British Journal of Educational Psychology (2019), 89, 524–537

The British Psychological Society

www.wileyonlinelibrary.com

Validating process variables of sourcing in an assessment of multiple document comprehension

Carolin Hahnel[1,2]*, Ulf Kroehne[1], Frank Goldhammer[1,2], Cornelia Schoor[3], Nina Mahlow[4] and Cordula Artelt[3,4]

[1]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany
[2]Centre for International Student Assessment (ZIB), Frankfurt am Main, Germany
[3]University of Bamberg, Germany
[4]Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany

Background. With digital technologies, competence assessments can provide process data, such as mouse clicks with corresponding timestamps, as additional information about the skills and strategies of test takers. However, in order to use variables generated from process data sensibly for educational purposes, their interpretation needs to be validated with regard to their intended meaning.

Aims. This study seeks to demonstrate how process data from an assessment of multiple document comprehension can be used to represent sourcing, which summarizes activities for the consideration of the origin and intention of documents. The investigated process variables were created according to theoretical assumptions about sourcing, and systematically tested for differences between persons, units (i.e., documents and items), and properties of the test administration.

Sample. The sample included 310 German university students (79.4% female), enrolled in several bachelor's or master's programmes of the social sciences and humanities.

Methods. Regarding the hierarchical data structure, the hypotheses were analysed with generalized linear mixed models (GLMM).

Results. The results mostly revealed expected differences between individuals and units. However, unexpected effects of the administered order of units and documents were detected.

Conclusions. The study demonstrates the theory-informed construction of process variables from log-files and an approach for empirical validation of their interpretation. The results suggest that students apply sourcing for different reasons, but also stress the need of further validation studies and refinements in the operationalization of the indicators investigated.

The assessment of competencies is increasingly required to reflect on how well students can apply their acquired knowledge and skills in real-world contexts. Hence, performance

**Correspondence should be addressed to Carolin Hahnel, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Strasse 6, 60323 Frankfurt am Main, Germany (email: hahnel@dipf.de).*

DOI:10.1111/bjep.12278

assessments became particularly popular as they focus on assessing skills and strategies as directly as possible, by evaluating what students actually do on authentic tasks (Stobart & Gibbs, 2010). Standardized assessments, in comparison, are often criticized for reducing competence to the achievement of multiple-choice scores. However, with the aid of computer-based assessment (CBA), response processes can even be observed in standardized assessments by means of process data, such as from log-files. Although it often seems plausible to use certain log data indicators to represent intentional behaviours of individuals, there is a pressing need to examine whether such indicators work empirically, as theoretically expected. The present study investigates this for sourcing (i.e., activities of considering the background of documents) and examines whether the intended interpretation (Kane, 2013) of theory-informed indicators as 'sourcing' is more or less valid. Sourcing is a critical aspect required for following a discourse, for example, when readers attempt to create a comprehensive understanding of a topic from multiple perspectives. This study therefore uses a standardized assessment of multiple document comprehension (MDC) to evaluate whether meaningful episodes ('states'; Kroehne & Goldhammer, 2018) of sourcing occur as expected by current theories on the functions of sourcing. These episodes are examined for their relations to characteristics of persons, documents, tasks, and the administration of the MDC assessment, to validate their intended meaning and reduce ambiguities in their interpretation (Goldhammer & Zehner, 2017).

Functions of sourcing

Sourcing is described as 'attending to, evaluating, and using available or accessible information about the sources of documents, such as who authored them and what kind of documents they are' (Bråten, Stadtler, & Salmerón, 2018, p.8). Empirical findings suggest that sourcing differs systematically across individuals and reading situations. For example, source information is often not considered spontaneously (Wiley *et al.*, 2009) and more likely to get confused when arguments are similar in different documents (Braasch, McCabe, & Daniel, 2016). Especially when dealing with multiple documents (Britt & Rouet, 2012), sourcing enables readers to interpret and weigh information in light of the source features, helping them to understand and reconcile potential conflicts between different perspectives (e.g., Strømsø, Bråten, & Britt, 2010). There are several approaches how sourcing can take place in discourse comprehension (Bråten *et al.*, 2018), resulting in distinct functions of sourcing. Three areas of interest include sourcing as a heuristic, sourcing as consolidation of memory traces, and sourcing as a reaction to task instructions.

