ambiguous words have a higher difficulty. It was found that such ambiguity (a) severely affects test-takers in answering the item correctly.

The fact that four out of eleven significant basic parameters consist of response-related radicals conclusively reinforces the suggestions of Embretson and Wetzel (1987) and Gorin and Embretson (2006) to consider such components when modeling complexity of reading comprehension items.

Being aware of the strong influence of formal components, a solution would be to control these effects by, for instance, holding the number of response options constant. Although this influence is essential, it could be neglected if all items are influenced in the same way. Differences in item difficulty could then be traced back to input related radicals. Furthermore, special attention should be paid to the formulation of response options, avoiding ambiguous words or phrases, and ensuring independency.

Conclusion

Although the successful explanation of 21 item parameters by 12 elementary operations is indeed a very good result when compared to initial attempts at modeling highly structured reasoning items, the findings can hardly be generalized if the postulated item-generating system is not cross-validated. Additional research is needed, verifying the stated item generating system on newly developed items and other samples. Overall, according to the results, the following principles can be recommended for further test developments in order to gather maximal information about the test-takers ability to read and minimize the influence of response format:

- The use of unambiguous words in item stems and related response options
- A constant number of response options
- A constant number of correct response options
- The consideration of inferences in the text in item construction
- The consideration of propositional density in item construction

To sum up, the study again demonstrates the considerable usefulness of the LLTM for clarifying cognitive demands of item processing even for very complex materials like verbal tasks. Even if the developed item-generating system is not ready for use at the moment, the underlying cognitive processes of the analyzed items of the LEVE-E were fully explained.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barshi, I., & Healy, A. (2002). The effects of mental representation on performance in a navigation task. *Memory and Cognition, 30,* 1189-1203.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Fredriksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-359). Hillsdale, NJ: Erlbaum.

Eckes, T. (2004a). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im « Test Deutsch als Fremdsprache » [Rater Agreement and Rater Severity: A Multifacett-Rasch-Analysis]. *Diagnostica*, *50*, 65-77.

Eckes, T. (2004b). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen [Facetts of Language testing: Severity and Consistence in Rating Language Performances]. In A. Wolff et al. (Eds.), *Materialien Deutsch als Fremdsprache* (pp. 451-484). Regensburg: FaDaf.

Eckes, T. (2005). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell [Evaluation of Ratings: Psychometric Quality Assurance with Multifacett-Rasch-Model]. *Zeitschrift für Psychologie, 213*, 77-96.

Embretson, S. E. (1999). Generating items during Testing: Psychometric Issues and Models. *Psychometrika, 64*, 407-433.

Embretson, S. E., & Gorin, J. (2001). Improving Construct Validity with cognitive Psychology Principles. *Journal of Educational Measurement, 38*, 343-368.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11,* 175-193.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. H., & Pendl, P. (1980). Individualized Testing on the Basis of Dichotomous Rasch Model. In L. J. Th. Van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds.), *Psychometrics for Educational Debates.* (pp. 171-187). Chichester, England: John Wiley Et Sons.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch Modeling. Handbook of Usage of LPCM – WIN 1.0.* Groningen: ProGAMMA.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*, 2-32.

Gittler, G. (1984). Entwicklung und Erprobung eines neuen Testinstruments zur Messung des räumlichen Vorstellungsvermögens. [Development and testing of a new instrument for measuring spatial imagination.] *Zeitschrift für Differentielle und Diagnostische Psychologie*, *5*, 141-165.

Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions: The Feasibility of Verbal Item Generation. *Journal of Educational Measurement, 42,* 351-373.

Gorin, J. S., & Embretson S. E. (2006). Item Difficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement, 30,* 394-411.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing Inferences during Narrative Text Comprehension. *Psychological Review, 101,* 371-395.

Grafinger, G. (2002). LEVE-E, Leseverständnistest für Erwachsene [Reading Comprehension Test for Adults]. *Unpubl. master thesis.* University of Vienna.

Hornke, L. F., & Habon, M. W. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben [Experiences in Rule-based Construction of Testitems]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 5,* 203-212.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369-380.

Hornke, L. F., & Rettig, K. (1989). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen [Rule-based item generation considering cognitive psychology]. In K.D. Kubinger (Ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen* [Modern psychometrics – A brief survey with recent contributions] (pp. 140-162). Munich: PVU.

Irvine, S. H. (2002). Item generation for test development: An introduction. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. xv-xxv). Mahwah, NJ: Erlbaum.

Jafarpur, A. (2003). Is the test constructor a facet? *Language Testing, 20,* 57-87.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale: New York.

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

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5,* 257-274.

Kintsch, W., & Rawson, K. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 209-226). Oxford: Blackwell.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*, 193-220.

Kubinger, K. D. (2005). Psychological Test Calibration using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*, 377-394.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From composing tests by item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,*311-327.

Kubinger, K. D., Frebort, M., Holocher-Ertl, S., Khorramdel, L., Sonnleitner, P., Weitensfelder, et al. (2006). Large-Scale-Assessment zu den Bildungsstandards in Österreich: Testkonzept, Testdurchführung und Ergebnisverwertung [Large-Scale-Assessment for educational standards in Austria: testconcept, administration and application of results]. *Erziehung und Unterricht, 7,* 588-599.

McKoon, G., & Ratcliff, R. (1992). Inference During Reading. *Psychological Review, 99,* 440-466.

Mispelkamp, H. (1985). Theoriegeleitete Sprachtestkonstruktion [Theory-based construction of a reading comprehension test]. *Unpubl. doctoral thesis,* University of Düsseldorf.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of Reading Comprehension Skill. In M. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 227-247). Oxford: Blackwell.

Proyer, R., Wagner-Menghin, M. M., & Grafinger, G. (2006). *Leseverständnis-Test* (LEVE) [Reading comprehension test for adults]. Test: Software and Manual. Mödling: Dr. G. Schuhfried GmbH.

Richter, T., & van Holt, N. (2005). ELVES: Ein computergestütztes Diagnostikum zur Erfassung der Effizienz von Teilprozessen des Leseverstehens [ELVES: a computer-based test for measuring efficiency of processes involved in reading comprehension]. *Diagnostica, 51,* 169-182.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science.* New York: Academic Press.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441-474.

Zwaan, R. A., Graesser, A. C., & Magliano, J. P. (1995). Dimensions of Situation Model Construction in Narrative Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 386-397.