TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Empirische Bildungsforschung
Zentrum für Internationale Bildungsvergleichsstudien (ZIB) e.V.

Automatic Processing of Text Responses in Large-Scale Assessments

Fabian Zehner

Vollständiger Abdruck der von der Fakultät *TUM School of Education* der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil)

genehmigten Dissertation.

Vorsitzende:	Prof. Dr. Kristina Reiss
Prüfer der Dissertation:	1. Prof. Dr. Manfred Prenzel
	2. Prof. Dr. Frank Goldhammer, Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
	3. Prof. Dr. Peter Hubwieser

Die Dissertation wurde am 31. März 2016 bei der Technischen Universität München eingereicht und durch die Fakultät TUM School of Education am 22. Juli 2016 angenommen.

Abstract

In automatic coding of short text responses, a computer categorizes or scores responses. In the dissertation, a free software has been developed that is capable of (i) grouping text responses into semantically homogeneous types, (ii) coding the types (e.g., *correct / incorrect*), and (iii) extracting further features from the responses. The software overcomes the crucial disadvantages of open-ended response formats, opens new doors for the assessment process, and makes raw responses in large-scale assessments accessible as a new source of information. Three studies analyzed $n = 41{,}990$ responses from the German sample of the *Programme for International Student Assessment* (PISA) 2012 with different research interests, which were balanced according to three pillars. Publication (A) introduced the software and evaluated its performance. This involved pillar (I), which investigates, optimizes, and evaluates the algorithms and statistical models for automatic coding. Publication (B), studying how to train the software with less data, also concerned pillar (I) but then covered pillar (II) too, which deals with potential innovations in the assessment process. The article demonstrated how coding guides can be created or improved by the automatic system incorporating the variety of the empirical data. Pillar (III), attempting to add to content-related research questions, was covered in publication (C). It analyzed differences in the text responses of girls and boys to shed light on the gender gap in reading. Results showed *fair to good* up to *excellent* agreement beyond chance between the software's and humans' coding (76–98%), according to publication (A). Publication (B) demonstrated that, on average, established PISA coding guides only covered about 28 percent of empirically occurring response types, and, at the same time, the software enabled the automatic expansion of the coding guides in order to cover the remaining 72 percent. Publication (C) concluded that the difficulties some boys face in reading were associated with a reduced availability and flexibility of their cognitive situation model and their struggle to correctly identify a question's aim. The analysis showed among others that boy-specific responses were characterized by remarkably fewer propositions, plus those few propositions turned out to be more often irrelevant than those of girl-specific responses. The findings of the studies in pillars (II) and (III) illustrate how the developed approach and software can advance research fields and innovate educational assessment. The three pillars raised by this dissertation will be fortified in further studies. One of the crucial challenges will be to balance use cases and further software development, in order to take advantage of the innovations in natural language processing while identifying the demands in the social sciences. This paper ends with a detailed discussion on limitations, implications, and directions of the presented research.

Keywords: Computer-Automated Scoring, Automatic Short Answer Grading, Automatic Coding, Open-Ended Responses, Short Text Responses, Reading Literacy, Coding Guides, Gender Gap

Acknowledgments

A research project merely evolves in a researcher's mind alone but rather within a diverse setting of circumstances and individuals. While having been introduced to the academic and scientific world, I was highly privileged to be embedded into the lively theme park of the PISA studies with all its thrilling roller coasters and their most competent as well as kind operators.

There is Professor Dr. Manfred Prenzel. He pointed out the demands of the park investors and engineers to me and evaluated the newest inventions from the lab. His way to communicate as well as to treat content-related issues hopefully left its mark on me. There are Professor Dr. Frank Goldhammer and Dr. Christine Sälzer. They navigated me through the labyrinth-like paths in the park. They helped me to focus on investigating only one roller coaster instead of riding on two at a time; well, actually not the whole roller coaster in one study but only the blue screw at the lower left wheel of the third car, which tended to stagger a little in the final bend.

Further companions made the park attractions interesting and comforting at the same time. There is the national PISA team with, again, Dr. Christine Sälzer, Dr. Anja Schiepe-Tiska, Stefanie Schmidtner, Jörg-Henrik Heine, Julia Mang, and Elisabeth González Rodríguez. It is a pleasure to furnish the park with additional signposts, to install the newest swing ride and bumper cars, and to discuss them with you. There are the innumerable research assistants who helped recording the park visitors' talking with enormous effort and engagement. There is Jörg-Henrik Heine, always available for a stimulating chatting about techniques of log rides, LaTeX, R, or the latest IRT gossip. There is Dr. Markus Gebhardt, who showed me how to thrill the park visitors with an aquaponic system. There is Professor Dr. Torsten Zesch, who helped me a lot in setting up the new and fancy roller coaster firmware. There is the park's Department of Communication in the form of the TUM English Writing Center with Jeremiah Hendren, its staff, and its tutors. They taught me not to unethically bother stakeholders with linguistic redundancy and complexity.

There is my family. They are responsible for the outcome of my research for obvious reasons.

There is my wife. You are the plain motive for me to be in the park.

Contents

1 The Multiple Faces of Language

The studies in this dissertation deal with the automatic processing of short text responses in assessments. Put in a nutshell, they used a software for the automatic processing in order to (A) identify correct responses, (B) empirically improve reference responses in coding guides, and (C) capture features in responses that are sensitive to subgroup differences. While only the richness of natural language enabled these innovations at all, at the same time, they were restricted by some facets of language. In this sense, language has multiple faces that influenced the studies either constructively or destructively.

Language is one of our main ways to express cognitions. Since the quality of the cognitions is often at the core of educational and psychological research, many assessment instruments operate through language by presenting the stimulus and capturing the response. For the latter, the response can be assessed via closed response formats, such as multiple choice, or open-ended response formats, such as free-text fields. Because closed formats allow data to be easily processed, they have become state of the art. However, sometimes only open-ended opposed to closed response formats assess the full scope of the intended construct (e.g., for reading: Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2011; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006; for mathematics: Birenbaum & Tatsuoka, 1987; Bridgeman, 1991), which is why open-ended questions are typically included despite their detrimental impact in established large-scale assessments, such as the *Programme for International Student Assessment* (PISA; OECD, 2013b). Closed response formats evoke different cognitive processes than open-ended ones do. The two formats involve at least two different faces of language. One is the severe judge in a robe, faced with facts. The judge strikes the gavel onto the lectern upon evaluating each given response option as either wrong or right. The other face is an old and complex three-headed tortoise oracle, with its three heads in charge of information querying, solution finding, and solution expressing, the three heads constantly arguing with each other in order to come up with a result.

PISA assesses hundreds of thousands of fifteen-year-old students, who are instructed to respond to questions about a read text, mathematical problem, or scientific matter. The variability and sheer mass of the responses necessitate huge manual efforts from trained human coders deciding on the correctness of the responses. The human coders, in turn, entail some disadvantages (cf. Bejar, 2012): their subjective perspectives, varying abilities (e.g., coding experiences and stamina), the need for consensus building measures (i.a., coder trainings), and in turn a high demand on resources (i.a., time).

In order to overcome these disadvantages of the open-ended response format, a computer can be used to automatically code the responses. Therefore in this dissertation, a software has been developed that is capable of coding and scoring responses. Furthermore, responses contain linguistic information that hitherto could not have been used in large-scale assessments due to the large volume of data. Besides grouping the responses into semantically homogeneous types, the software thus also captures response features that provide further insights into the respondents. For example, the software checks whether information given in a response has been repeated from the stimulus or constitutes newly added information from the respondent.

The studies in this dissertation are centered around the software and assemble at three pillars. Studies in pillar (I) concern the development of the software for processing responses. They aid in identifying, evaluating, and improving appropriate algorithms, techniques, and statistical models (Zehner, Goldhammer, & Sälzer, 2015; Zehner, Sälzer, & Goldhammer, 2016). Studies in pillar (II) deal with new possibilities in the assessment process introduced by automatic processing of text responses. For example, the just mentioned publication Zehner et al. (2015) demonstrated how established coding guides can be improved by sampling prototypical responses from the empirical data as new reference responses. In this way, the coding guides cover the broad range of response types that human coders meet during their work. In pillar (III), the studies use the software to contribute to open content-related research questions. Zehner, Goldhammer, and Sälzer (submitted) explored features in boys' and girls' responses to further explain

the gender differences repeatedly found in reading. The studies analyzed data from the German PISA 2012 sample and mainly focused on questions assessing reading literacy, but they also included first evidence for mathematical and scientific literacy. Further studies will fortify and elaborate on each of the three pillars in the future.

All the software developments and analyses described have been accompanied by the multiple faces of language. On the one hand, natural language is abundant in information—a white-haired, long-bearded wise man sitting at the campfire who can tell a dozen stories when stumbling across a single word. This facet of language allows the assessment of multiple features from responses that furnish information about a respondent. It also enables the automatic scoring of responses as opposed to manual scoring, because the given linguistic information suffices to make scoring decisions. On the other hand, natural language is not an absolute system with symbols (words, phrases) of which each uniquely refers to one and only one entity (e.g., an object). Rather, in the terminology of Peirce (1897/1955), a reader perceives a symbol (word) and forms an idea of this symbol in their own mind, which is the interpretant—this process is often called grounding (Glenberg, de Vega, & Graesser, 2008). The symbol (word) itself is related to an object. Since the grounding process with the resulting interpretant is related to the object but not conclusively identical across different readers, language needs to be understood as a communication medium with information loss. Hence, a writer's interpretant tends to differ from the reader's interpretant. In this light, language is also a mystical fortune teller swirling over a crystal ball and extrapolating questionable ideas from observable signs. This face of language is very salient for human coders when confronted with student responses. Fortunately, human language comprehension is designed to deal with this fuzziness. However, this is a source for inconsistencies in human ratings. Computers, in contrast, are designed to work precisely and reliably. Attempting to make a computer work fuzzy, computer engineers experience remarkable obstacles when, for example, trying to generate a true random number, which might appear as a trivial task compared to the processing of natural language.

The faces of language depicted here are not exhaustive. For example, there is the red-headed, smirking twelve year-old boy, excitedly shifting from one foot to the other, who is longing to tell the latest pun. Or there is the eager eleven year-old girl, starting to learn a foreign language with eyes wide open due to the surprise when a new world appears on her horizon. Good professional writers are experts in utilizing the natural language's different faces, in order to manipulate what their readers perceive while reading. Analogously, we as researchers should be able to use the opportunities coming along with open-ended response formats instead of only being constrained by its complications. This dissertation contributes to lowering the required manual effort through automatic processing and attempts to increase the usage frequency of the open-ended format and more comprehensive information gain from the responses. That said, it needs to be emphasized that closed response formats are not generally considered to be worse than open-ended ones. But the researcher should simply not base the decision for either response format on the coding effort but on the construct.

2 Methods and Results

In this dissertation, a software has been developed and the studies were balanced according to three pillars. Publication (A) introduced the software, presented empirical evidence regarding its performance, and showed which parameter values and methods worked best for the German PISA data. This involved pillar (I) which investigates, optimizes, and evaluates the algorithms and statistical models for automatic processing. Publication (B), studying how to train the cluster model with less data, also partly concerned pillar (I) but then covered pillar (II) too, which deals with potential innovations in the assessment process through automatic text response processing. The article demonstrated how coding guides can be improved by automatic identification of response types in the empirical data. Pillar (III), attempting to add to open, content-related research questions, was covered in publication (C). It analyzed differences in the responses of girls and boys to further shed light on the gender gap in reading. The following subsections outline the corresponding research interests, methodological approaches, and main findings.

2.1 Publication (A): The Software for Processing Text Responses

The software development, its evaluation, further analyses, the designs of experiments, and the structuring along with the writing of the manuscript were carried out in the context of the dissertation by the first author. Also, the initial interest in the study was raised by the first author. The two co-authors advised on the analyses, on strategic decisions, such as the selection of data and the target journal, and on the final editing of the manuscript. It has been submitted to the journal *Educational and Psychological Measurement* in December, 2014, and accepted in February, 2015. It had been made available online in June, 2015, and appeared in April, 2016, in a print issue.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. doi: 10.1177/0013164415590022

2.1.1 Context and Related Work

The paper proposed a collection of baseline methods that is capable of categorizing or scoring student responses automatically. The methods were employed by software components under open licenses. The feasibility and validity of the method collection was demonstrated by using data assessed in PISA 2012 in Germany. Large-scale assessments, in particular, typically require enormous effort and resources in terms of human coding of open-ended responses. Naturally, they are a highly suitable field to apply automatic coding. This seems particularly true for international studies, since their inherent endeavor is to maximize consistency across different test languages (cf. OECD, 2013b).

For twenty years by now (Burstein, Kaplan, Wolff, & Lu, 1996), the issue of natural language processing for automatic coding of short text responses has been addressed by several research groups. But in contrast to the strongly related but somewhat older field of essay grading (Dikli, 2006), its methods are not commonly used in practice. In fact, a large number of open natural language processing libraries are available. These can be combined with statistical components in order to conduct automatic coding. In the last two decades, software and approaches have been published that can be used for automatic coding of short text responses, such as *c-rater* (Leacock & Chodorow, 2003) and *AutoTutor* (Graesser et al., 1999). Only in the last years, automatic coding has been receiving notable attention by larger research projects such as *Smarter Balanced Assessment* (Smarter Balanced Assessment Consortium, 2014).

It is beyond the scope of this paper to provide an extensive overview of existing systems; the reader might want to refer to the comprehensive overview by Burrows, Gurevych, and Stein (2014). The proposed approach and software differ from existing systems in the following characteristics. First, the approach does not yet include the most powerful machine learning methods. This however, second, offers more flexibility and transparency for researchers of the social sciences, because they are familiar with the methods and tools for the most part. Third, the software will be made available free

and open to encourage researchers to use automatic coding. Fourth, the software can be adapted, for example, by using the script language *R*, in order to tailor the response processing to the study's research questions.

2.1.2 Proposed Collection of Methods for Automatic Coding

The proposed procedure for automatic coding of text responses can be split into three phases. First, each text response is transformed into a quantified representation of its semantics. Second, the responses are grouped into response types by their semantics in a clustering model. Third, the response types are assigned to interpretable codes, such as *incorrect*, by machine learning. Figure 1 illustrates the procedure.