As a heuristic (i.e., cognitive shortcut to make fast and frugal judgements; Gigerenzer & Gaissmaier, 2011), sourcing takes over the function of providing an anticipatory framework for the subsequent encoding of text. Wineburg (1991) described sourcing in his expert–novice comparisons as actions, where individuals first look at the source of a document to identify specific biases and predict upcoming information. We refer to this as *proactive sourcing*.

Sourcing can also update memory traces of sources in readers, either to strengthen connections between claims and sources (Britt & Rouet, 2012) or to reconcile conflicts for restoring textual coherence (content–source integration model; Stadtler & Bromme, 2014). Accordingly, readers were found to pay more attention to the source of information when reading inconsistent, rather than consistent stories (Braasch, Rouet, Vibert, & Britt, 2012), which corresponds to the assumption of discrepancy-induced source

comprehension (D-ISC; Braasch & Bråten, 2017). If memory traces of sources need to be updated in working memory, readers are required to re-access source information, which we refer to as *repeated sourcing*.

Finally, particular task requirements can determine what information is currently relevant and might trigger sourcing. Attention is focused more on currently relevant than irrelevant concepts (Kaakinen & Hyönä, 2008). Therefore, task instructions can shift readers' focus of attention towards specific information, making them aware of potential differences between documents. As a result, readers might access source information directly after receiving particular instructions, which we refer to as *task-related sourcing*.

The present study

In summary, sourcing can serve several functions for readers who take source information into account when reading multiple documents. Due to the nature of the different functions described above, it is possible to identify specific conditions for the operationalization of log data indicators of sourcing whose interpretation can be empirically tested. Going beyond the investigation of whether or not a source was merely accessed, but opposite to a purely explorative and data-driven procedure, we propose a theory-driven validation approach. By contextualizing the observed behaviour of source access through theoretical assumptions about the driving cognitive process of sourcing, hypotheses can be formulated that address whether this behaviour occurs systematically and as theoretically expected. If they are not falsified, these hypotheses support the interpretation of an indicator; otherwise, the intended interpretation cannot be held (Kane, 2013).

Based on different theoretical assumptions about sourcing, validity arguments for proactive, repeated, and task-related sourcing allow hypotheses about systematic variations due to characteristics of persons, documents, tasks, and the test administration, strengthening or falsifying the interpretation of these sourcing indicators. On the person level, trait-like individual differences are expected to be explained by MDC skill since sourcing is generally assumed to be an important part of MDC (Britt & Rouet, 2012).

H1: Better MDC is positively associated with all sourcing indicators.

In addition to students' MDC skill, proactive sourcing should be related to students' general strategies in dealing with information. Assuming that students acquire knowledge on systematic acquisition, structuring, and use of information during their education, school graduation grades of university students should indicate acquired information strategies.

H2: Better graduation grades predict a higher probability of proactive sourcing over and above students' MDC skill.

H3: Graduation grades do not relate to repeated and task-related sourcing over and above students' MDC skill.

With respect to documents and tasks, proactive sourcing should not be affected at all as it takes place before document processing. In contrast, repeated sourcing can be affected, because the more documents, conflicts, or tasks involving source information need to be processed, the more memory traces of source information need to be

consolidated. Task-related sourcing should only adapt to variations in task-specific requirements.

H4: Proactive sourcing does not relate to any characteristics of documents and tasks.

H5: The probability a student engages in repeated sourcing behaviour increases by the number of documents, the number of conflicts between documents, and the number of tasks that require the comprehension of source information.

H6: The probability a student engages in task-related sourcing increases only as a function of the number of tasks involving source information.

Finally, we examined relations of the sourcing strategies to the administered positions of topics and documents as properties of the test administration.

H7: Differences in the sourcing indicators do not relate to properties of the test administration.

Methods

Assessment environment

For capturing source access, the design of an adequate assessment instrument needs to allow a differentiation between relevant parts within documents, namely the content and the source. This requirement is met in the MDC test of Schoor *et al*., (in press), which displays source information of documents on separate pages accessible by button clicks (number 3 in Figure 1), creating distinct events of accessing source information in log-files. Log-files record process data collected during the processing of a computer-based task, such as mouse clicks with timestamps. These log events can be used to reconstruct the test-taking process (Kroehne & Goldhammer, 2018), allowing different representations of sourcing behaviours to be captured (Table 1).

