Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 2, 137–155
© 2017 Waxmann

Sandra Schladitz, Jana Groß Ophoff & Markus Wirtz

Effects of different response formats in measuring Educational Research Literacy

Abstract

The use of appropriate response formats in competency testing has been a topic of interest for the last few decades. Especially the comparison of multiple-choice items with free-response items has been widely examined. The present study examines objective and subjective difficulty of those two response formats and furthermore addresses the question of construct dimensionality based on response formats. Test items measuring Educational Research Literacy were presented to 600 university students in Educational Sciences. To eliminate possible distortions from memory effects, stem-equivalent items of both formats were distributed among two test booklets and linked together with anchoring items. Comparing the response formats did not reveal a clear result concerning objective difficulty. Free-response items were in most cases subjectively rated to be more difficult than multiple-choice items. Objective item difficulty was in most cases not related to subjectively rated difficulty, independent of response format. Model comparisons suggested no differing dimensionality when a method factor based on response format was defined additionally. The results show that in the domain of Educational Research Literacy, there is no distinct advantage of one format over the other in terms of difficulty. This and the established unidimensionality suggest that both formats may be used in competency tests in this content domain.

Keywords

Response formats; Item difficulty; Educational Research Literacy

Sandra Schladitz (corresponding author), Institute for Psychology, Philipps-University Marburg, Gutenbergstraße 18, 35032 Marburg, Germany
e-mail: sandra.schladitz@staff.uni-marburg.de

Dr. Jana Groß Ophoff, Institute for Educational Science/Institute for Psychology, University for Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany
e-mail: jana.grossophoff@ph-freiburg.de

Prof. Dr. Markus Wirtz, Institute for Psychology, University for Education Freiburg, Kartäuserstraße 47, 79102 Freiburg, Germany
e-mail: markus.wirtz@ph-freiburg.de

Sandra Schladitz, Jana Groß Ophoff & Markus Wirtz

Effekte verschiedener Antwortformate in der Erfassung Bildungswissenschaftlicher Forschungskompetenz

Zusammenfassung

Der Einsatz angemessener Antwortformate in der Kompetenzmessung ist ein viel diskutiertes Thema pädagogisch-psychologischer Forschung. Dabei steht vor allem der Vergleich offener und geschlossener Antwortformate im Vordergrund. Die vorliegende Studie vergleicht die objektive und subjektive Schwierigkeit dieser beiden Formate und prüft die sich daraus ergebende Dimensionalität des Konstrukts Bildungswissenschaftliche Forschungskompetenz. 600 Studierenden der Bildungswissenschaften wurden Testitems in geschlossenem und offenem Antwortformat vorgelegt. Um Verzerrungen durch Erinnerungseffekte auszuschließen, wurde inhaltsgleiche Items in beiden Formaten auf zwei Testhefte verteilt und mit Ankeritems untereinander verlinkt. Im Vergleich zeigte sich kein klarer Vorteil eines Formats in der objektiven Schwierigkeit; subjektiv wurden jedoch Items mit freien Antworten eher schwieriger eingeschätzt. Für die meisten Items zeigten sich keine Zusammenhänge zwischen objektiver und subjektiv eingeschätzter Schwierigkeit. Modellvergleiche auf latenter Ebene deuteten nicht auf eine spezifische Dimensionalität in Abhängigkeit von den Antwortformaten hin. Die Ergebnisse sprechen gegen einen klaren Vorteil eines bestimmten Antwortformats im Bereich der Bildungswissenschaftlichen Forschungskompetenz. Zusammen mit der gefunden Eindimensionalität des Konstrukts legen die Ergebnisse nahe, dass in dieser Domäne beide Formate in Tests eingesetzt werden können.

Schlagworte

Antwortformate; Aufgabenschwierigkeit; Forschungskompetenz; Bildungswissenschaften

1. Theoretical background

1.1 Educational Research Literacy

Educational Research Literacy (ERL; Groß Ophoff, Schladitz, Lohrmann, & Wirtz, 2014) is defined as the ability to purposefully access, comprehend, and reflect on scientific information, as well as to apply the resulting conclusions to arising problems (Shank & Brown, 2007). This *Engagement with Research* can be distinguished from *Engagement in Research*, with the latter describing an active participation in the scientific community by generating new knowledge (Borg, 2010). Several aspects are relevant for ERL: *Information Literacy* describes the ability to generate appropriate research questions and to use resources effectively to find relevant information to answer those questions (Blixrud, 2003). The ability to read and interpret the findings, especially in quantitative domains, is described with *Statistical Literacy* (Watson & Callingham, 2003). The final important step is being able to reflect on and critically evaluate the results, which can be described as *Evidence-Based Reasoning* (Brown, Nagashima, Fu, Timms, & Wilson, 2010). Students and professionals with high competency levels in these aspects are able to base their decision-making on well-founded arguments and keep up with the constantly changing knowledge society (Grundmann & Stehr, 2012). The project *LeScEd – Learning the Science of Education* (Schladitz et al., 2013) focuses on conceptualizing and measuring ERL to gain better insight into the structure and levels of this competency. This is the basis to evaluate the success of (university) education, because it allows depicting the development of abilities over the course of students' professional lives.