Phase One. In a first step of phase one, the text response is preprocessed (cf. part A in Figure 1). Beside further basic transformations, such as *punctuation removal*, *digit removal*, and *decapitalization*, *tokenizing* splits the response into word chunks. Next, *spelling correction* is applied. Functional words (*stop words*), not carrying crucial semantic information, are omitted. Finally, *stemming* cuts off affixes in the response. In the next step, representations of the responses' word semantics are needed. Therefore, a big text corpus is analyzed with a statistical method in order to build the machine's lexical knowledge. Wikipedia, for instance, can serve as an adequate corpus. *Vector space models* are commonly used to model semantics, called *semantic spaces*. These are hyper-dimensional spaces in which each word is represented by a vector. The semantic similarity of words is defined by the angle between two of these vectors. Semantic spaces can be computed by, for instance, Latent Semantic Analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) or Explicit Semantic Analysis (ESA; Gabrilovich & Markovitch, 2007). Now having semantic knowledge about words, the computer can extract the response semantics (cf. part B in Figure 1). The simplest way is to compute the centroid vector of all words in the preprocessed response. In the case of utilizing LSA, the result could be a 300-dimensional vector constituting the response's total semantics. This centroid vector represents the response in all further analyses.

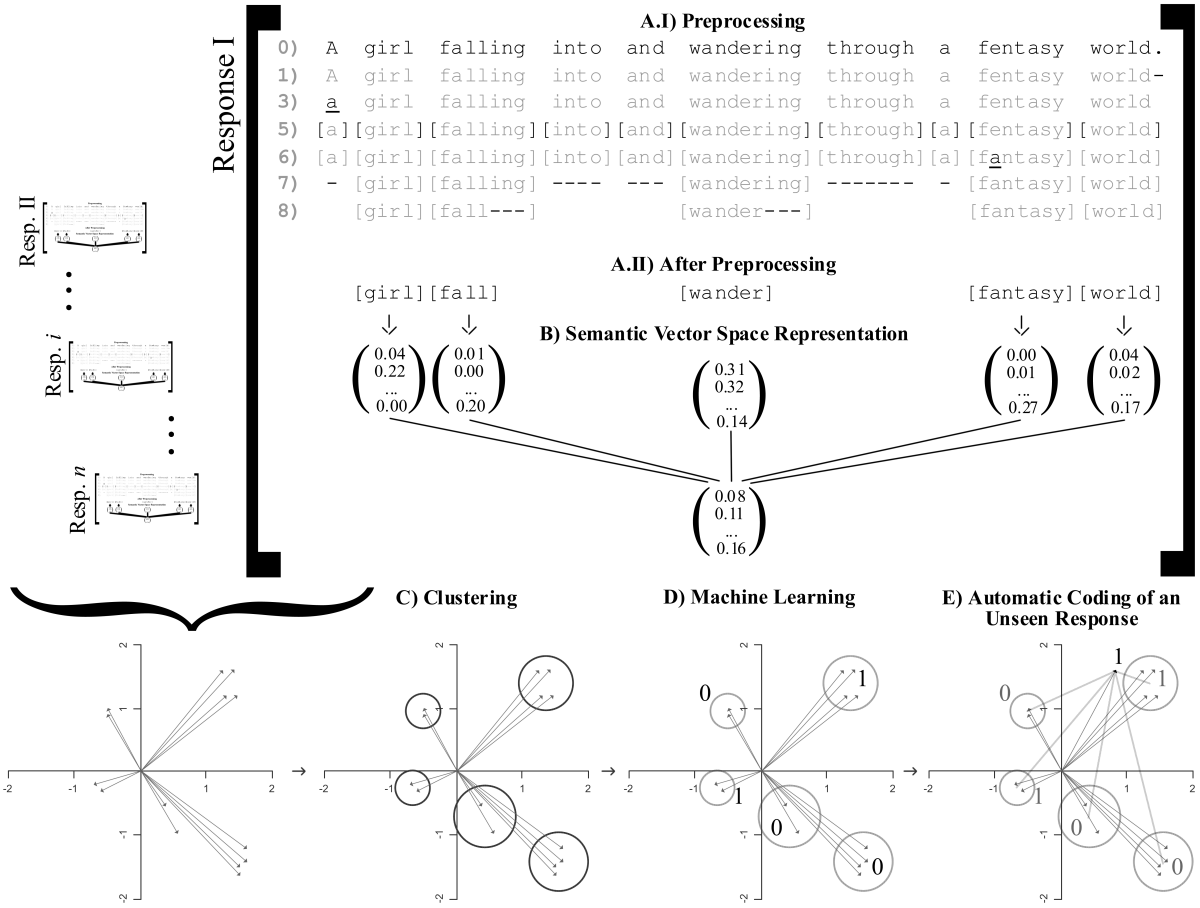

Figure 1: Schematic Figure of Automatic Coding (Zehner et al., 2016, p. 4)—In phase one, the computer preprocesses the response (A) and extracts its semantics (B). In the second and third phase, these numerical semantic representations are used for clustering (C) and for machine learning, in which a code is assigned to each cluster (D). Finally, a new response receives the code of the cluster which is most similar to it (E).

Phase Two. In the second phase, groups of similar responses are built by an agglomerative hierarchical cluster analysis (cf. part C in Figure 1). The groups constitute response types. The method has several adjustable parameters. First, (dis-)similarity is typically defined as the (arc-)cosine of two vectors. Second, it is suitable to use Ward's method (Ward, 1963) as the clustering algorithm. Third, the number of clusters is determined by the researcher in order to attain the best solution.

Up to this point, the procedure works with text response data only. This is especially reasonable for data sifting and is called unsupervised learning. The procedure does not utilize another criterion, such as a variable indicating whether a response is *correct* or *incorrect*. Such an external criterion would allow model optimization, which is then called supervised opposed to unsupervised learning. If some kind of supervised learning is the task's overall aim, such as scoring responses, the model parameter values should be varied systematically in order to choose the best performing one. This is reasonable, among others, for the most important parameter choice, which is the number of clusters.

Phase Three. Where the overall goal is automatic coding, meaning that a text response needs to be assigned to one of a fixed range of values (e.g., *correct*), an external criterion is used from which to learn the relation between the semantic vectors and the intended code (cf. part D in Figure 1). This is supervised learning and constitutes the third phase. The external criterion can be judgments by human coders. Supervised machine learning procedures are separated into two steps: *training* and *testing*. Each step in the procedure only uses a subset of the data and puts the rest aside. The method used in the paper is called *stratified, repeated tenfold cross-validation*. Details on the method can be found in Witten, Frank, and Hall (2011), and for comparisons with other methods see Borra and Di Ciaccio (2010).

For building the classification cluster model, the code of a response type (e.g., *correct*) is determined by the highest conditional probability of the codes for all responses that are assigned to the type. In this way, unseen responses can be classified by using the

cluster model and the codes of the response types (cf. part E in Figure 1). To do so, the highest similarity between the response and a cluster centroid determines the assignment of a response to a response type. The corresponding response type code is then assumed to be the response code. This is applied to the test data which had not been included in the model training, in order to compute the model performance. The simplest coefficient for the model performance is the percentage of agreement between computer and human. Another important coefficient for the inter-rater agreement is kappa (Cohen, 1960), which corrects for a-priori probabilities of code occurrences.

2.1.3 Participants and Materials

The analyzed n = 41,990 responses come from the German PISA 2012 sample. This includes a representative sample of fifteen-year-old students as well as a representative sample of ninth-graders in Germany. A detailed sample description can be found at Prenzel, Sälzer, Klieme, and Köller (2013) and OECD (2014). In PISA 2012, reading, mathematics, and science were assessed paper-based. Hence, the paper booklets needed to be scanned, and responses were transcribed by six persons. That is why not all items but only ten transcribed ones, including eight reading, one mathematics, and one science item, were at hand. All items were coded dichotomously, that is, responses either got full or no credit. Item and response contents could not be reported due to the items' confidentiality.

2.1.4 Research Questions

In order to evaluate the proposed collection of methods for automatic coding, the article answers three main research questions. The second research question splits up to seven different analyses that investigate the proposed collection of methods in detail.

1. How well does the proposed collection of methods for automatic coding perform using German PISA 2012 responses?

2. Which of the following steps and parameter configurations in the proposed collection of methods lead to higher system performance?

   (I) vector space model opposed to the existence of plain words
   (II) automatic opposed to none and manual spelling correction
   (III) Latent Semantic Analysis (LSA) opposed to Explicit Semantic Analysis (ESA)
   (IV) different text corpora as the base for the semantic space
   (V) different numbers of LSA dimensions for the semantic space
   (VI) different distance metrics for the clustering
   (VII) different agglomeration methods for the clustering

3. How well does the proposed collection of methods for automatic coding perform using smaller sample sizes?

2.1.5 Result Highlights

The agreement between human raters and the system reached high percentages from 76 to 98 percent ($M = 88\%$) across the ten items. According to Fleiss's definition (Fleiss, 1981), the kappa agreement ranged from *fair to good* to *excellent* agreement beyond chance ($.46 \leq \kappa \leq .96$). The models' reliabilities were acceptable with $\kappa \geq .80$ except of two items with $\kappa = .74$ and $\kappa = .77$.

The analyses studying the second research question demonstrated which parameter values or methods outperformed others. (I) Even for a question in which four terms needed to be repeated from the stimulus, a vector space model yielded higher performance than testing for the existence of plain words. The vector space model represents semantic concepts and, thus, among others, takes synonyms into account. (II) Spelling correction was important for some items but not for others. The automatic spelling correction reached similar performance levels like the manual correction. (III) LSA outperformed ESA. (IV) If text corpora were large enough, their domain-specificity did not matter anymore. (V) With at least 100 dimensions, the variation of the number of LSA dimensions did not impact the performance. (VI) The distance metrics *arccosine*, *Euclidean*

distance, and *Manhattan distance* performed equally well but significantly better than others. (VII) The agglomeration methods *Ward's* and *McQuitty's method* and *Complete Linkage* performed equally well but also significantly better than others.

In the sample size simulation experiment, the system performed equally well despite a large loss of data points from $n = 4152$ down to a sample size of about $n = 1600$. With a small but acceptable loss of performance, the system worked well with sample sizes down to about $n = 250$. With smaller sample sizes than this, the system should not be used for similar data.

2.2 Publication (B): The Use and Improvement of Coding Guides

The study was initiated by the first author in the context of the dissertation. The additional software development, analyses, and the structuring along with the writing of the manuscript were carried out by the first author. The two co-authors again advised on strategic decisions, such as the presentation of the manuscript. The article appeared in the conference proceedings of the *IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015)* that was held at the *IEEE International Conference on Data Mining* in Atlantic City, NJ, in November, 2015.

Zehner, F., Goldhammer, F., & Sälzer, C. (2015). Using and improving coding guides for and by automatic coding of PISA short text responses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015)*. doi: 10.1109/icdmw.2015.189

2.2.1 Context and Related Work

This study adapted the approach for automatic coding of text responses presented in publication (A). The adaptation reduces the manual effort for model training by using reference responses from so-called *coding guides* to start the model training. These are documents for manual coding. At the same time, the study attempted to show how to automatically improve the coding guides by adding empirical response types. The

procedure can also be used to systematically create coding guides from scratch. For the analyses, the data described for publication (A) were used.

In order to satisfy requirements of machine learning procedures, most systems for automatic coding rely on relatively large amounts of manually coded training data. The training data are expensive to collect and can also contain incorrect codes, mainly due to the required mass. Hence, different research groups have recently strived to find appropriate procedures to train models with less but most informative data (Zesch, Heilman, & Cahill, 2015; Dronen, Foltz, & Habermehl, 2014; Ramachandran & Foltz, 2015; Sukkarieh & Stoyanchev, 2009).

Conceptually comparable to the procedure proposed in this study, the *Powergrading* approach was published with partly remarkable performance (Basu, Jacobs, & Vanderwende, 2013). Despite the powerful method, its excellent performance might have partly stemmed from the relatively low language diversity in the analyzed responses evoked by the questions. Powergrading's drawback is, it uses a fixed number of clusters, which is not plausible for typical assessment needs, because different questions naturally evoke different numbers of response types. Moreover, the unsupervised system described in Mohler and Mihalcea (2009) also makes use of LSA similarities between responses, however, without grouping them. Their central idea is to expand the response key by words from the empirical responses with the highest similarity. A weakness of this is, instead of allowing for different lines of reasoning, it might only add synonyms to the response key, which already should have been considered similar by the vector space model. Since the processing of natural language is relevant to various different domains, such as dialog systems, a vast variety of implementations with different adaptations is available in the literature. For example, extrema are an interesting concept of weighting single words in vectors of a vector space model (Forgues, Pineau, Larchevêque, & Tremblay, 2014, December). Yet, the authors found that extrema do not outperform the baseline in the domain of questions.

2.2.2 Adaptation of the Processing of Responses

Briefly depicted, the approach described in publication (A) is used in order to group the empirical responses into types. Different to the original procedure, the researcher determines the number of clusters by the development of the residuals (Rasch, Kubinger, & Yanagida, 2011). In a next step, the software processes the reference responses from the coding guides in the same way as the empirical responses. Next, the reference responses are projected into the semantic space, and each is assigned to the most similar type. Ideally, all types then have unambiguously coded reference responses assigned. This model can serve for automatic coding, but it can contain the following conflicts.

Conflicts of type I represent response types without reference responses. Contrary, in Conflict II, multiple reference responses are assigned to one response type, however, they belong to different classes (e.g., *correct* and *incorrect*). This would reveal an insufficient semantic space or cluster model. In Conflict III, a reference response is excluded, because it is less similar to its cluster centroid than 95 percent of the responses are which had been assigned to the type. This conflict unveils reference responses that do not have empirical equivalents.

For Conflict I, a new empirical response needs to be sampled, so that it can be added to the coding guide. When a regression is carried out, the most informative responses can be selected by optimal design algorithms (e.g., Fedorov, 1972), which turned out to be highly effective (Dronen et al., 2014). For clustering approaches, the overall goal is to identify one response that is most prototypical for the whole type. Often, responses close to their centroid are simply assumed to be the most prototypical (e.g., Zesch et al., 2015). In Ramachandran and Foltz (2015), a list heuristic was used to find the response with the highest similarities and the most connections. In other terms, the heuristic seeks for the densest area within a cluster. In order to examine dense areas in clusters, the present study adopted an approximation similar to this list heuristic, making use of the fact that dense areas comprise many responses with relatively low

pairwise distances. For this, responses' pairwise distances to other responses within the cluster were sorted increasingly. Finally, those responses with the most relatively low distances belonged to the densest area. This procedure defines dense areas analogous to *kernel density estimates*, which generally provide powerful methods for identifying dense areas but cannot be applied to hyperdimensional spaces (cf. Scott, 1992).

2.2.3 Research Questions

The analyses in the study answered three research questions in order to evaluate established coding guides, provide empirical evidence about which responses should be sampled as new prototypes, and evaluate the newly proposed procedure.

1. How many Conflicts I, II, and III occur for PISA coding guides and the German PISA 2012 data?
   (I) empirical response type without reference response equivalent
   (II) contradicting reference responses within a response type
   (III) reference response without empirical response type equivalent
2. Which responses constitute the densest area within a response type and, thus, good prototypes?
3. How well does the new procedure perform ...
   (a) compared to the original procedure?
   (b) dependent on the number of clusters being extracted?