The MDC test assesses how university students deal with multiple documents. The test presents a computer-based, multi-page environment, in which students receive a set of two or three documents on a topic and items assessing how they compare, integrate, and evaluate information across documents and sources (Figure 1). Schoor *et al*., (in press)

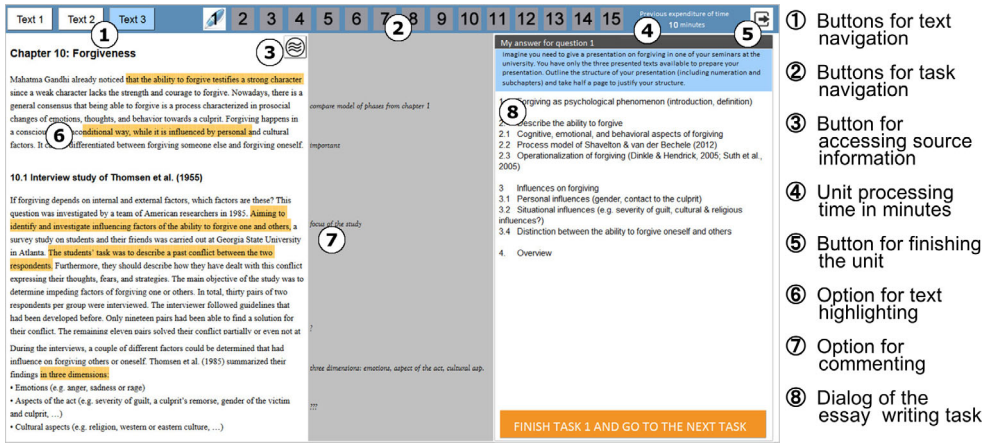

Figure 1. Example unit for assessing multiple document comprehension. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1. Overview over the process variables

Purpose	Process description	Operationalization of the process variable	ω_{RT}	*M*	(1)	(2)	(3)
(1) Proactive sourcing	Source information is accessed before a document is read	Dichotomous indicator of whether the source was accessed within the first 10% of the document processing time[a]	.94	0.34			
(2) Repeated sourcing	Source information is visited multiple times	Dichotomous indicator of whether the source was accessed multiple times in the reconstructed test-taking sequence	.91	0.34	.56		
(3) Task-related sourcing	Source information is accessed after item instruction	Dichotomous indicator of whether the state-trigram 'item–document–source' occurred, combined with a maximal duration of 10 s on the document[b]	.72	0.21	.14	.68	
General sourcing	Source information is accessed	Dichotomous indicator of whether the source of a document was accessed	.97	0.68	.77	.86	.74

Notes. The variables were derived per document (16 documents in total). ω_{RT} = Revelle's omega total. The means and tetrachoric correlations show the descriptive statistics of the variables in the long format (N = 2,485 observations).

[a]The 10% limit was chosen after a visual inspection of when students visited the source during document processing (Appendix A).

[b]The 10-s time limit was chosen after visually inspecting the distribution of time that students spent on the document in the trigram 'item–document–source' per unit (Appendix B).

developed six MDC units comprised of document sets with 174 corresponding items (Table 2). The units include topics from different domains (e.g., science, literary studies; Table 2); the contents are mostly fictitious (except for the unit 'Universe'), which limits possible prior knowledge. The students can move freely between document and item pages (numbers 1 and 2, Figure 1). There are options for text highlighting and commenting (numbers 6 and 7, Figure 1). In two units, the students are requested to write an essay before item completion (number 8, Figure 1). A comprehensive video-based tutorial introduces all functionalities.

Design and participants

The study focused primarily on the development of the computer-based MDC test (Schoor *et al.*, in press). After the participants gave their informed consent to participate, they filled in a questionnaire about demographic variables and received successively three randomly assigned MDC units (i.e., a total of 7 to 9 documents). The units were systematically varied according to a balanced incomplete block design, resulting in 60 testlets. The documents within units were always presented in the same order. The MDC units were not time-restricted. After each unit, the students could take a short break. A test session took about two hours.

An ad hoc sample of 310 university students (79.4% female) aged 18 to 31 years (M = 21.44, SD = 2.72) participated in the study (expense allowance of 20,-€). The students were enrolled in several programmes of the social sciences and humanities (70.0% Bachelor, 30.0% Master) of two German universities.