An important factor in competency assessment is the use of an appropriate response format that does not distort the results. In the focus of the current study, two commonly used response formats are investigated, which require participants either to choose a correct response from a given set of possibilities (multiple-choice [MC] format) or to formulate a response without any given options (free-response [FR] format). These different demands may correspond with different underlying cognitive processes that occur during item processing (Hancock, 1994; Martinez, 1999), e.g., the processing depth of the content. Based on Bloom's taxonomy of learning domains (Anderson et al., 2001), the most basic domain is being able to simply remember information, while creating information is the highest domain. This corresponds with the demands of response formats. Memory performance can easily be tested with MC items, while a production task necessitates a FR format. Another distinction between the response formats is the inherent guessing probability. MC items with their presenting response options allow the test taker to simply recognize the correct response even if they would not have been able to freely produce it without the prompt. In the worst case scenario, participants would simply guess correctly and the results would be distorted because responses were based on luck instead of actual knowledge. To sum up, different response formats are appropriate in different situations, depending on content, learning domain, and risk of guessing.

The purpose of this paper is to examine issues related to the use of different response formats in assessing ERL. As some testing situations can be considered high-stakes testing with future educational or career decisions depending on the test results (Powell, 2012; Wilson, 2007), the main concern is objective performance. If different tests using different item formats are applied but one format shows a generally higher difficulty than the other, comparability of test results may be severely limited. Even though large scale competency assessments like the PISA studies (Programme for International Student Assessment; OECD, 2016) use the same tasks and therefore same response formats for all participants, smaller assessments like examinations in different universities may employ different formats and thereby give an advantage to some test takers while impairing the performance of others (Becker & Watts, 2001). Therefore, we first examined the objective difficulty of our test items measuring ERL. In addition to objective item difficulty, the subjectively evaluated difficulty of the items may also affect performance. If the test takers experience trouble in dealing with the items, their performance may decrease directly (Adler & Benbunan-Fich, 2015; Maynard & Hakel, 1997) or mediated by other factors like self-efficacy (Mangos & Steele-Johnson, 2001). Subsequently, empirical evidence should be identified whether both formats can be used interchangeably in a test or whether format effects have to be considered. If the aim is to measure the same competency with all items, statistical analyses should yield a solution with only one dimension: A second dimension based on response formats should not enhance the prediction of students' performance on the test items (Wang, Drasgow, & Liu, 2016). Those aspects are analyzed in this study in order to contribute to the validation of the test instrument measuring ERL.

1.2 Comparing multiple-choice and free-response formats

Previous research on comparing different response formats yielded ambiguous results. Some studies suggest that performance on MC items is significantly better than performance on FR items (Funk & Dickson, 2011; Powell, 2012), but others found no consistently superior performance for one or the other response format (Bleske-Rechek, Zeug, & Webb, 2007; Chan & Kennedy, 2002).

Those studies have focused on very diverse domains, e.g., psychology (Funk & Dickson, 2011), mathematical problem solving (Powell, 2012), general scholastic aptitude (Bleske-Rechek et al., 2007), and economics (Chan & Kennedy, 2002). The ambiguous results show that performance on different response formats seems to depend on test content or domain. Since there are no studies comparing response formats in the context of ERL, it is not possible to predict performance in this domain. Therefore, an exploratory approach is used to investigate the following research question:

> Research Question 1: Do items measuring ERL differ in their difficulty based on their response format?

1.3 Subjective item difficulty

As test performance is not only dependent on objective item difficulty, but also on subjective factors regarding the test taker, the latter should be taken into account as well when evaluating a test instrument. Subjective ratings can generally be focused on two aspects: The difficulty of the item or the test takers' own ability to solve it correctly.

The first aspect is rooted in Cognitive Load Theory (Brünken, Plass, & Leutner, 2003; Sweller, 1988) and defines critical aspects of a given task that may contribute to the mental effort necessary to solve it. *Intrinsic cognitive load* refers to the load due to complexity that is inherent to the task itself and independent of the instruction. *Extraneous cognitive load* is caused by task and instruction format, and *germane cognitive load* refers to the effort of the person to process the given information (Brünken et al., 2003). To assess cognitive load, participants can be asked to judge the difficulty of the task (Kalyuga, Chandler, & Sweller, 1999). As expected, research shows that higher subjective difficulty ratings correspond to lower objective task performance (Adler & Benbunan-Fich, 2015; Maynard & Hakel, 1997). Further findings point to a mediating effect of self-efficacy as a third component explaining this relationship (Mangos & Steele-Johnson, 2001).

The second aspect of self-ratings would be aimed at the test takers' abilities. Those subjective competency measurements have been used in the past (Braun, Gusy, Leidner, & Hannover, 2008; Linninger et al., 2015). However, in order to assess actual competencies, this approach is limited in its informational value, because it has been shown that students are generally not able to realistically evaluate their own competencies. This applies to school education (Freiberger, Steinmayr, & Spinath, 2012) as well as to university level education (Chevalier, Gibbons, Thorpe, Snell, & Hoskins, 2009). Schladitz, Groß Ophoff, and Wirtz (2015) report no significant relationships between subjective (i.e., self-evaluated) and objective performance on ERL (in this case Statistical Literacy and Evidence-based Reasoning). In that preliminary study, the self-evaluation was gathered as a global assessment at the beginning of the competency test. Therefore, this discrepancy was attributed to the abstract content of the self-evaluative questions.