2.2.4 Result Highlights

A relatively small number of response types with contradicting reference responses (Conflict II) occurred across all items (0–2). These cases showed insufficiently specified response types in the corresponding semantic space or clustering model. A larger number, namely 21 percent on average, of the reference responses in the coding guides were redundant due to the lack of empirical equivalents (Conflict III). Analogously, 72 percent of the

empirical response types on average were not covered by the reference responses in the coding guides (Conflict I). The presented numbers are highly dependent on the number of extracted clusters, but nevertheless, they show the large potential for improvements in the coding guides.

The analysis focusing on the second research question demonstrated that responses close to the centroid belong to the relatively densest area within the response type. That was true across all response types and items. Because the employed list heuristic can be misleading in cases in which a very small dense area is present in the response type, the responses close to the centroid appear to constitute the optimal prototypes. They are located in the relatively densest area and are most representative for all responses in the type.

The performance of the new procedure turned out to vary unreliably if fewer than 100 clusters were extracted. From this point on, the performance became more accurate and reliable with not too much deviation from the original procedure, which uses all responses for training opposed to the new procedure which only uses about 2 percent when extracting 100 clusters. The requirement for a relatively high number of, and thus small, clusters probably stemmed from the fact that only one response was sampled as a new prototype for response types which lacked a reference response. It appeared to be an open question as to how to balance the number of clusters and the number of sampled prototypes.

2.3 Publication (C): The Reading Gender Gap in Text Responses

The initial interest in the research matter was brought up by the first author in the context of the dissertation. The additional software development, further analyses, and the structuring along with the writing of the manuscript were carried out by the first author. The two co-authors advised on the analyses, on strategic decisions, such as the presentation of the data and the target journal, and on the final editing of the manuscript.

It has recently been submitted to the journal *Reading Research Quarterly* in March, 2016, and the journal's peer review is currently pending.

Zehner, F., Goldhammer, F., & Sälzer, C. (submitted). Further Explaining the Reading Gender Gap: An Automatic Analysis of Student Text Responses in the PISA Assessment. *Reading Research Quarterly.*

2.3.1 Context and Related Work

Reading literacy is characterized by remarkably consistent gender differences of relevant magnitudes (e.g., Mullis, Martin, Foy, & Drucker, 2012; NCES, 2015; OECD, 2014). For this vivid research area, raw responses to open-ended questions have not been an accessible source of information hitherto. Particularly in large-scale assessments, their linguistic information could not be considered due to the mass of data. This study, first, attempted to add to the understanding of the gender gap in reading literacy by analyzing the natural language in short text responses. It identified responses that are typical for either of the genders and contrasted their features by processing the responses via the software described in publication (A). Second, it compiled a theoretical framework about how and which linguistic features in responses can be mapped to the preceding cognitive processes.

In PISA, girls consistently outperformed boys (OECD, 2014). That is the case across all countries—with only three non-significant exceptions—and across all cycles, constituting 268 comparisons in total. With the scale's standard deviation of 100, the gender gap average in the OECD ranged from 31 points in 2000 to 39 points in 2009. These numbers show an astonishingly stable figure that appears even more remarkable in the light that the effect was not always replicated in other studies (Elley, 1994; Hyde & Linn, 1988; Thorndike, 1973; White, 2007; Wolters, Denton, York, & Francis, 2014). The *Progress in International Reading Literacy Study* (PIRLS; Mullis et al., 2012) and the *National Assessment of Educational Progress* (National NAEP; NCES, 2015) found consistent but smaller effects. For adults, the *Programme for the International Assessment of Adult*

Competencies (PIAAC) did not find significant gender differences in reading literacy for most participating countries including Germany (OECD, 2013a).

All these variations across studies as well as the consistency within but not across PISA, PIRLS, and NAEP support the view of Lafontaine and Monseur (2009). They attributed the varying gender differences across studies to methodological decisions such as age- versus grade-based populations, and they emphasized the effect of whether the response format is open-ended or closed. Beside methodological reasons, mostly reading engagement and strategies have been emphasized (Artelt, Naumann, & Schneider, 2010).

Whether they are the original source for the differences or only a mediator of external influences, the difference would always be inherent to the students' cognitions. The theoretical framework depicts the features in responses that can be mapped back to the cognitions. Here, the framework is only briefly sketched. During reading and comprehending, the students build cognitive situation models comprising propositions explicitly given by, inferred from, or associated with the text, called micro- and macropropositions (Kintsch & van Dijk, 1978). In order to answer a question such as in the PISA assessment, the students identify the question focus and category. Then, they query their episodic and generic knowledge structures, including the situation model, and winnow them down to propositions that are relevant and compatible to the question focus and category (Graesser & Franklin, 1990). The students use these propositions in order to formulate the final response by concatenating and enriching them by linguistic structures specific to the question category (Graesser & Clark, 1985).

For the semantic measures (micro and relevance), gender types instead of the students' real gender served as the split criterion. That is, only semantic response types that comprised a significantly dominating proportion of one of the genders were included in these analyses and determined the group assignment. This was the rationale because it was not reasonable to assume that all responses by one gender were semantically homogeneous but that cognitive types exist that are dominated by one gender. For the analyses, the

data of the reading items described for publication (A) were used.

2.3.2 Automatic Processing of Features in Text Responses

Beside its capability to group responses into semantic types, the software now additionally annotated words with their parts of speech (POS). Specific words of particular POS were then considered proposition entities (PEs), genuinely referring to the situation model. The software captured (I) the proposition entity count (PEC), representing how many PEs were incorporated into a response. (II) The micro measure indicated the semantic similarity of the response's PEs to the stimulus and question. PEs with low values in the micro measure constituted macropropositions. (III) The relevance measure indicated the semantic similarity of the response's PEs to correct example responses from the PISA coding guides. Both, (II) and (III), used the maximum cosine similarity.

2.3.3 Research Questions

The analyses resulted from three research questions that investigated the gender gap in reading literacy and further explored the measures that were derived from the theoretical framework.

1. How do girls and boys differ in the number of propositions they use in their responses?

2. How do gender-specific responses differ with respect to the use of micro- vs. macropropositions and the extent to which these are relevant?

3. How are the measures for response features related to further variables, such as reading literacy and test motivation?

2.3.4 Result Highlights

A linear regression, controlling for the response correctness, revealed differences across gender types from 2.8 to 4.9 PEs. This shows that girl types always integrated significantly more elements from the situation model into their responses. That was true for correct as well as for incorrect responses. For four items, correct responses were additionally associated with more PEs, irrespective of the gender type. The analyses furthermore delivered empirical evidence for the plausibility to analyze gender types opposed to or on top of plain genders.

In order to not confound the measures with the PEC, the relative frequencies of relevant and irrelevant micro- and macropropositions within a response were analyzed. Summarized, girl types used more relevant and less irrelevant PEs than boys did. Similar to the PEC, this was generally the case for correct and incorrect responses, whereas the effect of the gender type on the frequency of relevant PEs was more pronounced for incorrect responses than for correct ones. This was similarly true for irrelevant PEs. Only for three items, the correct responses were similar across gender types, while the incorrect responses repeated the described figure of girl types using more relevant PEs. The girl types furthermore adapted more successfully to the level of reasoning within the situation model. That is, they used micropropositions if the question referred to the text base and macropropositions when the question asked for information that was not explicitly stated in the text. Boy types tended to do the exact opposite.

Briefly summed up, the boy types seem to involve a less stable situation model and struggle with retrieving and inferring from it. The high number of PEs integrated in girl-specific responses emphasizes that girl types allow to liberally juggle the information in the situation model. The relatively few PEs that boy types tend to use are more often irrelevant than those in girl-specific responses—and this is true regardless of the response correctness for almost all items.

3 Discussion

The first subsection disentangles the approaches and findings of the three studies and then links them. Corresponding to pillar (I), publication (A) constitutes the basic software development and evaluation. It forms the base for the two other studies. They are representatives of pillars (II) and (III), demonstrating by means of use cases how automatic processing of text responses can enhance assessment in research and how it can add to open content-related research questions. Both, the achievements and limitations are discussed. The second subsection takes up the constraints and outlines the implications and potential concrete future developments of the research initiated in the dissertation.

3.1 Discussing and Linking the Main Findings

Prior to publication (A), a software had been developed in the context of the dissertation that automatically categorizes and codes text responses. The software assigns, for instance, the codes *correct* and *incorrect* to responses. Despite the common terminology *scoring* (e.g., Leacock & Chodorow, 2003) and *grading* (e.g., Burrows et al., 2014), the dissertation's publications throughout use the term *coding*, in order to underline that scoring is a special case of coding and the proposed collection of methods is also capable of nominal, polytomous coding. Not all scoring techniques are capable of nominal coding, for example, those employing regression (e.g., Dronen et al., 2014). Generally, the plain development of the software in publication (A) served as the base for publications (B) and (C). The software's positive evaluation constituted a necessary requirement for legitimating further applications, such as the ones in the subsequent publications. The agreement beyond chance between the automatically generated codes and those produced by humans was judged as *fair to good* up to *excellent*, and the performance was judged as acceptable down to sample sizes of about 250 participants. Beside this minimum requirement of a positive evaluation, particularly the software's potential to group responses into semantically homogeneous types without requiring supervised training, which would

involve manually annotated data, catalyzed the wide-ranging possibilities the software offers. The significance of innovations attainable through the software is apparent in two key findings of the latter two publications. First, publication (B) showed that, on average, the analyzed established coding guides only covered about 28 percent of empirically occurring response types. At the same time, the software enabled the automatic extension of the coding guides in order to cover the remaining 72 percent. Second, publication (C) concluded that the difficulties some boys face in reading were associated with a reduced availability and flexibility of their cognitive situation model and their struggle to correctly identify a question's aim. The two examples show how the software can be utilized as a tool in assessments and the research process. Previously, text responses had been present in studies but had seldom been accessible as a source of information due to their large mass and unstructured form. Except for the goal of scoring, the volume of responses in large-scale assessments as well as the complications in objectively processing them made text responses a productive dairy cow that has rarely been milked.

The potential implications of the dissertation's development and findings are best regarded in the picture of the previously described three pillars. In future extensions of the presented research, it will be necessary to balance the further development of the software, on the one hand (pillar I), and its use cases for the assessment process and content-related research, on the other hand (pillars II and III). The technical aspect of natural language processing is a research matter of great interest in computer science at the moment (Cambria & White, 2014). Due to the growing demand for language processing in artificial intelligence (e.g., Woo, Botzheim, & Kubota, 2014), the analysis of big data (cf. Jin & Hammer, 2014), and dialog systems (e.g., Shrestha, Vulić, & Moens, 2015), an unwieldy body of corresponding technical innovations is being published and will continue to be published over the next decade. For the aim of improving human-machine interaction, the processing of natural language constitutes an indispensable component of complex computer systems, such as intelligent cars (e.g., Kennewick et al., 2015), and will become even more prevalent in every day human life with the further spreading of the

Internet of Things (cf. Ding, Jin, Ren, & Hao, 2013; Whitmore, Agarwal, & Da Xu, 2015). Claiming to be devoted to applied research, the extension of the presented studies will face the crucial challenge of selecting the relevant innovations from this technical research that can further enhance social science research and assessments. Thus, the balance of the three pillars will be important since only the identification of (feasible) demands in the social sciences, in the form of use cases, will point out which implementations of natural language processing are capable of advancing them.

In the assessment area, textual responses have customarily been used to extrapolate, for example, the participants' competency. Because this textual information is elevated to an aggregated level by the further processing through scoring and scaling, it might appear acceptable that the prevailing paradigm in educational and psychological assessment goes without an explicit theoretical framework about the underlying language. Contrary, for a study that uses basic linguistic information from text responses, it appears imperative to supply a corresponding theoretical framework for the operationalizations. *Which were the processes that resulted in the analyzed text? How are the processes and their outcome related to the construct / domain of interest?* The discussion of these questions has primarily been evolving in and is noticed by discourse and survey research (Graesser & Clark, 1985; Graesser & Franklin, 1990; Moore & Wiemer-Hastings, 2003; Tourangeau, Rips, & Rasinski, 2009). A corresponding theoretical framework for the context of reading assessments was specified in publication (C). It draws the path that information takes from the reading stimulus through the tested person's cognitions to finally show up in the text response. Nevertheless, further elaboration on the framework will be necessary and will add further insights about the observed measures.

Another crucial question has not been dealt with explicitly in the dissertation's articles yet and will be the subject of a work yet to come: *What is the conceptual understanding of semantics (meaning) in the implemented software and studies?* This matter is often critically discussed and considered in philosophy, discourse research, or computer science disciplines such as artificial intelligence (cf. Cambria & White, 2014; de Vega, Glenberg,

& Graesser, 2008; Foltz, 2003; Peirce, 1897/1955). In the dissertation, the basic concept for operationalizing semantics utilizes LSA. The founders of LSA believe that we gain the knowledge about the meaning of words mainly by the occurrences of words themselves (Landauer & Dumais, 1997; Landauer, 2011), opposed to the sensory world experiences we associate with the objects related to the words. These two model types are referred to as *amodal symbol models* versus *embodiment models* (Glenberg et al., 2008). Both are supported by empirical evidence. LSA's heart is the singular value decomposition. It is a powerful, purely mathematical method to implement an amodal symbol model, because it sensitively traces relationships of words via their co-occurrences and, even more important, via their non-co-occurrences (Martin & Berry, 2011). In the end, each word's meaning is represented by the direction its vector points at. To put it more precisely, the meanings are encoded by the angles between the vectors. Thus, the words are assumed to be pure symbols, and their interrelations represent their meanings. Studies that work at the linguistic level of text responses need to explicitly specify their implied understanding of language and meaning in order to verify the appropriateness of the employed operationalizations.

In addition to the symbolic concept of word meaning, further characteristics of the applied language models need consideration. For the most part, the dissertation employed the paradigm *bag of words*, which is inherent to LSA as well. That is, words are assumed to be isolated units not interacting with each other and syntactic information is most often neglected. All words in a response or in a text corpus document are thrown into the bag, irrespective of their original order and relations. It is apparent to every individual with a minimum of verbal abilities that there is more to natural language than the oblivious concatenation of syntactically context-free words. Such language use could be pictured as a beheaded chicken, wildly running around in its barn and crashing into every wall and object it meets on its way. At the same time, every individual might have experienced situations with improper communication, such as a phone call with a bad signal, trying to make a foreign language speaker understand the way to the underground, or the

chatting with a very small child. Sometimes these situations come along with momentous misunderstandings, sometimes they run surprisingly well and clear. The limitations of the bag of words are intuitive, diverse, and routinely criticized, among others, in the context of co-occurrence, which is the second paradigm in LSA (e.g., Glenberg & Mehta, 2008; Higgins et al., 2014; Leacock & Chodorow, 2003; Perfetti, 1998). However, the studies presented in the dissertation are only three representatives of a large body of studies (a) using the paradigms bag of words as well as co-occurrence and yet (b) resulting in remarkable, valid findings that advance the studies' research areas (e.g., Graesser et al., 1999). Apparently in spite of their shortcomings, these paradigms accomplish to capture the relationships of words sufficiently well in order to produce new, valid, reliable, and objective knowledge. For researchers of applied disciplines, the trade-off between feasibility and complexity of models is a critical challenge. The bag of words is a prototypical example for this trade-off when the condition of feasibility needs to be met. That said, it is most important to emphasize that, at the same time, it is a fine line to not simply justify any concept by its easy implementation, but the models need to prove their effectiveness in producing knowledge of scientific value. As shown, the applied paradigms are capable of doing so.