Measures

Dependent variables

To quantify sourcing, we decomposed the test-taking process into states of processing documents, sources, and items, and examined the empirical occurrence of sourcing under certain conditions. Table 1 provides details on the operationalization; it should be noted that the operationalization of proactive and task-related sourcing includes time constraints. Reliabilities ranged between .72 and .97 (Revelle's omega total; McNeish, 2017). For demonstrating the value of the theory-informed process variables, we also examined a general but less informative variable providing information on whether readers accessed generally the source information of a document or not (general sourcing). It should be noted that all process variables are to some degree interdependent. For example, if readers apply proactive sourcing, they will also show general sourcing, but not necessarily vice versa.

Person characteristics

Students' MDC skill was derived by modelling their performance on 67 dichotomously scored MDC items, which were selected for their psychometric properties and Rasch-scaled prior to this study (Schoor *et al.*, in press; correct response rate: 16.3% to 90.7% across items). Based on the documents model framework (Britt & Rouet, 2012), the items were constructed to mainly reflect one of four major MDC-specific aspects—namely (1) comparing and (2) integrating content information across documents, (3) comparing and evaluating sources, and (4) representing content as generated by particular sources. For a correct item response, information from at least two documents of a unit had to be

Table 2. Description of the units

Unit	Content description	*N* documents	*N* conflicts	*N* source-related items	Unit difficulty	*N* items	*N* observations
2134	Descriptions of the arrival of aliens on Earth in 2134	3	5	12	−0.92	35	156
Catalano	Biographical descriptions of the Mafia boss Catalano	2	2	3	−0.29	22	154
Nothing	Book reviews on the novel 'Nothing'	2	3	11	−0.99	36	151
Animals	Introductory texts about literary approaches for interpreting animals in novels	3	0	24	−0.37	36	153
Universe	Popular science texts about scenarios of how the universe will die	3	0	4	−0.98	17	160
Forgiving	Introductory texts about psychological theories of forgiving	3	1	22	−0.55	35	156

Note. The variables refer to all administered items and not to the subset of items selected for estimating students' MDC skill. The unit difficulty is the average of the item difficulty estimates per unit (Rasch model); higher values indicate that the items of a unit were harder to solve on average.

considered. The response formats included verification and multiple-choice formats. Weighted likelihood estimates served as MDC scores (WLE reliability = .69, SD = 0.75; Warm, 1989). Furthermore, students' graduation grades were assessed (M = 2.20, SD = 0.62). Note that German graduation grades range from 1.0 to 4.0 by .1 intervals, with lower grade points reflecting higher marks .

Unit characteristics

Unit characteristics were obtained to investigate differences between documents and tasks (Table 2). The following were considered: (1) the number of documents included in an MDC unit (two vs. three documents), (2) the number of conflicts between documents (how often does information in one document directly contradict information in another document), and (3) the number of items that required students to compare, evaluate, and consider source information when interpreting document information.

Properties of test administration

The position of a unit during a test session (first, second, or third position) and the position of documents within a unit (text 1, text 2, text 3) were considered.

Analysis

Generalized linear mixed models (GLMM) were conducted using the *R* package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). These models allow the probability of a binary outcome to be predicted by fixed effects (regression weights that are constant across observed units such as persons) and random effects (deviations specific to observed units), taking into account hierarchical data structures. Fixed effects were specified for all independent variables, units, and the unit difficulty (the average difficulty of items; Table 2), which was taken into account as a control variable; random effects were specified for persons. The regression coefficients represent the predicted change of the probability to show a particular sourcing behaviour in log odds. All continuous independent variables were z-standardized.

Results

Compared to general sourcing (M = 0.68), proactive, repeated, or task-related sourcing occurred less often (M = 0.21–0.34); their correlations were generally positive (Table 1). Figure 2 shows a detailed picture on the level of the MDC unit documents, showing first indications of situational differences in accessing source information.

The relations of the process variables with the characteristics of persons, units, and the test administration are reported in Table 3. Concerning the person characteristics, the results are in line with our expectations. All process variables were positively associated with the MDC test score (b = 0.27–.53; H1), and the prediction by graduation grades was significant for proactive sourcing (b = −0.42; H2), but not for repeated and task-related sourcing (H3). It is noteworthy that the graduation grades predicted proactive sourcing independently, although they were found to correlate moderately with students' MDC skill (r = −.40; Schoor *et al.*, in press).