Based on those findings and since the current study addresses the item properties rather than the persons' competency, participants were asked to rate the subjective difficulty of every single item. Studies usually focus on either objective or subjective competency and do not combine these two aspects. Combined with the content specificity mentioned above, this leads to the fact that there are no predictions as to whether objective and subjective item difficulty are related. The corresponding research question is of exploratory nature as well.

> Research Question 2: Are objective and subjective difficulties of items measuring ERL related?

1.4 Method factor

Additionally to item difficulty, the dimensionality of the respective competency may depend on response formats. This line of inquiry is less focused on performance on the items but rather on a method factor. A method factor would imply that response formats affect item difficulties distinctively. Multitrait-multimethod analyses (Campbell & Fiske, 1959) are of great importance in the development of new instruments as they help to define the underlying construct and to identify appropriate modes of assessment (Eid, Nussbeck, & Lischetzke, 2006). In assessing ERL with both MC and FR formats it is possible to determine convergent validity as both methods are supposed to measure the same latent trait.

Findings from large scale studies on language competency support the assumption of a method factor based on response format. Rauch and Hartig (2010) identified a two-factor model (with MC and FR items loading on one dimension and FR items loading on another nested dimension) as being superior to a one-factor model. On the other hand, in other subject domains like computer science (Bennett et al., 1990; Bennett, Rock, & Wang, 1991; Thissen, Wainer, & Wang, 1994), different response formats did not lead to multidimensionality. A meta-analysis (Rodriguez, 2003) suggests a dependency on test design: The author argues that the close relationship of stem-equivalent items in MC and FR format can be seen as an indicator for measuring the same construct. However, if the comparison was based on items with different stems, the relationship was less close, which indicates different constructs based on response format.

Based on the assumptions of the multitrait-multimethod approach and the fact that stem-equivalent items were used in this study, it is expected that both item formats measure the same construct.

> Research Question 3: Does a unidimensional model provide a superior fit to the data than a multidimensional model assuming the two response formats to be a distinct determinant of performance?

2. Methods

In a previous study within the project *LeScEd – Learning the Science of Education* (Schladitz et al., 2013), a competency test was developed to measure ERL, consisting of the three competency facets Information Literacy, Statistical Literacy, and Evidence-Based Reasoning (Groß Ophoff et al., 2014). In total, 148 test items were distributed to 20 test booklets and administered to *N* = 1,360 participants in an incomplete booklet design (Frey, Hartig, & Rupp, 2009). Both MC and FR formats were incorporated, which provided the basis for the current study.

2.1 Item selection

The items from the main study were used to develop stem-equivalent items for the current study. Items that were previously presented in MC format were – with appropriate adaptations in wording – transformed into FR items. To derive MC from FR items, distractors for the multiple-choice items were drawn from participants' incorrect responses. In most cases, for each item stem a minimum of four statements was given that had to be rated “right” vs. “wrong” or “yes” vs. “no”, respectively (complex MC format; Pohl & Carstensen, 2013). This was done to minimize the guessing probability as well as to allow for comparing different scoring patterns in upcoming studies. Only a few items feature different response formats, including simple MC format (e.g., "Which one of the following four statements is correct?") or ratings with three options (e.g., "A is correct" vs. "B is correct" vs. "A and B are correct").

2.2 Design

A common difficulty in comparing response formats is to ensure comparability of test items. A major problem in comparing MC and FR items arises because it would be invalid if participants had to complete two stem-equivalent items (i.e., items with the same content but different response formats) within the same testing situation. With this approach, the assumption of local independence, which is a necessary requirement for item analyses based on item response theory (IRT), cannot be guaranteed (Ferrara, Huynh, & Michaels, 1999). Some studies try to solve this problem by presenting FR items to the participants first, followed by stem-equivalent MC-items later (Funk & Dickson, 2011). This ensures that at least no response options influence the freely generated responses. Another possibility is to present items with different stems and compare performance on the MC items with performance on the FR items (Bennett et al., 1990; Bennett et al., 1991; Bleske-Rechek et al., 2007). This requires a very careful test construction, because it has to be ensured that both sets of selected items measure the same competency (Chan & Kennedy, 2002). In the presented study another design was used to avoid any of these critical effects: Identical anchoring items were given to two groups of participants in combination with stem-equivalent items of MC and FR format. Hence, the responses on the anchoring items allow calibrating items and participants on the same competency dimension using IRT-based linking procedures.

2.3 Materials

Competency items, overall 34 (17 MC items and 17 stem-equivalent FR items) were distributed among two test booklets (Table 1). By this means, strain for the participants was minimized and it was ensured that no stem-equivalent items were presented to the same participant. Additionally to those 34 items, six items (three of each format) were used in both booklets to allow for test-linking. As those six items only serve the purpose of linking the booklets and were not presented in two formats, they will not be included in the presentation of the results.