3.2 Limitations and Future Directions

The limitations as well as the general structure of future works stemming from the presented studies have already been touched on in the previous subsection. This final subsection names concrete research questions that need to be answered in subsequent studies. The discourse is structured along the three pillars.

The central limitation of the dissertation's software and analyses is introduced in pillar (I) in the general layout of the software, using the bag of words. Among others, it neglects syntax, relations between words, pronoun correlates, and negation by dedicated words such as *not* (opposed to negation by using antonyms). As the result of the use of

the bag of words, so-called *stop words* are often ignored, because they are assumed to only carry unimportant semantic information. As already mentioned, the techniques in natural language processing still show excitingly steep improvements with regard to their conceptual scopes. For example, *textual entailment* is a promising concept that is able to verify whether one sentence can be transformed into another one without loosing or biasing the original information; the interested reader might want to check the EXCITEMENT project (`https://sites.google.com/site/excitementproject/`, [2016-03-09]). Interestingly, it is a well-known phenomenon in the area of machine learning that the steep conceptual improvement is often not mirrored in accordingly steep performance improvements on tasks when the new concepts are implemented. Rather, very simple models, like the bag of words, already attain remarkable performance levels, whereas the use of additional, more sophisticated, and maybe also more complex techniques often only adds a fraction of incremental performance (e.g., see Higgins et al., 2014). Textual entailment is a prototypical example for the phenomenon. Although one would assume that the consideration of linguistic dependencies, negation, and linguistic inferences would give large raise to systems' performance levels, that is often not the case, if at all (e.g., Dzikovska, Nielsen, & Leacock, 2016). The main weakness of more fine-grained computational linguistic techniques, like textual entailment, often is their high dependency on well-formed language (Dzikovska et al., 2016; Higgins et al., 2014); even with regard to punctuation, about which students in low-stake assessments typically do not care at all. Especially responses in low-stake assessments, such as in the analyzed PISA data, are teeming with linguistic mistakes (spelling, syntax, punctuation), and regularly even human coders are not able to identify the response intention. Although it seems paradox, for this data imprecise methods like the bag of words appear to work more precise, because the necessary fuzziness for extrapolating the response intention is inherent to their low precision. Further developments for normalizing or recovering the language (similar to automatic spelling correction) might aid in making the fine-grained techniques usable for this area in the next decade. In addition to the discussed issues, further progress in the context

of pillar (I) will be possible when considering further models such as *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003; e.g., Saif, Ab Aziz, & Omar, 2016) or using hybrid techniques such as LSA combined with fuzzy logic (Mittal & Devi, 2016). Several other steps in the automatic coding process can and will be improved; for example, by utilizing more powerful machine learning methods such as *support vector machines*. Nonetheless, the balance between methodological power and transparency as well as flexibility for practitioners (researchers in the social sciences) is a central objective of the presented research and needs to be considered carefully. Next, the software was only evaluated by German data. Generally, the collection of methods and the software are easily scalable to other languages. This will be the core of another line of according studies. SNOWBALL can serve as a showcase for the collection's scalability, being one implementation of one of many stemming algorithms. It is available for six Germanic, six Italic, two Slavic, and two Uralic languages as well as for Turkish, Irish, Armenian, and Basque (Porter, 2001). Finally, the publication of the free software is due, which will allow practitioners to take advantage of automatic coding of short text responses without any costs.

In the context of pillar (II), further innovations regarding the assessment process are to be expected. For example, automatic coding is highly suitable to be employed in computerized adaptive testing (CAT), because it allows to extend CAT's scope to open-ended response formats (e.g., He, 2015, April). Since the aim of CAT is high economical efficacy, the corresponding studies will need to investigate the threshold of how many coding errors are acceptable before the adaptive testing procedure is not efficient anymore. Another possible innovation of the assessment process could be a probing mechanism during assessment. The background is that some test taker responses lack one single bit of information in order to exceed the threshold from incorrect to correct. But this does not mean that the test taker was not able to access this last bit. Since assessment is typically interested in the *latent* construct, it indeed should distinguish between those test takers who *are* able to give the missing bit and those who are *not*. It might turn out that a specific automatically identified response type contains such

kinds of responses; that means, they are almost correct but miss the last important bit of information. Automatic coding employed in computer based assessment would then give the opportunity to probe the test taker for the missing bit. Of course, the differential implications of such a mechanism, selectively probing specific students, would need thorough investigation whether it harmed the comparability across test cases. This would be similar to the underlying concept of learning tests (Guthke, 1992). The learning tests offer feedback to the student about the misconception that is observable in the (multiple-choice) response. The difference, of course, is that the learning tests' goal is to indeed teach the student by evoking specific cognitive processes, whereas this would be a confounding aspect for the probing mechanism, in which the feedback only serves as another stimulus to check if the student actually has access to the missing bit of information. This probing would be similar to what human test administrators do when test persons give borderline responses. More exemplary ideas, how the developed software can innovate assessments, can be found in publication (A).

For pillar (III), aiming to enhance content-related research questions, the main innovation of the dissertation is the new accessibility of raw responses as a new source of information. First, new assessment instruments can be developed that use open-ended response format. This way, they could improve the assessed scope of the construct. For example, personality questionnaires often force the respondents to evaluate their personality for given categories. But in the sense of Kelly (1955), it is reasonable to assume that some personality characteristics might be constructs that are rather peripheral relevant for some persons, others might be central to the behavioral tendencies. Open-ended assessment would allow to capture personality profiles that are not forced into a previously specified dimensionality structure, such as in the Big Five (Costa & McCrae, 1985) or 16 PF (Cattell, Cattell, & Cattell, 1993). Nonetheless, at the same time, the automatic processing would allow grouping the responses in order to know where the responses are located relative to others' responses. Analogous instruments in educational research with other constructs are similarly reasonable. Second, more objective assessment of person-

ality traits is possible by the automatic coding of text responses. For example, He (2013) impressively demonstrated how written texts can be used to predict psychiatric classifications. Third, cultural comparisons in international studies can be carried out at a new level of information; namely, semantic types of responses can be compared. Fourth, not only research could profit from the new developments in natural language processing, but also assessment in the educational context can be improved by digital learning environments that operate through written language (e.g., Graesser et al., 1999). Particularly, most recent innovations in the processing of speech (e.g., Zechner, Higgins, & Xi, 2007) promise to foster developments in the area of learning environments. Fifth, further construct domains will be of interest and require the development of other approaches of processing of the responses. For example, the processing of mathematical equations is a recently addressed topic (e.g., Lan, Vats, Waters, & Baraniuk, 2015; Smarter Balanced Assessment Consortium, 2014), as are scientific reasoning (e.g., Guo, Xing, & Lee, 2015) and programming source texts (e.g., Glassman, Singh, & Miller, 2014; Nguyen, Piech, Huang, & Guibas, 2014; Srikant & Aggarwal, 2014). Particularly the last research interest is recently driven by *Massive Open Online Courses* (MOOCs) platforms, but the entire research field of automatic coding will be dominated by this application in the next years (e.g., Jorge Diez, Alonso, Troncoso, & Bahamonde, 2015; Mi & Yeung, 2015). Similarly to large-scale assessments, in MOOCs, partly thousands of students need to be evaluated, and this mass of data challenges the evaluation procedures.

References

Artelt, C., Naumann, J., & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme et al. (Eds.), *PISA 2009* (pp. 73–112). Münster u.a.: Waxmann.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*(4), 385–395. doi: 10.1177/014662168701100404

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, *54*(12), 2976–2989. doi: 10.1016/j.csda.2010.03.004

Bridgeman, B. (1991). A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examinations General Test. *ETS Research Report Series*, *1991*(2), i–25. doi: 10.1002/j.2333-8504.1991.tb01402.x

Burrows, S., Gurevych, I., & Stein, B. (2014). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 1–58. doi: 10.1007/s40593-014-0026-8

Burstein, J., Kaplan, R. M., Wolff, S., & Lu, C. (1996). Using lexical semantic techniques to classify free-responses. In E. Viegas (Ed.), *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons* (pp. 20–29). Santa Cruz, California: Association for Computational Lingustics.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. doi: 10.1109/MCI.2014.2307227

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *16PF Fifth Edition Questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi: 10.1177/001316446002000104

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory.* Odessa: Psychological Assessment Resources.

de Vega, M., Glenberg, A. M., & Graesser, A. C. (Eds.). (2008). *Symbols and embodiment: Debates on meaning and cognition.* Oxford: Oxford Univ. Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, *5*(1).

Ding, Y., Jin, Y., Ren, L., & Hao, K. (2013). An intelligent self-organization scheme for the Internet of Things. *IEEE Computational Intelligence Magazine*, *8*(3), 41–53. doi: 10.1109/MCI.2013.2264251

Dronen, N., Foltz, P. W., & Habermehl, K. (2014). Effective sampling for large-scale automated writing evaluation systems. *arXiv preprint arXiv:1412.5659*.

Dzikovska, M. O., Nielsen, R. D., & Leacock, C. (2016). The joint student response analysis and recognizing textual entailment challenge: Making sense of student responses in educational applications. *Language Resources and Evaluation*, *50*(1), 67–93. doi: 10.1007/s10579-015-9313-8

Elley, W. B. (1994). *The IEA study of reading literacy.* Oxford: Pergamon Press.

Fedorov, V. V. (1972). *Theory of optimal experiments.* New York: Academic Press.

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss, B. Levin, & M. C. Paik (Eds.), *Statistical methods for rates and proportions* (pp. 598–626). New York: John Wiley & Sons.

Foltz, P. W. (2003). Quantitative Cognitive Models of Text and Discourse Comprehension. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.),

Handbook of discourse processes (pp. 487–523). Mahwah, NJ: Erlbaum. doi: 10.4324/9781410607348

Forgues, G., Pineau, J., Larchevêque, J.-M., & Tremblay, R. (2014, December). *Bootstrapping Dialog Systems with Word Embeddings.* Montreal. Retrieved from `http://www.cs.cmu.edu/~apparikh/nips2014ml-nlp/camera-ready/forgues_etal_mlnlp2014.pdf`

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (pp. 1606–1611).

Glassman, E. L., Singh, R., & Miller, R. C. (2014). Feature engineering for clustering student solutions. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 171–172).

Glenberg, A. M., de Vega, M., & Graesser, A. C. (2008). Framing the debate. In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment* (pp. 1–9). Oxford: Oxford Univ. Press.

Glenberg, A. M., & Mehta, S. (2008). The limits of covariation. In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment* (pp. 11–32). Oxford: Oxford Univ. Press.

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge* (Vol. 17). Norwood, NJ: Ablex.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*(3), 279–303. doi: 10.1080/01638539009544760

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., Tutoring Research Group, et al. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, *1*(1), 35–51.

Guo, Y., Xing, W., & Lee, H.-S. (2015). Identifying students' mechanistic explanations in textual responses to science questions with association rule mining. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* doi: 10.1109/icdmw.2015.225

Guthke, J. (1992). Learning tests-the concept, main research findings, problems and trends. *Learning and Individual Differences*, *4*(2), 137–151. doi: 10.1016/1041-6080(92)90010-C

He, Q. (2013). *Text mining and IRT for psychiatric and psychological assessment.* University of Twente.

He, Q. (2015, April). Combining automated scoring constructed responses and computerized adaptive testing. *Presented at the 2015 Annual Meeting of the National Council on Measurement in Education (NCME)*, Chicago, IL.

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., ... Blackmore, J. (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *CoRR*, *abs/1403.0801*. doi: arXiv:1403.0801v2

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69. doi: 10.1037/0033-2909.104.1.53

Jin, Y., & Hammer, B. (2014). Computational intelligence in big data [guest editorial]. *IEEE Computational Intelligence Magazine*, *9*(3), 12–13. doi: 10.1109/MCI.2014.2326098

Jorge Diez, O. L., Alonso, A., Troncoso, A., & Bahamonde, A. (2015). Including content-based methods in peer-assessment of open-response questions. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* doi: 10.1109/icdmw.2015.256

Kelly, G. A. (1955). *The psychology of personal constructs: Volume 1: A theory of personality.* New York: WW Norton and Company.

Kennewick, R. A., Locke, D., Kennewick, M. R., Kennewick, R., Freeman, T., & Elston, S. F. (2015). *Mobile systems and methods for responding to natural language speech utterance.* Google Patents. Retrieved from `https://www.google.com/patents/US9031845`

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363. doi: 10.1037/0033-295X.85.5.363

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, *8*(1), 69–79. doi: 10.2304/eerj.2009.8.1.69

Lan, A. S., Vats, D., Waters, A. E., & Baraniuk, R. G. (2015). Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*

(pp. 167–176). doi: 10.1145/2724660.2724664

Landauer, T. K. (2011). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Erlbaum: Mahwah. doi: 10.4324/9780203936399

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi: 10.1037/0033-295x.104.2.211

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*, 389–405.

Martin, D. I., & Berry, M. W. (2011). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Erlbaum: Mahwah. doi: 10 .4324/9780203936399

Mi, F., & Yeung, D.-Y. (2015). Temporal models for predicting student dropout in Massive Open Online Courses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* doi: 10.1109/ icdmw.2015.174

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., & McNamara, D. S. (2011). Assessing and improving comprehension with Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 207–225). Erlbaum: Mahwah. doi: 10.4324/9780203936399

Mittal, H., & Devi, M. S. (2016). Computerized evaluation of subjective answers using hybrid technique. In S. H. Saini, R. Sayal, & S. S. Rawat (Eds.), *Innovations in Computer Science and Engineering* (pp. 295–303). Singapore: Springer Singapore. doi: 10.1007/978-981-10-0419-3_35

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567–575). doi: 10.3115/1609067 .1609130

Moore, J. D., & Wiemer-Hastings, P. (2003). Discourse in computational linguistics and artificial intelligence. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 439–485). Mahwah, NJ: Erlbaum. doi: 10.4324/9781410607348

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International results in reading.* Boston College: Chestnut Hill, MA.