Concerning the unit characteristics, as expected, proactive sourcing was not predicted at all by the unit characteristics (H4). The hypothesized positive relations with repeated sourcing were only partly found as expected (H5). The students were more

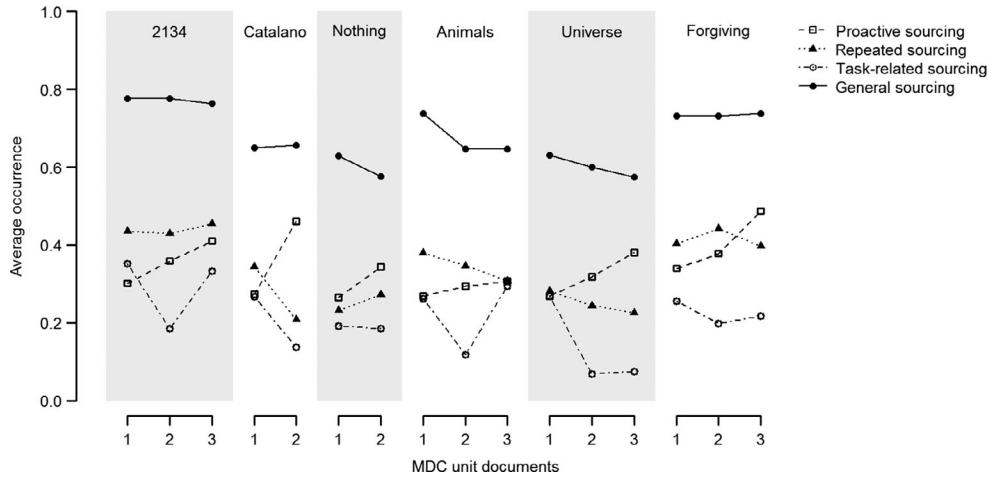

Figure 2. Description of the process variables, averaged across persons and unit positions.

likely to re-access source information when three, instead of two documents, were provided (b = 1.56), and the number of conflicting information between documents increased (b = 0.91). However, repeated sourcing was not significantly associated with the number of source-related items (b = 0.10). The prediction of task-related sourcing also just partly met our expectations. The document-related characteristics were not significantly associated with task-related sourcing; however, the number of source-related items was not either (H6).

Contrary to our expectations, the properties of test administration affected the process variables selectively (H7). Sources were visited more often when units were administered

Table 3. Results of the explanatory models

	Proactive sourcing	Repeated sourcing	Task-related sourcing	General sourcing
Intercept	−3.01 (0.35)***	−2.40 (0.31)***	−1.29 (0.26)***	1.33 (0.44)**
Unit difficulty	0.14 (0.13)	0.33 (0.11)**	0.14 (0.11)	0.12 (0.14)
Person characteristics				
MDC score	0.41 (0.18)*	0.53 (0.14)***	0.27 (0.08)**	0.91 (0.27)**
Graduation grade	−0.42 (0.18)*	−0.09 (0.14)	−0.02 (0.08)	−0.58 (0.27)*
Unit characteristics				
N documents	0.55 (0.67)	1.56 (0.59)**	−0.20 (0.55)	0.30 (0.78)
N conflicts	0.60 (0.47)	0.91 (0.41)*	0.07 (0.38)	0.09 (0.55)
N source-related items	−0.04 (0.15)	0.10 (0.13)	0.20 (0.12)	0.34 (0.17)*
Properties of test administration				
Position 2	1.32 (0.16)***	0.66 (0.14)***	0.12 (0.13)	0.70 (0.17)***
Position 3	1.63 (0.16)***	0.73 (0.14)***	0.27 (0.13)*	1.27 (0.18)***
Document 2	0.66 (0.14)***	−0.16 (0.13)	−0.83 (0.13)***	−0.33 (0.16)*
Document 3	1.04 (0.17)***	−0.25 (0.15)	−0.30 (0.14)*	−0.47 (0.19)*

Notes. The results are based on N = 2,485 observations. The regression coefficients represent the predicted change of the probability of sourcing in unstandardized log odds. Fixed effects of units were included in the model, but are not reported.
*p < .05; **p < .01; ***p < .001.

later than at the first position within a test session (b = 0.27–1.63; except for task-related sourcing, where no significant difference was found between the first and second unit positions). Compared to the first document within a unit, students accessed sources of other documents more often proactively (b = 0.66–1.04) and less likely in a task-related way (b = −0.30 to −0.81); only repeated sourcing was not significantly associated with the position of documents.