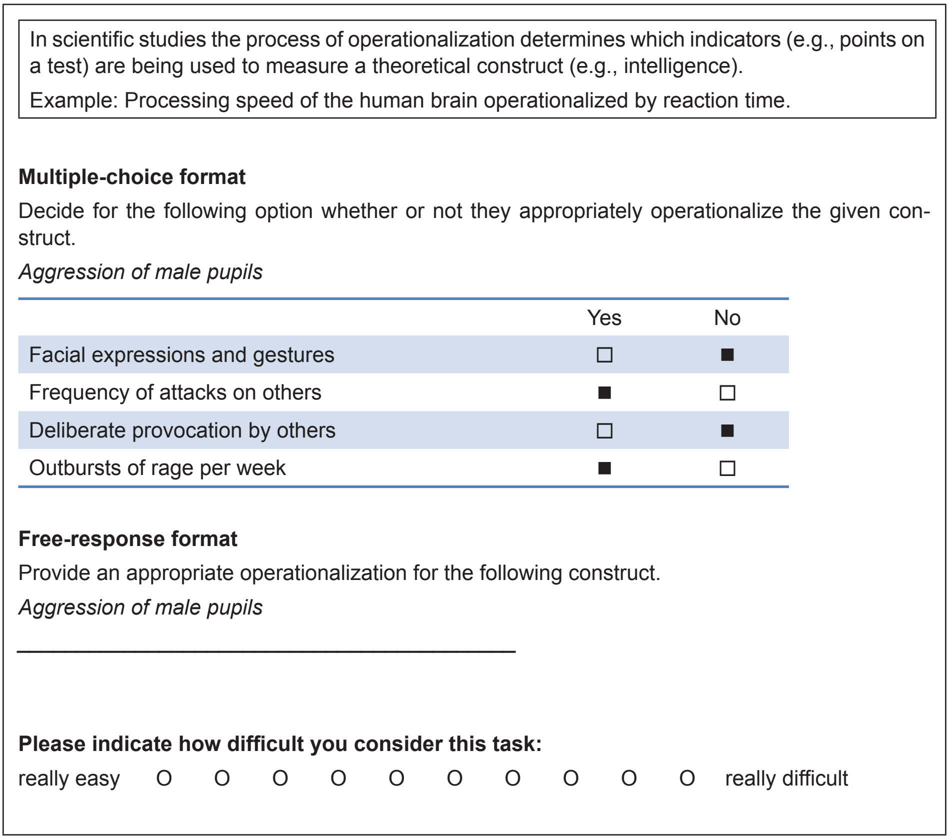
Table 1: Schematic representation of the booklet design

Booklet A	Booklet B
anchoring item 1 – anchoring item 6	anchoring item 1 – anchoring item 6
item 1 – item 9 MC format	item 1 – item 9 FR format
item 10 – item 17 FR format	item 10 – item 17 MC format

Note. MC = Multiple-choice format; FR = Free-response format.

To ensure a matched processing time, the quantity of MC and FR items was the same in each booklet. The order of the test items was chosen randomly. To avoid sequence effects, each booklet was implemented in two versions: forward and backward sequence. A sample item displaying both response formats to one item stem can be found in Figure 1.

Figure 1: Sample item to objectively assess Educational Research Literacy in both MC and FR format and subjectively rated item difficulty. Correct responses are marked in black



In scientific studies the process of operationalization determines which indicators (e.g., points on a test) are being used to measure a theoretical construct (e.g., intelligence).

Example: Processing speed of the human brain operationalized by reaction time.

Multiple-choice format

Decide for the following option whether or not they appropriately operationalize the given construct.

Aggression of male pupils

	Yes	No
Facial expressions and gestures	□	■
Frequency of attacks on others	■	□
Deliberate provocation by others	□	■
Outbursts of rage per week	■	□

Free-response format

Provide an appropriate operationalization for the following construct.

Aggression of male pupils

____

Please indicate how difficult you consider this task:

really easy O O O O O O O O O O really difficult

In addition to the competency test, demographic information was gathered at the beginning. After completing the booklet, participants were also asked to rate their enjoyment, diligence, and motivation during the test.

2.4 Sample

Participants were recruited during different courses in Educational Sciences (convenience sample). A total of *N* = 604 students filled out the paper-pencil test. Four students had to be excluded from further analyses because of disregard to the instructions. As is common in Educational Sciences, the majority of the participants were female (80.8 %). Students were on average 22.77 years old (*SD* = 4.03) and attending their third semester. Most of them were enrolled in the university's Teacher Training (58.0 %) and Pedagogy (28.5 %) programs; the minority was enrolled in other educational courses, e.g., Health Education and Early Childhood Education.

2.5 Rating of the competency items

The process of rating the responses will be described exemplarily using item #12 (Figure 1). In the MC format, each of the four single responses was rated correct or incorrect and in case of four correct single responses the whole item was rated correct (All-or-nothing scoring; cf. Pohl & Carstensen, 2013). Concerning the FR format, a correct response could be one of the correct ones from the MC responses, i.e., "Frequency of attacks on others" or "Outbursts of rage per week". Any other response that contained the important information was considered correct as well. For this item, this includes all kinds of quantifiable indicators of aggressive behavior, e.g., "How often he insults his classmates". Two raters were trained in determining correct responses and subsequently rated each response independently.

For inter-rater agreement, a Cohen's κ greater than .60 was considered sufficient (Fleiss & Cohen, 1973; Wirtz & Caspar, 2002). Consequently, five items below that threshold were excluded from further analyses – along with their stem-equivalent MC counterparts. Evidently, the respective guidelines were insufficient for a reliable rating and have to be revised before further analyzing these items. Following this selection process, 24 remaining items (12 of each format; excluding the anchoring items) were used for the following analyses.