NCES. (2015). *The nation's report card: 2015 mathematics and reading assessments.* Retrieved 16.01.2016, from `http://www.nationsreportcard.gov/reading_math_2015`

Nguyen, A., Piech, C., Huang, J., & Guibas, L. (2014). Codewebs: scalable homework search for massive open online programming courses. In *Proceedings of the 23rd international conference on World wide web* (pp. 491–502).

OECD. (2013a). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills.* OECD Publishing. doi: 10.1787/9789264204256-en

OECD. (2013b). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. doi: 10.1787/9789264190511-en

OECD. (2014). *PISA 2012 results: What students know and can do (volume I, revised edition, February 2014).* Paris: OECD Publishing. doi: 10.1787/9789264208780-en

Peirce, C. S. (1897/1955). Logic as Semiotic: The Theory of Signs. In J. Buchler (Ed.), *Philosophical writings of Peirce* (pp. 98–119). New York NY: Dover Publications.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language. *Discourse Processes*, *25*(2-3), 363–377. doi: 10.1080/01638539809545033

Porter, M. (2001). *Snowball: A language for stemming algorithms.* Retrieved from `http://snowball.tartarus.org/texts/introduction.html`

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Ramachandran, L., & Foltz, P. W. (2015). Generating reference texts for short answer scoring using graph-based summarization. In Association for Computational Linguistics (Ed.), *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 207–212). doi: 10.3115/v1/w15-0624

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011). *Statistics in psychology using R and SPSS.* Chichester, West Sussex, UK: John Wiley & Sons. doi: 10.1002/9781119979630

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354–379.

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*(4), 441–474. doi: 10.1191/0265532206lt337oa

Saif, A., Ab Aziz, M. J., & Omar, N. (2016). Reducing explicit semantic representation vectors using latent dirichlet allocation. *Knowledge-Based Systems*. doi: 10.1016/j.knosys.2016.03.002

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley. doi: 10.1002/9780470316849

Shrestha, N., Vulić, I., & Moens, M.-F. (2015). Semantic role labeling of speech transcripts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 9042, pp. 583–595). Springer International Publishing. doi: 10.1007/978-3-319-18117-2_43

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced pilot automated scoring research studies*. Retrieved 2016-02-25, from `http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/09/Pilot-Test-Automated-Scoring-Research-Studies.pdf`

Srikant, S., & Aggarwal, V. (2014). A system to grade computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1887–1896). doi: 10.1145/2623330.2623377

Sukkarieh, J. Z., & Stoyanchev, S. (2009). Automating model building in c-rater. In *Proceedings of the 2009 Workshop on Applied Textual Inference* (pp. 61–69). doi: 10.3115/1708141.1708153

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study. International studies in evaluation, III*. Stockholm: Almqvist and Wiksell.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2009). *The psychology of survey response* (10. print ed.). Cambridge: Cambridge Univ. Press. doi: 10.1017/cbo9780511819322

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *American Statistical Association Journal*, *58*(301), 236–244. doi: 10.1080/01621459.1963.10500845

White, B. (2007). Are girls better readers than boys? Which boys? Which girls? *Canadian Journal of Education/Revue canadienne de l'éducation*, 554–581. doi: 10.2307/20466650

Whitmore, A., Agarwal, A., & Da Xu, L. (2015). The Internet of Things: A survey of topics and trends. *Information Systems Frontiers*, *17*(2), 261–274. doi: 10.1007/s10796-014-9489-2

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed ed.). Burlington, MA: Morgan Kaufmann. doi: 10.1016/B978-0-12-374856-0.00018-3

Wolters, C. A., Denton, C. A., York, M. J., & Francis, D. J. (2014). Adolescents' motivation for reading: group differences and relation to standardized achievement. *Reading and Writing*, *27*(3), 503–533. doi: 10.1007/s11145-013-9454-3

Woo, J., Botzheim, J., & Kubota, N. (2014). Conversation system for natural communication with robot partner. In *2014 10th France-Japan/ 8th Europe-Asia Congress on Mecatronics (MECATRONICS)* (pp. 349–354). doi: 10.1109/MECATRONICS.2014.7018624

Zechner, K., Higgins, D., & Xi, X. (2007). Speechrater: A construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education.*

Zehner, F., Goldhammer, F., & Sälzer, C. (2015). Using and improving coding guides for and by automatic coding of PISA short text responses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* doi: 10.1109/icdmw.2015.189

Zehner, F., Goldhammer, F., & Sälzer, C. (submitted). Further Explaining the Reading Gender Gap: An Automatic Analysis of Student Text Responses in the PISA Assessment. *Reading Research Quarterly.*

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. doi: 10.1177/0013164415590022

Zesch, T., Heilman, M., & Cahill, A. (2015). Reducing annotation efforts in supervised short answer scoring. In Association for Computational Linguistics (Ed.), *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 124–132). doi: 10.3115/v1/w15-0615

A Dissertation Publications

Publication (A)

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. doi: 10.1177/0013164415590022

Publication (B)

Zehner, F., Goldhammer, F., & Sälzer, C. (2015). Using and improving coding guides for and by automatic coding of PISA short text responses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015)*. doi: 10.1109/icdmw.2015.189

Publication (C)

Zehner, F., Goldhammer, F., & Sälzer, C. (submitted). Further Explaining the Reading Gender Gap: An Automatic Analysis of Student Text Responses in the PISA Assessment. *Reading Research Quarterly*.

Article

Educational and Psychological Measurement
2016, Vol. 76(2) 280–303

DOI: 10.1177/0013164415590022
epm.sagepub.com
SAGE

Automatic Coding of Short Text Responses via Clustering in Educational Assessment

Fabian Zehner[1,2], Christine Sälzer[1,2], and Frank Goldhammer[2,3]

Abstract

Automatic coding of short text responses opens new doors in assessment. We implemented and integrated baseline methods of natural language processing and statistical modelling by means of software components that are available under open licenses. The accuracy of automatic text coding is demonstrated by using data collected in the *Programme for International Student Assessment* (PISA) 2012 in Germany. Free text responses of 10 items with $n = 41,990$ responses in total were analyzed. We further examined the effect of different methods, parameter values, and sample sizes on performance of the implemented system. The system reached fair to good up to excellent agreement with human codings $(.458 \leq \kappa \leq .959)$. Especially items that are solved by naming specific semantic concepts appeared properly coded. The system performed equally well with $n \geq 1,661$ and somewhat poorer but still acceptable down to $n = 249$. Based on our findings, we discuss potential innovations for assessment that are enabled by automatic coding of short text responses.

Keywords

computer-based assessment, automatic coding, automatic short-answer grading, computer-automated scoring

[1]Technische Universität München, Munich, Germany
[2]Centre for International Student Assessment (ZIB) e.V., Munich, Frankfurt am Main, Kiel, Germany
[3]German Institute for International Educational Research, Frankfurt am Main, Germany

Corresponding Author:
Fabian Zehner, TUM School of Education, Centre for International Student Assessment (ZIB) e.V., Arcisstr. 21, 80331 Munich, Germany.
Email: fabian.zehner@tum.de

In automatic coding of short text responses (AC), a computer categorizes or scores short text responses. Such open-ended response formats, as opposed to closed ones, can improve an instrument's construct validity (e.g., for mathematics: cf. Birenbaum & Tatsuoka, 1987; Bridgeman,1991; for reading: cf. Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2011; Rauch & Hartig, 2010; Rupp, Ferne, & Choi, 2006). Yet today, closed formats have become state of the art, because they allow data to be easily processed and are more objective in nature. They do not require human coders, who entail some disadvantages (cf. Bejar, 2012): their subjective perspectives, varying abilities (e.g., coding experience and stamina), the need for consensus building measures, and, in turn, a high demand on resources. Taking a step back and reconsidering the response format's impact on construct validity, one realizes that these well-founded practical reasons in favor of closed formats might not be the most important decision criteria. Construct validity is the very base of all subsequent outcomes (e.g., analysis, conclusions, treatments) and, hence, can be claimed to be the most important criterion. Closed response formats should be used when the construct definition demands, or is supported by, closed response formats, and open-ended formats should be used when the construct definition requires open-ended ones.

One way to avoid or to compensate weaknesses of human coding is AC. Its employment has several implications. First, it is internally more consistent, because the final system is deterministic and always takes the same decision paths. This, in turn, circumvents consensus building measures. Second, the computer has stamina not to fatigue even at big data. Third, the coding of a new, unseen text response is done instantly and, thus, opens new doors in assessment. For example, computer adaptive testing (CAT) is constrained to the use of response formats that can be evaluated immediately. AC enlarges the scope of CAT to open-ended text responses. On the other hand, fourth, an AC system has its own limitations, errors, and costs (e.g., software development, manual coding for training data, etc.). The reader might notice that we use the term *coding* instead of the common *scoring*. That is because we regard scoring as a special case of coding and do not want to restrict AC's scope. In coding, a nominal category is assigned to an element; for example, a response could be categorized as either dealing with a financial or social aspect. On the other hand, in scoring, the categories additionally have a natural order; in the example, the social response could be favored over the financial one.

Since Carlson and Ward (1988), several research groups have been dealing with natural language processing (NLP) striving to automatically code short text responses. But in contrast to the strongly related field of essay grading, the respective methods are not commonly used in practice. Reasons for this, among others, may be proprietary licenses, under which most off-the-shelf software in this field is published, as well as a general human skepticism toward machines processing natural language (cf. Dennett, 1991).

In fact, a large number of open NLP libraries are available that can be combined with statistical components. In this article, we propose a collection of baseline

methods and related software components that can be used under open licenses. We implemented and integrated these components and report analyses demonstrating the method collection's feasibility and accuracy by applying it to data assessed at the Programme for International Student Assessment (PISA) 2012 in Germany. PISA is initiated by the Organisation for Economic Cooperation and Development (OECD) and assesses scientific, mathematical, and reading literacy in a large-scale context, partly by administering open-ended items. Large-scale assessments, in particular, typically require enormous effort and resources in terms of human coding of open-ended responses. Naturally, they are a highly suitable field to apply AC. This seems particularly true for international studies, since they inherently endeavor to maximize consistency across different test languages (cf. OECD, 2013).

The three main constructs assessed in PISA exemplarily show the importance of incorporating open-ended response formats. First, PISA defines mathematical literacy as ''the capacity to formulate, employ, and interpret mathematics'' (OECD, 2013, p. 25); particularly the first two elements can hardly be validly assessed in closed formats. Second, scientific literacy is, among others, defined as ''an individual's scientific knowledge and use of that knowledge … [as well as their] awareness of how science and technology shape our material, intellectual and cultural environments'' (OECD, 2013, p. 100). Particularly the latter clearly shows the requirement for respondents to integrate individual knowledge, ideas, and values, which is best assessed in open-ended formats; otherwise, the given response options would crucially influence the respondents' cognition. This is also true for, third, the construct of reading literacy, defined as ''understanding, using, reflecting on and engaging with written texts … [which] requires some reflection, drawing on information from outside the text'' (OECD, 2013, p. 61).

It is beyond the scope of this article to provide an extensive overview about existing AC systems. Rather, we would like to refer to Burrows, Gurevych, and Stein (2014) who give a comprehensive overview. The main differences of the approach presented here and software opposed to existing systems are the following. On the one hand, the approach does not yet include the most sophisticated machine learning methods with the advantage of higher flexibility and transparency for researchers of the social sciences by, among others, applying clustering. On the other hand, the software will be made freely available with a graphical interface to encourage researchers to use AC.

Our study was led by three research questions that investigate, first, the performance of the AC system; second, the need of single steps in the proposed collection of methods as well as possible alternate configurations; and third, the required sample size for its employment. This article aims to demonstrate the performance of an AC system for short text responses that relies solely on baseline methods. Thereby, researchers shall be encouraged to design instruments with open-ended response format where appropriate and assessment practitioners to use them. The following sections present the collection of proposed methods for AC, how the empirical analysis of these methods was conducted, and the obtained results. The results illustrate the

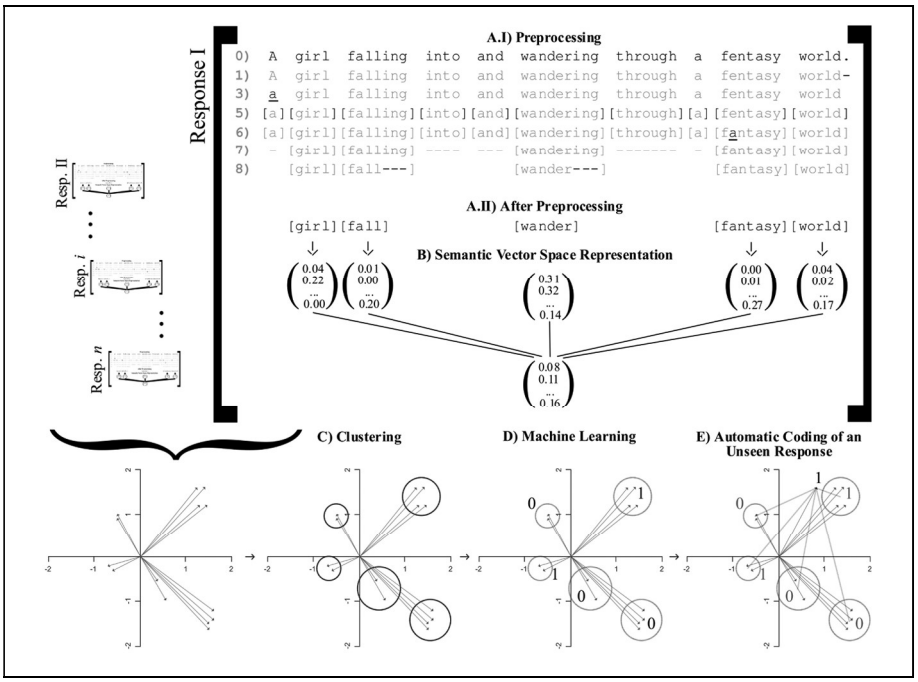

Figure 1. Schematic figure of automatic coding (AC).

Note. First, the computer preprocesses the response (A) and extracts its semantics (B). These numerical semantic representations are used for clustering (C) and machine learning in which a code is assigned to each cluster (D). Finally, a new response receives the code of the cluster that is most similar to it (E).

overall performance of AC and how it depends on the item evoking the free text response, the selection of a particular method, and its configuration as well as sample size. The final discussion explicates exemplary ways in which AC could enhance educational assessment.