The comparative analysis of general sourcing revealed a mixture of patterns that were observed for the theory-informed process variables. It was related to MDC skill (b = 0.91) and graduation grades (b = -0.58); it showed to be unrelated to document-specific characteristics but associated with the number of source-related items (b = 0.34); and it was positively predicted by the administered position of units (b = 0.70–1.27) and negatively by the position of documents (b = −0.33 to −0.47).

Discussion

This study investigated process variables derived from different theoretical considerations about sourcing as a component of MDC. We assumed that the defined behavioural measures of proactive, repeated, and task-related sourcing represent different aspects of sourcing. Supporting this assumption, different correlational patterns were observed for the process variables that reflected the occurrence of these states during the test-taking process. Comparably, the indicator of general sourcing revealed a mixture of these patterns. Generally, the results support the overall postulation that data from log-files can be filtered and integrated in a theory-driven way to meaningfully represent individuals' strategies or skills (Goldhammer & Zehner, 2017). However, they also show that the interpretation of process variables strongly depends on the framework which conceptualizes a construct of interest.

Implications for interpreting the sourcing variables

Although the comprehensiveness of the MDC test tutorial might have triggered students' visits to sources via written instructions, a 7-second video clip, and a tryout dummy button, source information was not always accessed; only about one third was considered proactively, which is in line with previous research (e.g., Britt & Aglinskas, 2002; Wiley *et al.*, 2009). Compared to students who generally ignored source information, students who accessed sources were found to be more skilled in MDC, suggesting that they were more likely to represent an adequate documents model (Britt & Rouet, 2012). However, it should be noted restrictively that the MDC test includes items that require students to access source information for a correct item solution, which could artificially increase the association between the sourcing indicators and MDC. Therefore, it would be desirable to assess MDC and sourcing with independent instruments in future studies and cross-validate their relationship. Furthermore, the separate representation of document content and source information might be artificial for some multiple document settings, such as reading scientific papers. Although a separation is not unusual (e.g., About Us sections of websites), differences related to text types and expectations established in particular reader groups should be part of further investigations.

Concerning the investigated sourcing behaviours, students seem to access sources for different purposes, such as heuristic pre-evaluation (Wineburg, 1991), updating of memory traces (Braasch *et al.*, 2012), or due to shifts in task relevance (Kaakinen & Hyönä, 2008). The results of proactive sourcing showed that some students accessed source information comparably early in their course of document processing. Since it

was related to graduation grades over and above MDC skill and not affected by the unit characteristics, our results are in line with the assumption that students engage in proactive sourcing to pre-evaluate and structure newly encountered information (Wineburg, 1991). In contrast, repeated sourcing is rather a reaction of students to documents and tasks as they unfold during processing. The results support the interpretation that students revisited source information to reactivate memory traces of sources (Braasch *et al.*, 2012), despite the missing association with the number of source-related items, which might be an artefact of the MDC test construction since conflicts were mainly generated to create items. The results suggest for both proactive and repeated sourcing, source information is used rather than just passively received (e.g., by the positive relation of repeated sourcing and the number of conflicts). However, the interpretation of the indicators is still clearly restricted. The results cannot rule out alternative interpretations (e.g., students engage in proactive sourcing out of curiosity), and there is no direct evidence supporting other important aspects of sourcing (e.g., students use source information to interpret the specific content of a document).

Concerning task-related sourcing, the missing relationship with the number of source-related items is unexpected. The 10-second time limit on documents in the trigram 'item–document–source' might have been too long to actually reflect triggered source access. However, since this is supposed to be the critical characteristic of this sourcing indicator, we currently advise against interpreting it until its definition and operationalization are refined.

The observed but unexpected effects of the properties of the test administration allow for a number of explanations. Basically, all process variables were positively predicted by later administered unit positions, implying that later in the test-taking process, students were more likely to access sources. This might indicate that students needed to 'warm up' with the assessment situation or that they were capable of developing other strategies during the test-taking process facilitating their use of sources (test-wiseness; Downing & Haladyna, 2006). Although this does not directly threaten the intended interpretation of the sourcing indicators, it shows that the process variables might be valuable for investigating a reader's potential to learn how to deal efficiently with multiple documents while working with them.