2.6 Statistical analyses

Nonresponse was treated twofold: Up until the last processed item of each participant, nonresponse was coded as incorrect. After that breaking point, it was assumed that participants were not engaged with the remaining items. Accordingly, nonresponse was coded as missing afterwards. The treatment of missing values is a complex topic in competence modelling and is discusses in further detail in Groß Ophoff, Wolf, Schladitz, and Wirtz (2017). As for this study, participants completed on average 74 % and 75 % of the presented MC and FR items, respectively.

Objective item difficulty was derived from the frequency of correct responses to the item. This frequency was then inversed, so that 0 represented an item never solved correctly and 1 represented an item solved correctly by every participant. Subjective difficulty was assessed with a 10-point Likert scale ranging from "really easy" to "really difficult" for each item at the bottom of the respective page (see Figure 1). Hence, in both cases higher values indicate higher item difficulty.

Research question 1, whether there are differences between response formats, was analyzed separately for objective and subjective difficulty. First, responses were matched by item stem to allow comparison. To account for the dichotomous nature of the data on objective difficulty, we used logistic regression to test whether correctly solving the item could be predicted from response format. The interval scaled ratings of subjective difficulty were compared using independent *t*-tests. In order to analyze the relationship between objective and subjective item difficulties (research question 2), Pearson correlation coefficients were calculated. The basis for the third research question is an analysis of the test items based on IRT. Two Rasch models were compared using the statistical analysis software ConQuest (Wu, Adams, Wilson, & Haldane, 2001). The first, unidimensional model assumes all items measure the same construct, regardless of their response format. The second, between-item multidimensional model assumes two latent dimensions based on the two response formats MC and FR (Hohensinn & Kubinger, 2011). The information criteria Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC) were employed, with lower values indicating superior model fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

3. Results

3.1 Research Question 1: Effects of response format on objective and subjective item difficulty

There was no clear pattern to be found with regard to the objective difficulty of stem-equivalent items with different response formats, as shown in Table 2. Performance on four items could not be predicted from the response format. For the other eight items, significantly higher difficulty is almost equally distributed among MC and FR formats. Four items showed no differences in subjective difficulty between the response formats. In case of significant differences, the FR format was in six of eight cases regarded more difficult than the MC format.

Table 2: Comparison of item difficulties in multiple-choice and free-response formats

	Objective difficulty				Subjective difficulty			
# Item	MC	FR	Odd's Ratio[a]	*p*	MC	FR	*t*	*p*
1	.16	.21	1.38	.18	2.80	2.79	0.04	.97
2	.43	**.68**	2.80	<.001	3.38	**3.80**	2.11	<.05
3	.65	**.82**	0.41	<.001	4.39	4.21	0.89	.37
4	**.97**	.56	24.58	.001	**4.92**	4.28	2.77	<.01
5	**.85**	.74	2.03	<.01	4.49	4.83	1.57	.12
6	**.92**	.79	0.32	<.001	4.90	**5.62**	2.85	<.01
7	.98	**1.00**	0.00	.99	5.68	**7.13**	6.39	<.001
8	.65	.73	1.42	.09	3.95	3.73	1.11	.27
9	.74	**1.00**	74.55	<.001	4.78	**6.13**	5.78	<.001
10	.24	**.85**	0.06	<.001	4.60	**6.58**	7.45	<.001
11	**.94**	.41	0.05	<.001	**4.39**	3.76	1.49	<.01
12	.70	.78	0.68	.09	4.69	**5.72**	3.81	<.001

Note. MC = Multiple-choice format; FR = Free-response format. Objective difficulties range from 0 to 1 and subjective difficulties from 1 to 10, with higher values indicating higher difficulty. Items with higher difficulty are shown in bold print. *N* = 133-283.

[a] Significances were calculated using logistic regression to account for dichotomous data.

3.2 Research Question 2: Relationship of objective and subjective item difficulty

Correlational analyses revealed that only in some cases objective performance and subjectively rated difficulty were related. Of the MC items only two showed a significant correlation and of the FR items six proved to be significantly related (Table 3).

Table 3: Results of correlational analyses on objective and subjective item difficulties

	Multiple Choice		Free Response	
# Item	*r*	*p*	*r*	*p*
1	.43	<.001	.44	<.001
2	-.04	.62	.14	.06
3	.18	<.05	-.01	.85
4	.05	.46	.15	<.05
5	.06	.37	.11	.18
6	.04	.64	.16	<.05
7	.02	.77	–[a]	–
8	.04	.60	-.19	<.01
9	.05	.53	.12	.15
10	.10	.20	.39	<.001
11	-.01	.90	.15	.07
12	.05	.54	.18	<.05

Note. N = 133-235.

[a] Coefficient not computable because of variance being zero.

3.3 Research Question 3: Method factor

The IRT analysis revealed good fit indices of all items to the proposed model (0.80 ≤ Weighted Mean Square ≤ 1.20, as suggested by Linacre, 1994). The IRT-based item difficulties ranged from -1.73 to 5.48.

The model comparison revealed a better fit of the data for the unidimensional model, as indicated by the lower values of the information criteria (Table 4).