Proposed Collection of Methods for Automatic Coding

The following subsections describe methods that step-by-step process a test taker's free text response represented as a character string. First, NLP methods are used to transform each text response into a quantified representation of its semantics. Second, clustering methods are used to build a clustering model by grouping similar responses based on the numerical representations of the responses. Third, machine learning methods are applied to assign an interpretable code to each of these groups. For easier comprehensibility, an example response leads through the various steps: Let ''A girl falling into and wandering through a fantasy world'' be a test taker's response to an item asking what the main plot of Lewis Carroll's *Alice in Wonderland* is about. Figure 1 illustrates the entire procedure using this example response.

The method selection was led by two main paradigms that helpfully balance the simplicity of assumptions and their effectiveness in practice. First, the so-called *bag of words*–approach is chosen. This means, all terms given in a response are considered separately, irrespective of their order and references to each other. Second, the *co-occurrence* paradigm serves as a tool for automatic extraction of semantic relationships between words. The idea is to use the phenomenon that semantically similar terms occur in similar contexts—not necessarily in exactly the same context but at least indirectly in similar ones (cf. Landauer, McNamara, Dennis, & Kintsch, 2011).

Making Sense of Responses via NLP Techniques

In a first step, the computer preprocesses the response. Therefore, NLP techniques split the given text response into tokens (words) and transform these into more normalized ones, that is, their variation is reduced. The preprocessing steps are described in the following, and each step's effect on the example response is visualized in Part A of Figure 1.

(1) *Punctuation removal* omits punctuations, similar to (2) *digit removal* omitting digits. Next, (3) *decapitalization* converts all chars to lowercase. For some languages, language-specific techniques need to be added. For instance, in German (4) *umlaut-normalizing* is used to change umlauts into their corresponding vowel (e.g., *ä* into *a*). Splitting the response into tokens, meant to be the level of analysis, is called (5) *tokenizing*. Next, it is advisable to apply (6) *spelling correction*. Then, functional words are omitted that do not carry crucial semantic information themselves. This is called (7) *stop word removal* and typically applies to words such as articles and conjunctions. As last very important variation reduction, (8) *stemming* is applied, which means to cut off affixes. The example now reads: *girl, fall, wander, fantasy, world.*

These preprocessing techniques are typically useful. Nevertheless, in some cases, it might be reasonable to adapt their assembly. If, for example, digits are important in the item's solution, they, of course, should not be omitted but considered. Furthermore, it is possible to replace stemming by the more sophisticated *lemmatization* (using the basic word form instead of just cutting off affixes), and dependent on language and application, it can be useful to apply *compound splitting* (splitting words that are made of more than one stem). The above listed eight techniques have been used for analysis presented in this article.

Now, having reduced linguistic variation in responses, the computer needs some kind of numerical representation of the remaining tokens' semantics (ignoring order information as a *bag of words*–approach is used). Therefore, a big text corpus is automatically analyzed by a statistical method—using the co-occurrence paradigm—to build the machine's lexical knowledge. This is often called *semantic space*. Wikipedia, for instance, can serve as an adequate corpus. If feasible, multiple sources can lead to higher performance (Szarvas, Zesch, & Gurevych, 2011). The semantic spaces used for analysis reported in this article were built on basis of a German Wikipedia dump[1] (from June 13, 2013) as German text responses were to be coded.

Solely in cases of items that only evoke a strongly limited amount of different words in the responses—for example, if test takers are required to repeat four terms explicitly given in the stimulus—it can be conceivable to disregard the words' semantics. In such cases, it is also possible to test the plain existence of words and to continue with clustering (see next subsection). Nevertheless, in most cases considering the semantics is worth the effort, and in the empirical part, we investigate whether the usage of plain words can outperform the usage of semantics at all.

A common way to model semantics is through *vector space models*. These are hyperdimensional spaces in which each term (e.g., *fantasy*) is represented by a vector. Similar vectors (terms) are semantically similar (e.g., *fantasy* and *imagination*), defined as two vectors having a small angle, or more precisely as a high cosine of two vectors. Such semantic spaces can be computed by, for instance, *Latent Semantic Analysis* (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) or *Explicit Semantic Analysis* (ESA; Gabrilovich & Markovitch, 2007). Both start with a co-occurrence matrix, called ''term × document-matrix,'' carrying frequencies of how often which term occurs in which context (i.e., document, paragraph, or sentence). This matrix is weighted based on the idea of relevance feedback (Salton & Buckley, 1990), giving more weight to terms occurring in rather few, specific contexts over those occurring diffusely across many contexts. This way, words that adhere to a specific semantic context—which typically is rather true for autosemantic opposed to function words—become important carriers of this context's semantic information. Typically, relevance feedback is applied using the log-Entropy function (Dumais, 1991). Creating a semantic space by means of ESA includes several steps (for details, see Gabrilovich & Markovitch, 2009) to finally gain a vector model describing the semantic space. The ESA is based on a manually structured text corpus, that is, documents (such as articles in Wikipedia) are included as intact entities. Within this natural structure (e.g., composed of articles) each element corresponds to one dimension in space, often leading to a very high number of dimensions. For instance, a term such as *fantasy* is represented by a vector comprising over one million (total number of articles) weights, each indicating the relatedness between an article and the term. The second approach to create a semantic space is LSA, which performs a singular value decomposition on the weighted term × document-matrix. Then, as text corpora are samples of language containing noise, and as contexts are assumed to be correlated, the decomposed matrix is reduced to less—often 300—dimensions, comparable with factor analysis.[2] For example, a term such as *fantasy* is represented by a vector comprising 300 values, each indicating the relatedness between a latent semantic concept and the term. For details of LSA, see Landauer (2011) and Martin and Berry (2011).

An important matter in LSA is domain specificity of text corpora. When taking Wikipedia as a basis, one can make use of its internal, explicitly connected structure. One way to gain a domain-specific corpus is to start at an article that is strongly related to the item's topic (e.g., *Alice's Adventures in Wonderland*) and crawl the incoming and outgoing links, also called ''children,'' to related articles (e.g.,

Fantasy). Dependent on the desired corpus size and the initial article's connectedness, it is reasonable to do this within a degree of connectedness of 1 to 3 (e.g., 2 meaning to take articles' children as well as children's children into the corpus). This procedure can easily be adapted. For example, more than one starting point can be chosen to gather bigger corpora, or filters can be applied as to which kind of articles must not be included (e.g., articles only dealing with a specific date; for more such considerations, see Gabrilovich & Markovitch, 2009).

Having provided the computer with semantic knowledge, the semantics of the responses can be extracted by computing the centroid vector of all tokens in the preprocessed response. At this stage, the example response comprises five tokens (cf. Part B in Figure 1). Each is represented by, assuming LSA was used, a 300-dimensional vector. Finally, the five vectors' centroid is computed—that is one 300-dimensional vector constituting the example response's total, or average, semantics. In all further analyses this centroid vector is the response's representation.

Model Building via Clustering

Represented by their semantic centroid vector, all responses given in the data are spread across the semantic space. If an appropriate semantic space has been built, the responses now typically form groups (cf. Part C in Figure 1). This means that some responses are more semantically similar to each other than to others—to put it more precisely, responses in one group contain semantically similar words. This perspective on the data can easily be applied by conducting a cluster analysis. In case the data are meant to be explored only, all responses are taken into account for model building. Otherwise, if some sort of supervised learning takes place—which is going to be explained in the next subsection—only a subset of responses (training data) is considered for this model building step.

Here, an agglomerative hierarchical cluster analysis is used that comprises three steps.[3] First, the distance between every pair of responses is computed. Second, every response is regarded as being its own cluster at the beginning. An agglomeration method iteratively decides which two clusters are minimal distant from each other and are to be merged to one cluster in the next iteration step. The agglomeration stops when all responses are assigned to one single cluster. Third, the researcher needs to decide which number of clusters is the best solution. This is an empirical matter and can be determined by plotting the distances (''rest''-component) of clusters for each solution and applying the elbow criterion. The respective number of clusters constitutes the desired solution at which the rest-component decreases significantly higher, relative to solutions with more clusters. See Rasch, Kubinger, and Yanagida (2011) for a concise description how this is done using R or SPSS.

A cluster analysis has several adjustable parameters. One important decision is how to define similarity, which can either be done by Euclidean distance or—worthwhile although computationally more expensive—cosine. The latter means to divide the vectors' dot product (Martin & Berry, 2011) by the product of their vector

lengths to apply an L2 norm, which is done to neutralize different document lengths (Gabrilovich & Markovitch, 2009). Cosine's advantage over Euclidean distance is to take only the vector directions into account—and not their lengths—because in the semantic space a vector's length mainly depends on the term's frequency, and as a term's frequency does not carry semantic information, the distance metric should incorporate only the vector's direction. Clustering algorithms base on *dis*similarity, and hence, the cosine's inverse, called *arccosine*, should be used. Besides choosing a measure of dissimilarity, the researcher needs to decide on the number of clusters as well as on an agglomeration method, such as Ward's method, which merges those two clusters leading to the smallest increase of variance within clusters (Ward, 1963). Nevertheless, all these parameters' optima are susceptible to the data's nature and, therefore, can be adapted with respect to theoretical or empirical knowledge about the respective item.

Up to this point, it is possible to work with text response data only. This is reasonable for data exploration and is called unsupervised learning; no further criterion is available (e.g., a variable indicating whether a response is *correct* or *incorrect*) that would allow model optimization with regard to this criterion. For example, the researcher can choose the number of clusters in an unsupervised manner as described above. But if the overall aim is to do some kind of supervised learning (e.g., scoring responses), it is reasonable to vary the number of clusters and choose the best performing one in terms of human–computer agreement. The following subsection about supervised learning describes how an interpretation such as *correct* or *incorrect* can be assigned to every cluster.

Assigning Codes to Neutral Categories via Supervised Machine Learning

Where the overall goal is AC, meaning that one of a fixed range of values—called *class*—needs to be assigned to a text response (e.g., *correct*), an external criterion is used by which to learn the relationship between the semantic vectors and the intended code. This kind of procedure is called supervised learning. Such an external criterion can be judgments by trained human coders, also called classification. Ideally, the whole data set is classified.

Supervised machine learning procedures are separated into two phases: *training* and *test*. In each phase only a subset of data is used, while the rest is put aside to control overfitting. Overfitting is given if a model is optimized too thoroughly, because it then performs very well on the data it was trained on but terribly poor on unseen data; while to apply models on unseen data usually is the purpose of building them. Hence, the data is split into 10 pieces, called *folds*. Nine folds are used together at a time to train the model while its performance is tested on the remaining 10th fold afterward. This is repeated 10 times so that every fold serves as test data once. Yet two things need to be considered. First, the data's overall class distribution should be represented in each fold; this is called *stratification*. Second, because still then chance highly impacts performance while choosing the folds, the whole procedure is again repeated

10 times, resulting in 100 performance estimates. This is called *stratified, repeated 10-fold cross-validation*; details can be found at Witten, Frank, and Hall (2011), and for comparison with other methods, see Borra and Di Ciaccio (2010).

In the training phase, clusters are built as described in the previous subsection. Then, to each cluster one code is assigned to (cf. Part D in Figure 1), determined by the within-cluster class distribution. For example, while scoring, a cluster might mainly comprise responses labelled as *correct*; hence, the cluster code is: *correct*. The decision which code is assigned to a cluster is based on the highest conditional probability using Bayes theorem,

$$P(c_i|g_j)_{k,j} = \frac{P(g_j|c_i)^* P(c_i)}{P(g_j)},$$

representing the probability that response k, belonging to the jth cluster g, has been assigned to the ith class c by the external criterion (e.g., human coder). For example, the probability of a response that is known to be member of Cluster 13, which is true for $\hat{P}(g_j) = 25\%$ of all responses, that it had been assigned to the class *correct* by the human coder, which is true for $\hat{P}(c_i) = 76\%$ of all responses, equals 96% if $\hat{P}(g_j|c_i) = 31\%$ of all correct responses belong to Cluster 13.

In the test phase, the unseen responses are classified by using the cluster model and the cluster codes. At first, the highest similarity, for instance, in terms of cosine, between the unseen response and a cluster centroid determines to which cluster the response is assigned to. Then, this cluster's code is assumed to be the response's code (cf. Part E in Figure 1).

At the end of the day, the automatic classifier needs to be evaluated. Its performance is often called *accuracy*. The simplest accuracy coefficient is percentage of agreement between computer and human. Another important one is kappa, which corrects for skewed class distribution.

Method

Measures of Coder Agreement

The Results section reports percentages of human–computer agreement ($\%_{h:c}$) as well as Cohen's kappa ($\kappa_{h:c}$) for data from PISA 2012. These statistics were computed by averaging the results from repeated 10-fold cross-validation. The kappa coefficient was transformed prior to averaging accordingly to Fisher (1915, 1921) and retransformed for reporting, on which the definition of Fleiss (1981) is applied: Values of $.40 \leq \kappa \leq .75$ constitute *fair to good*, lower values *poor*, and higher ones *excellent agreement beyond chance*. Moreover, a corrected coefficient of percentage of agreement,

$$\lambda_{h:c} = \frac{\%_{h:c_i} - \%_{h:c_1}}{100 - \%_{h:c_1}},$$

is reported, giving the proportion of the actual human–computer agreement's increase to the highest attainable increase; therefore, the percentage of agreement for the model with only one cluster ($\%_{h:c_1}$) is subtracted from the percentage of agreement for the final model with i clusters ($\%_{h:c_i}$), and this difference is divided by 100 minus the percentage of agreement for the model with one cluster, constituting the highest attainable increase. Unlike Kappa, this coefficient does not only correct for agreement by chance based on the class distribution but also for the trivial coding of empty responses. It further indicates how clustering affects performance in the data. Finally, Fleiss' kappa for multiple raters is used to report the system's internal reliability across repetitions ($\kappa_{c:c}$).

Materials

Items in PISA typically comprise a stimulus and a question referring to it. Responses used for analysis stem from 10 dichotomous items; eight items assessing reading literacy as well as one assessing scientific and mathematical literacy each. That is, there are items of three different constructs and specific coding guides for human coders for every item guiding them to code responses as either *correct* or *incorrect*. Although only items with dichotomous coding were used in this study, constituting the vast majority in PISA, the chosen approach to AC is also applicable to polytomously coded items. Because of repeated measurements, partly using the same items across cycles, the item contents are confidential and cannot be described here. The 10 items are listed in Table 1, each given a name for better traceability in the following. Prior to item selection, a theoretical framework had been developed as to which item characteristics might potentially influence AC performance using the proposed methods. According to this scheme, the selected items were intended to vary heterogeneously in their characteristics to test the AC method's scope.

As presented in Table 1, the items ranged from difficult (10%) to easy (83%) with the majority being medium difficult and a slightly skewed distribution toward easiness. For the reading items, the assessed aspect according to the PISA framework (OECD, 2013) mainly varied between the two of three aspects *Integrate and Interpret* and *Reflect and Evaluate* that both typically evoke more complex answers in linguistic terms than the third aspect does. The third aspect, *Access and Retrieve*, was only represented by one item and is assigned to items asking the test taker to find the relevant information given in the stimulus and repeat it, resulting in low language diversity in the responses.