With regard to the position of documents, students were more likely to access the source of the second and third document at the beginning of document processing (proactive sourcing), but less likely to access this source information after receiving task instructions (task-related sourcing). For proactive sourcing, it suggests that students have learned to process the documents within a unit efficiently over time, but also interestingly that the operationalized time limits work differently for documents within a unit. For task-related sourcing, the result seems odd, but might reflect a recency effect of students' memory for sources if they processed the second and third document last. Alternatively, if they decided to access source information after receiving task instructions, students might have systematically accessed the document sources in the suggested order, leading to more frequent visits of the first document's source. Source-to-source visits are not covered by the current operationalization of task-related sourcing.

Limitations

Our study is a first step towards a validation approach for indicators extracted from log data. Although this approach reduces the ambiguities in the interpretation of the indicators to some extent, the validation itself can only be regarded as preliminary. As

pointed out above, other alternative interpretations are still possible and require the further examination of validity arguments in order to ensure that inferences on the use of source information are eligible. There are various opportunities in the consideration of other third variables (e.g., prior knowledge, epistemic beliefs), experimental designs (e.g., comparing fictitious and real documents, varying the presentation of sources), and other methodologies (e.g., eye-tracking, micro-genetic methods). Related to this, although the MDC assessment was designed to record source access, the operationalized process variables do not claim exhaustiveness in the identification of possible cognitive functions of sourcing, as they cannot capture students' attention to, for example, implicit or indirect cues (e.g., text and surface properties as vocabulary, text comprehensibility, embedded sources; Bråten *et al.*, 2018). Furthermore, the definitions of at least proactive and task-related sourcing involved time criteria (Table 1). Although not arbitrarily chosen, they still might not be considered optimal since there is no precise definition of 'before reading' and 'triggered by task instructions'. The 10% limit has the advantage of taking into account individual differences in reading speed, but especially with long processing times, the limit can become too long to reflect source access prior to document processing. For the 10-second rule, it can be argued that this period is too long to represent sourcing triggered by task instructions. Finally, since our results are based on the analysis of an ad hoc sample, they cannot be generalized to the student or other populations.

Conclusions

The present study showcased the theory-informed construction of process variables from log-files and an approach for the empirical validation of their interpretation. The underlying rationale was to use an assessment instrument to measure both the outcome of students' comprehension of multiple documents and strategies that are supposed to closely relate to successful comprehension. Especially, the variables of proactive and repeated sourcing have proven useful in informing researchers and educators about how university students deal with source information and why some students perform poorly when working with multiple documents. Taken together, the process variables and the validation approach presented promise significant contributions to supplement different kinds of educational and psychological assessments.

Acknowledgements

The reported study was funded by the German Federal Ministry of Education and Research, funding number 01PK15008, within the research programme of KoKoHs ('Modeling and Measuring Competencies in Higher Education'). The data were assessed and analysed as part of the MultiTex project ('Process-based assessment of multiple documents comprehension'). The responsibility for the content of this publication lies with the authors. We want to thank two anonymous reviewers for their valuable insights and constructive comments.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Braasch, J. L. G., & Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: Basic assumptions and preliminary evidence. *Educational Psychologist*, *52*, 167–181. https://doi.org/10.1080/00461520.2017.1323219

Braasch, J. L. G., McCabe, R. M., & Daniel, F. (2016). Content integration across multiple documents reduces memory for sources. *Reading and Writing*, *29*, 1571–1598. https://doi.org/10.1007/s11145-015-9609-5

Braasch, J. L. G., Rouet, J.-F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition*, *40*, 450–465. https://doi.org/10.3758/s13421-011-0160-6

Bråten, I., Stadtler, M., & Salmerón, L. (2018). The role of sourcing in discourse comprehension. In M. F. Schober, D. N. Rapp & M. A. Britt (Eds.), *Handbook of discourse processes*. New York, NY: Taylor & Francis.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, *20*, 485–522. https://doi.org/10.1207/S1532690XCI2004_2

Britt, M. A., & Rouet, J.-F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781139048224

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 128–132. https://doi.org/10.1080/15366367.2017.1411651

Kaakinen, J. K., & Hyönä, J. (2008). Perspective-driven text comprehension. *Applied Cognitive Psychology*, *22*, 319–334. https://doi.org/10.1002/acp.1412

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. https://doi.org/10.1037/met0000144

Schoor, C., Hahnel, C., Artelt, C., Reimann, D., Kroehne, U., & Goldhammer, F. (in press). Entwicklung und Skalierung eines Tests zur Erfassung des Verständnisses multipler Dokumente von Studierenden [Developing and Scaling a Test of Multiple Document Comprehension in University Students]. *Diagnostica*.