Table 4: Comparison of one- and two-dimensional models

	one-dimensional model	two-dimensional model (additional method factor)
Final Deviance	6938.51	6939.74
$N_{\text{parameters}}$	28	30
AIC	**6994.51**	6999.74
BIC	**7117.63**	7131.65
CAIC	**7145.63**	7161.65

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion. *N* = 600.

Rasch analysis of the data revealed a reliability (EAP/PV) of .32 for the unidimensional model. In examining the multidimensional model, the two dimensions MC format and FR format showed reliabilities (EAP/PV) of .30 and .31, respectively. This measure of reliability can be interpreted similarly to Cronbach's alpha. The recommended value of α ≥ .70 was not met in the current analyses. The values also failed to exceed the threshold of α ≥ .55, which is recommended for instruments used in group comparisons (Rost, 2013). The variance for the two dimensions MC format and FR format is .33 and .26, respectively, and they were correlated by .84.

4. Discussion

The presented study examined three aspects in the measurement of ERL: differences in item difficulty between MC and FR formats, the relationship between objective and subjective item difficulties, and the assumption of a specific method factor based on the two response formats. The particular strength of the design is the allocation of stem-equivalent competency items to two test booklets and linking them together using anchoring items. This allows for a direct comparison of items identical in content without memory distortions (Hohensinn & Kubinger, 2011; Rodriguez, 2003) or violations of local independence (Ferrara et al., 1999).

As performance on different item formats seems to be dependent on test and item content, no directed hypotheses regarding superior performance in either format were formulated. Results of the current study are somewhat ambiguous: In some cases the MC format was more difficult to solve and in other cases the FR format was more difficult to solve. This pattern was found for subjectively rated difficulties as well. These findings suggest that there is no clear advantage of either response format in the domain of ERL on an objective or subjective level. This is in line with previous research that could not identify a systematic effect as well (Bleske-Rechek et al., 2007; Chan & Kennedy, 2002). As mentioned before, different underlying cognitive processes may influence the processing of the competency items (Anderson et al., 2001). It is possible that one response format fits a certain aspect of ERL better than another. For example, in measuring Information Literacy, FR items proved to be more difficult than MC items. This may be due to the fact that the more technical aspects of, for example, using key words in search processes are harder to generate freely, because can rely on a given set of possibilities. On the other hand, the items measuring Evidence-based Reasoning show no systematic effect, disputing the notion that the assessment critical thinking is tied to FR format (Hancock, 1994; Martinez, 1999). Of course, other factors than difficulty must be considered as well, which should be the focus of future in-depth item analyses incorporating more items for each facet.

Regarding the subjective rating of item difficulty, there was a lack of research that allowed for a clear hypothesis about the direction of an effect. Our approach followed research in the field of Cognitive Load Theory (Sweller, 1988) as it focused on participants rating the difficulty of specific items rather than their own ability. This combines the advantages of a less abstract evaluation of tangible items with the possibility to determine the adequacy of students' difficulty perceptions. With a few exceptions, for the presented study we found that the objective performance on items measuring ERL was not related to the subjectively rated difficulty of the items. The results add more proof to previous findings on poor evaluative abilities of students (Chevalier et al., 2009; Freiberger et al., 2012; Schladitz et al., 2015). Evidently, those are not limited to the evaluation of their own ability, but also influence the evaluation of test items. On one hand, the underestimation of item difficulty might enhance students' test motivation, as tasks of very high difficulty may discourage them and negatively affect performance (Adler & Benbunan-Fich, 2015; Maynard & Hakel, 1997). However, it was also found that self-efficacy may mediate the relationship between subjective task complexity and performance (Mangos & Steele-Johnson, 2001). Therefore, future research should take both perspectives into account and check for relationships between subjective item difficulty, subjective competency, and objective performance.

The question of dimensionality of the competency construct was answered by conducting IRT-based model comparisons. The model incorporating only one dimension provided a superior fit to the data than a two-factor model based on response format. Evidently, both response formats seem to measure the same construct. Following the multitrait-multimethod approach (Campbell & Fiske, 1959), the results seem to indicate a high convergent validity of the assessment method. Validity concerning the measured trait was already established in studying the relationships of ERL and measures of general intelligence (discriminant validity) and self-evaluated ERL (convergent validity; Schladitz et al., 2015).

4.1 Practical implications

An important question in the development of a competency assessment is its practical use. The current analyses are not only supposed to provide information about statistical properties of the items, but to give indications about their use in real-life testing situations. This aspect is also part of the current conceptualization of validity as proposed in the 1999 Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). The major conclusions of the present research are the following.

As mentioned before, results on performance and dimensionality seem to depend on the content domain. For this reason, generalizing properties of a test measuring ERL to tests measuring other competencies should be treated with caution. Furthermore, Norman, Swanson, and Case (1996) suggest a number of at least 50 items to ensure that results not only apply to the specific items, but are generalizable to a broader range of items. If this criterion is not met, the specific item content possibly is a greater influence than the response format. This is especially relevant considering the underlying cognitive processes during item processing

(Hancock, 1994; Martinez, 1999) that occur based on the interaction of content and format. Since the current study does not contain that number of items, the following suggestions should be confirmed in future studies with a higher number of items.

From our findings, we can assume that in assessing ERL, both MC and FR formats may be used as there seems to be neither a distinct advantage of one format nor a specific format effect that would introduce another dimension into the construct. It is common practice to use multiple response formats in educational assessment (Wang et al., 2016), either for trying to assess different cognitive processes (Hancock, 1994; Martinez, 1999) or to make the sometimes tedious process more varied for the test takers.