Analyses

According to the first research question, we report the measures of coder agreement introduced above for all items to show the AC performance. For each item, we varied the number of clusters from 1 to 1,000 and report the best performing solution using cosine and Ward's method for clustering. We cross-validated all measures by

Table 1. Item Characteristics.

Item	Domain[a]	Aspect[b]	Correct	*n*	Words[c]
1. Explain protagonist's feeling	read	B	83%	4,152	12.3 (4.6)
2. Evaluate statement	read	C	43%	4,234	15.6 (9.0)
3. Interpret the author's intention	read	B	10%	4,234	12.5 (6.3)
4. List recall	read	A	59%	4,223	5.6 (3.0)
5. Evaluate stylistic element	read	C	56%	4,234	14.7 (6.2)
6. Verbal production	read	B	80%	4,152	12.4 (6.9)
7. Select and judge	read	C	68%	4,152	13.6 (7.0)
8. Explain story element	read	B	69%	4,223	14.4 (5.5)
9. Math	math	M	35%	4,205	14.0 (6.8)
10. Science	scie	S	58%	4,181	11.1 (5.2)
Total			56%	41,990	12.6 (6.1)

[a]One of *read*ing, *math*ematics, *scie*nce. [b]According to PISA framework (OECD, 2013), A = Access & Retrieve; B = Integrate & Interpret; C = Reflect & Evaluate; M = Uncertainty & Data; S = Explain Phenomena Scientifically. [c]Word count in nonempty responses on average (with *SD*).

stratified, repeated (10 times) 10-fold cross-validation, which is also true for the analyses described next.

Following the second research question, the relationship between performance and different methods or their configurations were examined. Therefore, seven analyses were carried out and are reported for single representative items. In Analysis I, the need for semantic space building opposed to the use of plain words is demonstrated. Analysis II illuminates the impact of spelling correction on AC performance by comparing uncorrected, automatically corrected, and manually corrected responses. The latter was possible by means of the transcription process (cf. next subsection). Semantic space building was varied in Analysis III using ESA or LSA, in Analysis IV by using different text corpora, and in Analysis V by varying the number of LSA dimensions. In Analysis VI, various common distance metrics were applied (*cosine*, *Euclidean*, *Chebyshev*, *Manhattan*, *Canberra*), and in Analysis VII, we studied the impact of different agglomeration methods (*Ward's*[4] and *McQuitty's method*; *Single*, *Complete*, *Average*, *Median*, and *Centroid Linkage*). For each variation, all other parameters and methods were set constant to those from the analyses for the first research question despite the number of clusters that was set with respect to the performance optimum. Percentages of agreement served as dependent variable in these experiments. The runs within one experiment were compared by considering the difference of means and the dispersions, for which percentage of agreement is the optimal measure.

Finally, in the third research question, we went about the relationship between performance and sample size, investigating to which extent the methods were applicable to studies with less data by conducting a simulation. To reduce random errors, we designed the simulation similar to repeated 10-fold cross-validation, in that, we first split the data into 10 stratified parts (called metafolds in the following) and subsequently excluded one metafold at a time until only one was left. At each of these

reduction steps, all left metafolds were merged and an independent stratified, repeated (10 times) 10-fold cross-validation was run on them, disregarding the previous metafold assignments. For example, after the first reduction $n = 4{,}152 - 415 = 3{,}737$ responses contributed to the analysis, in the next step only $n = 3{,}737 - 415 = 3{,}322$ did. Next, when only 415 responses were left, the last remaining metafold was split into another 10 smaller metafolds to gain more fine-grained reduction steps within the last 10th; and again, we subsequently excluded these smaller metafolds until only one was left and ran independent repeated 10-fold cross-validations on the remaining data at each step. This procedure was repeated 10 times, each time shifting the order of metafold exclusion to reduce the impact of the metafold sampling itself. Figure 2 visualizes this design. The simulation used the data of Item 7, *Select and judge*, which turned out to be representative regarding the obtained results in various analyses and was automatically coded medium well—with that, being sensible to improvements and worsening caused by sample size variation. Again, all parameters and methods were set constant to those from the analyses for the first research question despite the number of clusters that was set with respect to the performance optimum.

Participants and Procedure

The responses we analyzed come from the German PISA 2012 sample. This includes a representative sample of 15-year-old students as well as a representative sample of ninth graders in Germany. A detailed sample description can be found at Prenzel, Sälzer, Klieme, and Köller (2013) and OECD (2014). Due to a booklet design, the numbers of test takers varied for each item ($4{,}152 \geq n \geq 4{,}234$; cf. Table 1).

In PISA 2012, *reading*, *maths*, and *science* were assessed paper based. Hence, booklets needed to be scanned and responses were transcribed by six persons. To reduce typos, an additional mechanism was employed. Misspelled terms were annotated with their correct spelling, which additionally served as upper bound of achievable improvement by spelling correction in Analysis II (cf. subsection *Relationship Between Performance and Alternation of Parameter Values or Methods*). Only for the additional typo reduction mechanism, the misspelled terms were replaced by their annotation and these manually corrected data were put into Microsoft Word 2013 and checked for left misspellings (that now could only stem from transcription typos) using spelling correction. In total, 0.2% of all words were revealed to contain a typo this way, which were corrected in the data. In the analyses, the original responses were used disregarding the manual correction but applying an automatic spelling correction on students' misspellings.

Software

The software programmed for the aforementioned procedure implements open software, libraries, and packages. First, a database storing the downloaded Wikipedia

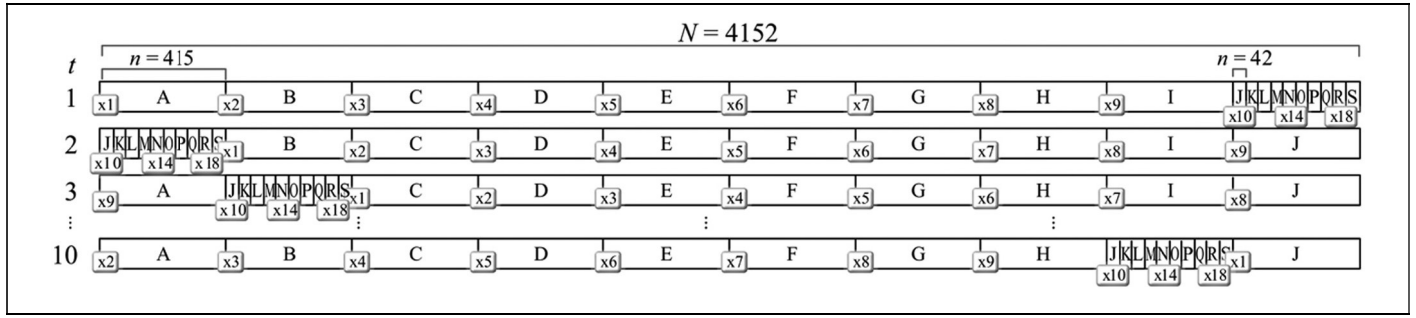

Figure 2. Sample size simulation design.

Note. The *x*-labels indicate which metafold was excluded in which order; for example, in the second repetition ($t = 2$), the data in Metafold E is included in the first three runs and excluded for all subsequent ones ($\times 4$) of this repetition. See the text for a more detailed description.

Table 2. Accuracy and Reliability.

Item	$\%_{h:c}$ [a]	$\lambda_{h:c}$ [b]	$\kappa_{h:c}$ [c]	$\kappa_{c:c}$ [d]
1. Explain protagonist's feeling	91.2[±0.2]	16.5	.533[.519; .547]	.814
2. Evaluate statement	86.5[±0.3]	63.8	.729[.723; .735]	.910
3. Interpret the author's intention	90.4[±0.2]	9.6	.458[.444; .472]	.738
4. List recall	98.4[±0.1]	84.2	.955[.952; .959]	.983
5. Evaluate stylistic element	76.2[±0.4]	12.4	.503[.495; .511]	.771
6. Verbal production	92.7[±0.2]	39.7	.704[.695; .713]	.925
7. Select and judge	86.6[±0.3]	39.2	.679[.672; .686]	.875
8. Explain story element	86.7[±0.3]	41.5	.676[.668; .683]	.886
9. Math	85.0[±0.3]	56.6	.669[.661; .676]	.822
10. Science	88.4[±0.3]	45.4	.761[.754; .767]	.918

Note. 95% confidence intervals given in brackets. [a]Percentage of human–computer agreement. [b]Relative accuracy increase by clustering; $\lambda_{h:c} = \frac{\%_{h:c_j} - \%_{h:c_1}}{100 - \%_{h:c_1}}$ (for details cf. Methods section). [c]Cohen's kappa for human–computer agreement. [d]Fleiss' kappa for within-computer reliability across repetitions.

dump (cf. subsection *Making Sense of Responses via NLP Techniques*) is built by using JWPL (Zesch, Müller, & Gurevych, 2008), which also comes with an application programming interface to access the corpus data. DKPro Similarity (Bär, Zesch, & Gurevych, 2013), which in turn primarily utilizes S-Space (Jurgens & Stevens, 2010), is used to build a vector space model. The response processing makes use of components offered in DKPro Core (Gurevych et al., 2007), which fit into the Apache UIMA Framework (Ferrucci & Lally, 2004). For stemming, Snowball (Porter, 2001) is used. For statistical matters, such as clustering, the software evokes R (R Core Team, 2014).

Results

Overall Performance and Its Relationship to Item Characteristics

As indicated in Table 2, the AC can lead to good up to excellent human–computer agreement; still, its performance varied by item remarkably. Across all items, the system reached high percentages of agreement from 76% to 98%, whereas the kappa values ranged from fair to good—Item 3, *Interpret the author's intention* ($\kappa_{h:c} = .458$)—up to excellent agreement beyond chance—Items 10, *Science* ($\kappa_{h:c} = .761$), and Item 4, *List recall* ($\kappa_{h:c} = .955$). Besides Item 3, Item 1, *Explain protagonist's feeling* ($\kappa_{h:c} = .533$), and Item 5, *Evaluate stylistic-element* ($\kappa_{h:c} = .503$), attained the poorest agreement beyond chance, and the five other items reached good agreement beyond chance ($.669 \leq \kappa_{h:c} \leq .729$). Whereas the kappa coefficient identifies the just named three items as performing poorest but still to a fair to good extent, the coefficient $\lambda_{h:c}$ underlines that their AC performance did not crucially improve by clustering ($9.6 \leq \lambda_{h:c} \leq 16.5$). A detailed analysis showed that the responses' correctness to items 3 and 5 could partially be affected by other linguistic

elements than pure semantics, while the reasons for items' rather poor AC were not obvious. On the other hand, especially for Item 4, *List recall*, the human–computer agreement was excellent and it increased remarkably by clustering ($\lambda_{h:c} = 84.2$) as it also did for Item 2, *Evaluate statement* ($\lambda_{h:c} = 63.8$), and Item 9, *Math* ($\lambda_{h:c} = 56.6$). These items share one important characteristic, namely, the evoked response's correctness is determined by the test taker expressing specific semantic concepts. In Item 4, *List recall*, this is done in a very simple way by just listing terms while in the other items the verbal constructions need to be more complex to be able to express the required semantics. Finally, the system's reliability mainly attained excellent agreement with $.771 \leq \kappa_{c:c} \leq .983$, except of the AC of the already discussed Item 3 ($\kappa_{c:c} = .738$).

Relationship Between Performance and Alternation of Parameter Values or Methods

Competing methods and parameter values were compared in seven analyses with regard to percentage of human–computer agreement ($\%_{h:c}$) whereas the sets of 100 percentage values from stratified, repeated 10-fold cross-validation served as sample for the respective experimental conditions. Most analyses were conducted using the data of Item 7, *Select and judge*, constituting a medium well working item for AC. This way, the measure of agreement could optimally be affected by the experimental variation, which would not have been the case for an item performing perfectly or not at all. Only in the first analysis, the data of another item were used.

To take a conservative approach, we based Analysis I on Item 4, *List recall*, which asked test takers to simply name four terms that are explicitly given in the stimulus text. In such a case, one could assume that testing for the bare presence of terms might be sufficient. The analysis showed that using a semantic space opposed to simply using the presence of words results in significantly better AC, $t(197) = 3.34, p < .001, d = 0.94 (\hat{=}\ 0.3\%)$—although with small effect due to the conservative approach. Whereas the semantic space needed 65 clusters to reach its optimal performance, the usage of raw words already required 500 clusters for this simple item to do so.

Analysis II studied the impact of spelling correction on performance comparing three conditions: no, automatic, and manual spelling correction. The latter constituted an upper bound for the possible effect of spelling correction on performance of how much an almost perfect spelling correction could improve the AC. Using an ANOVA, the analysis revealed that the automatic spelling correction neither improved performance significantly opposed to dispensing with it nor did the performance with manual spelling correction differ significantly, $F(297, 2) = 2.1, p = .130$. To further investigate if this finding can be generalized across items for the PISA data, cross-checks on other items were run. These showed the necessity of spelling correction depends on the item's nature—the same analysis with the data of Item 10, *Science*, yielded in similarly insignificant results, $F(297, 2) = 2.4, p = .089$, whereas

performance with and without spelling correction for Item 6, *Verbal production*, differed significantly with relevant effect size, $t(197) = 5.81, p < .001, d = 0.82 (\hat{=} 0.9\%)$.

In Analysis III, the semantic space building was varied between the two methods ESA and LSA. Results showed that ESA performs poorer in comparison with LSA, $t(196) = 3.32, p = .001, d = 0.47 (\hat{=} 0.8\%)$.

Analysis IV varied the text corpus on which the semantic space was based on. Five different text corpora were compared of which three closely dealt with the item contents and two did not. Additionally, they varied in size. Results showed that the selection of a text corpus affects AC-performance significantly, according to a one-way ANOVA, $F(4, 495) = 580.20, p < .001$, and a Newman Keuls procedure (SNK) revealed significant differences between every pair of corpora except for one pair that was equal in size but with one corpus having contents close to the item contents and the other one not, $SNK = -0.4\%, p = .094$. However, another corpus, which also closely dealt with the item contents but was bigger in size, gained a significantly higher performance level, $SNK = 0.9\%, p < .001$. The smallest corpus, not dealing with the item contents, performed poorest, $SNK = -7.7\%, p < .001$ (compared with the next better performing one). Disregarding this exceptionally poor performing corpus, the performances differed in a range of 1.8%. Thus, primarily, corpus size seems to matter, and content similarity only matters secondarily, as long as the corpus is big enough still to deal with the most important terms of the item contents.