Stadtler, M., & Bromme, R. (2014). The content–source integration model: A taxonomic description of how readers comprehend conflicting scientific information. In D. N. Rapp & J. Braasch (Eds.), *Processing Inaccurate Information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379–402). Cambridge, MA: MIT Press.

Stobart, G., & Gibbs, C. (2010). Alternative assessment. In P. Peterson, E. Baker & B. McGraw (Eds.), *International encyclopedia of education* (3rd ed., pp. 202–208). Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00307-9

Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction*, *20*, 192–204. https://doi.org/10.1016/j.learninstruc.2009.02.001

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, *46*, 1060–1106. https://doi.org/10.3102/0002831209333183

Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*(1), 73–87. https://doi.org/10.1037/0022-0663.83.1.73

Received 30 November 2018; revised version received 9 March 2019

Appendix A

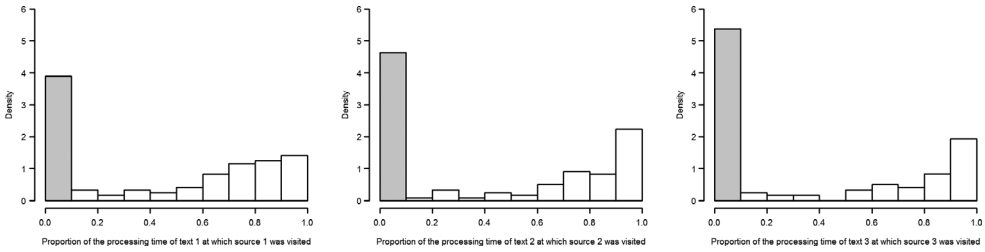

Figure A1. Source access during the time on a document (x-axis: normalized time on document); exemplary for the unit '2134′ (left: document 1, middle: document 2, right: document 3). The grey bars depict source access within the first 10% of students' document processing time.

Appendix B

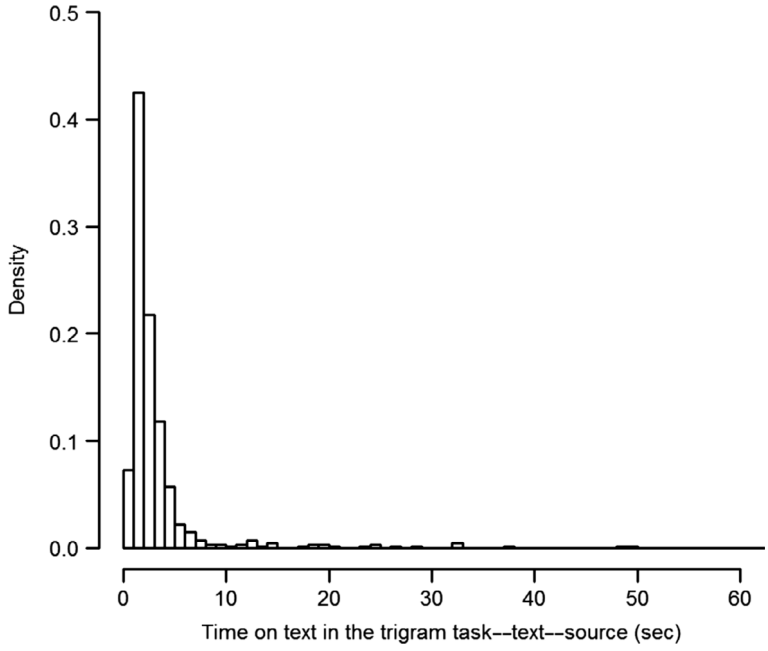

Figure B1. Distributions of the times that students spent on documents in the trigram 'item–document–source' (across all units). The figure shows the section of the first 60 s (*Min* = 0.75, *Max* = 403.75).