Designing the test in a way that keeps participants interested is especially relevant regarding subjective item difficulty. A subjectively too difficult or too complex test may negatively influence self-efficacy which in turn proved to be negatively related to performance (Adler & Benbunan-Fich, 2015; Maynard & Hakel, 1997). Therefore, it is important to choose an item format that minimizes extraneous cognitive load (Brünken et al., 2003). Again, our results suggest that neither format shows a systematic advantage over the other, as in some cases MC and in other cases FR format was rated as being more difficult.

Independent of the content domain, such analyses contribute to the growing body of research on test and item properties. With the increasing focus on the output of education systems (e.g., Klieme et al., 2010) and high-stakes testing situations (Powell, 2012; Wilson, 2007), it is necessary to examine each assessment in detail to avoid biases because of response formats. To overcome the problem of content specificity, it is possible to adapt the instrument by re-formulating the items while keeping the content equivalent. For instance, reframing the items with examples from medical instead of educational sciences would allow assessing Research Literacy in health sciences, where evidence-based reasoning has been a subject of investigation for a long time (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996).

4.2 Limitations and prospects

Because of the low item number in the presented study, a multidimensional model based on competency aspects (Information Literacy, Statistical Literacy, and Evidence-based Reasoning; Groß Ophoff et al., 2014) would not have been applicable. This may be a reason for the low reliability as items are not supposed to measure the same aspect but comprise several aspects of ERL. Furthermore, the low item number itself and the fact that five of those items display difficulties in the upper percentile are other possible reasons for the low reliability. Since this aspect is not the focus of the presented hypotheses, the low value is not detrimental to the results. Though, for future analyses, putting greater emphasis on good test targeting may improve those values.

This unidimensional structure also is a deviation from the competency structure model used in previous analyses within this project (see Groß Ophoff et al., 2017). When including the whole item pool into the analyses, a bi-factor model (with one dominant factor representing the generic aspect of ERL and secondary factors representing the three competence aspects) proved to be the best fitting to the data. The existence of the general factor allows the use of a unidimensional model in this study. Nevertheless, in future analyses it would be ideal to include a higher number of items featuring both response formats and then compare different dimensional models, including a multidimensional model based on response formats and the bi-factor model from the large-scale study.

Especially those items that show an extremely high objective difficulty (4, 6, 7, 9, and 11; see Table 3) should be subject to further investigation. In these cases, the All-or-Nothing method of scoring may be detrimental and should be reconsidered in favor of Partial Credit scoring (Masters, 1982). Previous research suggests that this may yield different results (Kastner & Stangl, 2011). The complex MC format chosen in the current study enables us to apply such scoring methods in future analyses in order to determine possible differences in results.

The applied scoring method of treating nonresponse as incorrect before and as missing after the last processed items was used in order to align the current analysis to the previous study (Groß Ophoff et al., 2017). Comparing different scoring methods again with the current data set may yield further information on the properties of the competency items and on possible position effects within the test (e.g., Hahne, 2008).

In addition to the scoring methods it is also crucial to further examine the influence of the distractors in the MC items as they determine the difficulty of the item. To further contribute to the generalizability of the results regarding item difficulty, it would be beneficial to compare different sets of incorrect responses and analyze possible changes in the item properties.

References

Adler, R. F., & Benbunan-Fich, R. (2015). The effects of task difficulty and multitasking on performance. *Interacting with Computers, 27*(4), 430–439. doi:10.1093/iwc/iwu005

AERA – American Educational Research Association, APA – American Psychological Association, & NCME – National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Rahts, J., & Wittrock, M. C. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Becker, W. E., & Watts, M. (2001). Teaching economics at the start of the 21st century: Still chalk and talk. *American Economic Review, Papers and Proceedings, 91*(2), 446–451.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*(2), 151–162. doi:10.1177/014662169001400204

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77–92.

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education, 32*(2), 89–105. doi: 10.1080/02602930600800763

Blixrud, J. C. (2003). Project SAILS: Standardized assessment of information literacy skills. *ARL: A Bimonthly Report on Research Library Issues & Actions, 230*, 18–19.

Borg, S. (2010). Language teacher research engagement. *Language Teaching, 43*(4), 391–429. doi: 10.1017/S0261444810000170

Braun, E., Gusy, B., Leidner, B., & Hannover, B. (2008). Kompetenzorientierte Lehrevaluation – Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica, 54*(1), 30–42.

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*(3-4), 142–174. doi: 10.1080/10627197.2010.530562

Brünken, R., Plass, J., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. doi: 10.1207/S15326985EP3801_7

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. doi: 10.1037/h0046016

Chan, N., & Kennedy, P. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal, 68*(4), 957–971. doi: 10.2307/1061503

Chevalier, A., Gibbons, S., Thorpe, A., Snell, M., & Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review, 28*(6), 716–727. doi: 10.1016/j.econedurev.2009.06.007

Eid, M., Nussbeck, F. W., & Lischetzke, T. (2006). Multitrait-Multimethod-Analyse. In F. Petermann & M. Eid (Eds.), *Handbuch der Psychologischen Diagnostik* (pp. 332–345). Göttingen, Germany: Hogrefe.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement, 36*(2), 119–140. doi: 10.1111/j.1745-3984.1999.tb00550.x

Fleiss, J., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Freiberger, V., Steinmayr, R., & Spinath, B. (2012). Competency beliefs and perceived ability evaluations: How do they contribute to intrinsic motivation and achievement? *Learning and Individual Differences, 22*(4), 518–522. doi: 10.1016/j.lindif.2012.02.004

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53. doi: 10.1111/j.1745-3992.2009.00154.x

Funk, S., & Dickson, K. (2011). Multiple-Choice and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*(4), 273–277. doi:0.1177/0098628311421329

Groß Ophoff, J., Schladitz, S., Lohrmann, K., & Wirtz, M. (2014). Evidenzorientierung in bildungswissenschaftlichen Studiengängen: Entwicklung eines Strukturmodells zur Forschungskompetenz. In W. Bos, K. Drossel, & R. Strietholt (Eds.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (pp. 251–276). Münster, Germany: Waxmann.

Groß Ophoff, J., Wolf, Schladitz, S., Wirtz, M. (2017). Assessment of Educational Research Literacy in Higher Education: Construct Validation of the Factorial Structure of an Assessment Instrument comparing Different Treatments of Omitted Responses. *Journal for Educational Research Online, 9*(2).

Grundmann, R., & Stehr, N. (2012). *The power of scientific knowledge: From research to public policy*. Cambridge, United Kingdom: Cambridge University Press.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379–390.

Hancock, R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education,* 62(2), 143–157. doi: 10.1080/00220973.1994.9943836

Hohensinn, C., & Kubinger, K. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732–746. doi: 10.1177/0013164410390032

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351–371. doi:10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Kastner, M. & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia Social and Behavioral Sciences, 12*, 263–273. doi: 10.1016/j.sbspro.2011.02.035

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (Eds.). (2010). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster, Germany: Waxmann.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions, 7*(4), 328.

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D., Seidel, T., Terhart, E., & Kunter, M. (2015). Messung des Bildungswissenschaftlichen Wissens von Lehrkräften: Konstruktspezifikation und Validierungsansätze. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(2), 62–74. doi: 10.1026/0049-8637/a000126

Mangos, P. M., & Steele-Johnson, D. (2001). The role of subjective task complexity in goal orientation, self-efficacy, and performance relations. *Human Performance, 14*(2), 169–185. doi: 10.1207/S15327043HUP1402_03

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. doi: 10.1207/s15326985ep3404_2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Maynard, D. C., & Hakel, M. D. (1997). Effects of objective and subjective task complexity on performance. *Human Performance, 10*(4), 303–330. doi: 10.1207/s15327043hup1004_1

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine, 8*(4), 208–216. doi: 10.1080/10401339609539799

OECD – Organization for Economic Co-operation and Development. (2016). *PISA 2015. Assessment and analytical framework: Science, reading, mathemat-*

ic and financial literacy. Paris, France: OECD Publishing. doi: http://dx.doi.org/10.1787/9789264255425-en

Pohl, S., & Carstensen, C. (2013). Scaling of competency tests in the National Educational Panel study – Many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*(2), 189–216.

Powell, S. (2012). High-Stakes testing for students with mathematics difficulty: Response format effects in mathematics problem solving. *Learning Disabilities Quarterly, 35*(1), 3–9. doi: 10.1177/0731948711428773

Rauch, D., & 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*(4), 354–379.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184. doi: 10.1111/j.1745-3984.2003.tb01102.x.

Rost, D. H. (2013). *Interpretation und Bewertung pädagogsich-psychologischer Studien* (3rd ed.). Bad Heilbrunn, Germany: Klinkhardt.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*(7023), 71–72. doi: 10.1136/bmj.312.7023.71

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Schladitz, S., Groß Ophoff, J., & Wirtz, M. (2015). Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz. In S. Blömeke & O. Zlatkin-Troitschanskaia (Eds.), *Kompetenzen von Studierenden* (Zeitschrift für Pädagogik, 61. Beiheft, pp. 167–184). Weinheim, Germany: Beltz.

Schladitz, S., Rott, B., Winter, A., Wischgoll, A., Groß Ophoff, J., Hosenfeld, I., Leuders, T., Nückles, M., Renkl, A., Stahl, E., Watermann, R., Wirtz, M., & Wittwer, J. (2013). LeScEd – Learning the science of education. Research competency in educational sciences. In S. Blömeke & O. Zlatkin-Troitschanskaja (Eds.), *The German funding initiative "Modeling and Measuring Competencies in Higher Education"* (pp. 82–84). Berlin & Mainz, Germany: Humboldt Universität & Johannes Gutenberg Universität.

Shank, G., & Brown, L. (2007). *Exploring educational research literacy*. New York, NY: Routledge.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi:10.1016/0364-0213(88)90023-7

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2), 113–123. doi: 10.1111/j.1745-3984.1994.tb00437.x

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor item response theory approach. *Frontiers of Psychology, 7*(270), 1–11. doi: 10.3389/fpsyg.2016.00270

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3–46.

Wilson, L. D. (2007). High stakes testing in mathematics. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1099–1110). Charlotte, NC: Information Age Publishing.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen, Germany: Hogrefe.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2001). ConQuest (Version 2.0): Assessment Systems Corporation.