Analysis V compared different numbers of extracted dimensions in LSA (50, 100, 200, 300, 400, 500, and 550). Results of an overall significant ANOVA, $F(6, 693) = 4.48, p < .001$, indicated that the number of dimensions only matter inasmuch as they should not decline below 100. The only significant differences stemmed from comparisons with 50 dimensions, $SNK = -0.5\%, p = .03$ (compared with next better performing one).

In Analysis VI, we studied the impact of different clustering distance metrics on performance. Results demonstrated significant differences between distance metrics in an ANOVA, which were also remarkable in effect size, $F(4, 495) = 121.30$, $p < .001$. Three groups crystallized, in that cosine, Euclidean distance, and Manhattan distance performed equally well while Chebyshev performed significantly poorer, $SNK = -2.2\%, p < .001$ (compared with Euclidean distance), and Canberra performed worse still, $SNK = -1.0\%, p < .001$ (compared with Chebyshev).

Finally in Analysis VII different agglomeration methods for clustering were compared. A similar pattern like in the previous analysis showed up as the agglomeration methods also formed three groups of performance levels with significant deviation in an ANOVA, $F(7, 792) = 125.60, p < .001$. No difference in performance could be found between Ward's method, Ward.D2, Complete Linkage, and McQuitty's method, $SNK = -0.5\%, p = .116$ (comparing the best with the poorest performer in this group). Another group was formed by Single, Median, and Centroid linkage that performed poorest. In between these two groups, Average linkage performed intermediately well; $SNK = -1.3\%, p < .001$ (compared with the poorest method of the

well performing group), $SNK = 2.1\%, p < .001$ (compared with the best method of the poorly performing group).

Relationship Between Performance and Sample Size

In the sample size simulation, sample sizes were varied to study the performance's dependency on sample size. As results showed, AC performance was affected by sample size both significantly and relevant in effect size—$F(18,18081) = 82.1, p < .001$ (ANOVA); $SNK = 6.3\%, p < .001$, comparing the best with the poorest performance ($n = 3,322$ and 84). Applying an SNK procedure, five homogeneous groups with respect to performance were built. It appeared that the proposed AC methods can be applied with three different sample size ranges (I) to (III) down to $n = 249$ that result in slightly different but acceptable performance levels (cf. Figure 3). As illustrated in the figure, performance started to decline crucially with a smaller sample size than 249. In the first range of $n = [1661, 4152]$ (I), AC performed best and without significant within-differences. Second, $n = 1,246$ (II) performed significantly poorer compared to the second poorest performance of the top performers, $SNK = -1.1\%, p = .009$. Third, opposed to this intermediately performing $n = 1,246$, a last range of sample sizes (III) with reduced but acceptable performance comprised the range of $n = [249, 831]$, $SNK = -0.9\%, p = .022$. Still further reducing the sample size to $n = 207$ (IV) again led to a significant decline, $SNK = 0.9\%, p = .015$, and finally, the range of poorest performance with $n = [42, 166]$ (V) lost another $SNK = 0.8\%$, $p = .040$. While for Range I clustering highly improved AC performance ($\lambda_{h:c} = 29.2$ for poorest performance), it still did so to a lesser extent for Range III ($\lambda_{h:c} = 13.7$ for poorest performance), but the performance increase by clustering was not significantly different from 0 at all when using data with $n \leq 125$.

Discussion

In terms of agreement with human coders, the implemented AC-system performed fair to good up to excellently on all 10 items. This suggests that AC is ready for a broader application in the field, bearing the potential for new opportunities in assessment. With regard to the accuracy increase by clustering as well as the poorest performances in terms of Kappa agreement (cf. Table 2), three items stuck out as being partially inadequate for the conducted AC: Item 3, *Interpret the author's intention*, Item 5, *Evaluate stylistic element*, and Item 1, *Explain protagonist's feeling*. While for the latter the reasons for AC failures were not obvious, responses on the first two of these items were partially coded improperly by AC because their correctness is crucially determined by other linguistic elements than just naming specific semantic concepts. In Item 3, *Interpret the author's intention*, the fact whether a response is correct or incorrect can be reliant on word order and relations, which is not detected by our approach to AC due to the use of the *bag of words*–paradigm, stemming, and stop word removal that often result in neglecting such information. This was also

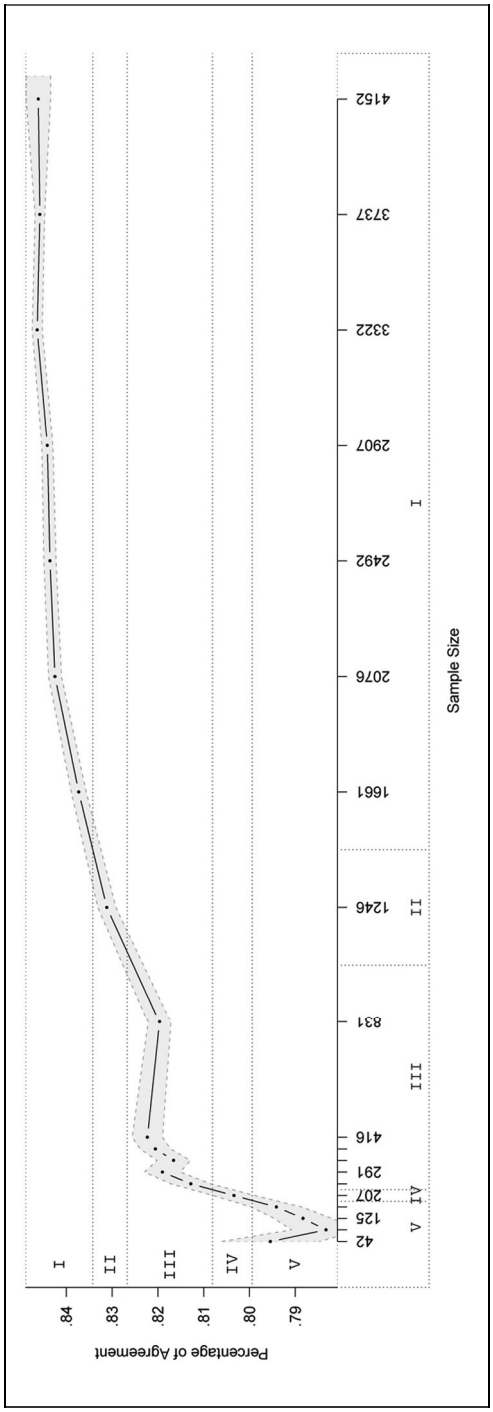

Figure 3. Results of the sample size simulation.
Note. The grey shaded area illustrates the means' 95% confidence intervals.

similarly true for Item 5, *Evaluate stylistic element*, in which responses' correctness is sometimes determined by synsemantic words (i.e., words that are not meaningful without other [autosemantic] words or a context). In this item, words such as ''how,''''what,'' and ''why,'' as well as which sentence subject they refer to carry important information. In contrast to that, most of the analyzed items were coded properly by AC mainly taking the responses' semantics into account. Not only Item 4, *List recall*, was automatically coded excellently, being an extreme case of demanding isolated semantic concepts but also other items, such as Item 10, *Science*, and Item 6, *Verbal production*, which require a complex response for solving it, were coded satisfactorily by AC. Beside the reading items and the just mentioned item assessing scientific literacy this was also true for Item 9, *Math*. Hence, this kind of AC is not limited to reading as a construct heavily based on natural language but it is also adequate to more formal domains when using verbalizable contents.

Furthermore, we showed that some methods and parameter values can be used interchangeably (e.g., number of LSA dimensions) whereas others significantly affect performance (e.g., LSA outperformed ESA, which we believe appears because spaces built via LSA are optimized for a single item and in turn are experts of the item content, whereas an ESA space is based on a universal corpus such as the whole Wikipedia and, thus, carries a lot of irrelevant information for one specific item so that it is less sensitive for the relevant information). Importantly, we also showed that the optimal performance is not attained by a fixed set of methods and parameter values but by finding the most appropriate ones for the data. As a rule of thumb, those described in the section *Proposed Collection of Methods for Automatic Coding* were always within the best performing ones in our analyses: LSA for semantic space building (with at least 100 but mostly used 300 dimensions), cosine as distance metric and Ward (.D2) as agglomeration method. Spelling correction was shown not to be necessary for all of the 10 PISA items, which of course again vastly varies by the test takers' average spelling skills and words the item content evokes. Experience shows that only a thoroughly developed and adapted spelling correction component should be applied; otherwise performance might even decrease by wrong ''corrections.''

While improving with more data in general, AC can also be applied in studies with smaller data than large-scale assessment data. There was a nonlinear relationship with an exponential performance increase in the range of typical sample sizes of social science studies, at about $n = 250$, and an asymptotic development toward optimal performance starting at about $n = 1{,}700$, which is about the typical sample size of studies for psychological instrument development. The item we used for this simulation was a medium complex one in terms of evoked language diversity and, thus, should be representative. For items with less or more diversity, this curve will be compressed or stretched, respectively.

The proposed methods should be seen as a baseline still to be improved in the future. The unsupervised clustering approach is very conservative because during clustering, the responses' classes are not considered at all. Yet this is advantageous as completely unsupervised applications of AC is conceivable (e.g., if textual data

are only meant to be categorized for analytical purposes; for more potential applications see the following list). In other contexts, so-called *Support Vector Machines* (e.g., see Theodoridis & Koutroumbas, 2009) constitute another machine learning method that will mostly outperform the proposed clustering approach. In the future, we intend to study the gain of semisupervised clustering in AC (e.g., see Witten et al., 2011).

The computational costs of the proposed AC are very low as soon as the model has been built once, and it should easily be possible to incorporate AC into existing procedures of large-scale assessments as well as into smaller studies. As the proposed methods appear to perform well, we want to encourage researchers as well as practitioners to use open-ended response format where it is appropriate and point out the following eight ways of using AC as examples that might improve existing and open doors to new assessment methods:

1. *Exploration*. Efficient exploration of textual data by automatic categorization is possible (e.g., for data sifting or to set up coding guidelines).
2. *Automatic Coding/Scoring*. By means of classified training data, codes can be assigned to each category (e.g., category of correct responses).
3. *More Differentiating Assessment*. Responses of different clusters may carry different information about the test taker. For instance, different clusters labeled all as correct may contain responses with different lines of reasoning. This might also lead to the possibility to synchronously assess a second construct enriching the principal construct (e.g., personality traits in a test which actually tests skills).
4. *Coding Guidelines Validity Check*. Multiple choice distractors are typically checked for their correlation to the construct. Now, for example, response clusters can be checked analogically for an unexpected high number of highly able test takers in a cluster that is interpreted as comprising incorrect responses. Such a case might reveal invalid coding guidelines.
5. *Control Loop for Human Coding*. If sticking to human coding, responses that are coded differently by human and computer can be revisited by a human (reducing errors from exhaustion, etc.).
6. *Reduced Human Coding*. Humans could only code, for example, 7 to 8 prototypical responses per cluster, which are defined as the most similar to the cluster centroid. In large-scale assessments, such as the presented PISA data, it then would not be necessary anymore to code more than 4,100 responses (per item!), but only, for instance, 30 clusters times 8 prototypical responses, resulting in 240 responses (an effort reduction of about 94%).
7. *Computer Adaptive Testing*. Adaptive testing requires to instantly code test taker responses. Therefore, usually closed response formats are used. Using AC expands Computer Adaptive Testing's scope to open-ended response format.

8. *Improving Assessment While Assessing*. Some test taker responses lack one single bit of information to exceed the threshold from incorrect to correct. But does this mean the test taker is not able to access this last bit? Since we are typically interested in the *latent* construct we indeed should distinguish between those test takers who *are* able to give us the missing bit and those who are *not*. It might turn out that a cluster contains such kind of responses; that means, there could be a cluster of responses that are almost correct but miss the last important bit of information. Automatic coding employed in computer-based assessment would then give the opportunity to probe the test taker for the missing bit. Of course, the implications of adding such a mechanism would need to be thoroughly studied. Nevertheless, it might be one more way to improve measures.

Limitations and Future Directions

We used German data only, but it should entail only a small to medium effort to apply the methods to further languages, since we utilized only baseline NLP methods that are available off-the-shelf for many different languages; as a showcase, being one implementation of one of many stemming algorithms, Snowball is available for six Germanic, six Italic, two Slavic, and two Uralic languages as well as for Turkish, Irish, Armenian, and Basque (cf. Porter, 2001). This is very interesting for international (large-scale) studies as using one coding system for multiple languages probably would strongly increase coding consistency between test languages, although the feasibility for all participating test languages would need to be examined.

Moreover, we only analyzed dichotomously coded items, whereas the used AC is also applicable to polytomous ones. For polytomously coded items it will be interesting to investigate the relationship between AC performance and sample size again with respect to the number of coding levels. For PISA 2012, only paper-based data were available that needed to be transcribed and, thus, is a possible source of noise due to typos. Also, computer-based assessed responses typically differ from paper-based assessed ones on the language level. We will soon be able to study the differences to computer-based data with the upcoming PISA cycle 2015. Furthermore, subgroup analyses like those of Bridgeman, Trapani, and Attali (2012) are needed to find whether this kind of AC performs equally for different test taker groups (native and nonnative speakers, high and low performers, different countries, etc.). Finally, it is worthwhile to study the possibility of using other external criteria for learning than human codings. Especially in the context of semisupervised learning it might be possible to use existing coding guidelines as learning criterion.

The newly implemented software used for analyses still lacks an interface that will be developed shortly in the context of an associated research group. As soon as ready, the software will be made freely and openly available and will enable researchers and practitioners to conduct AC on their data.[5] The software will not only contain the methods described here but also a more sophisticated NLP method called *Textual*

Entailment (cf. Androutsopoulos & Malakasiotis, 2010), which overcomes shortcomings of the *bag of words*–approach by performing inferences.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Wikipedia dumps are available at http://dumps.wikimedia.org/
2. Building an analogy to factor analysis, the extracted semantic concepts would correspond to the factors in factor analysis, contexts (documents) would represent variables, and terms would stand for persons. The resulting matrix could be interpreted accordingly in that a single value of a term's vector would be the variable's factor loading, carrying the information of how semantically important the term (variable) reveals to be for the semantic concept (factor).
3. Cluster analyses comprise a large method family with several adaptations. Only one commonly used is described here.
4. As implemented in R's hclust-function, the analysis distinguishes Ward.D2, which squares cluster distances before updating, and Ward.D, which does not.
5. If interested, you can check the corresponding author's website (http://zib.education/en/pisa/mitarbeiter-pisa/fabian-zehner.html) for reference to the upcoming software's website.

References

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research, 38*, 135-187. doi:10.1613/jair.2985

Bär, D., Zesch, T., & Gurevych, I. (2013). DKPro Similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 121-126). Sofia, Bulgaria: Association for Computational Linguistics.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9. doi:10.1111/j.1745-3992.2012.00238.x

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats: It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395. doi:10.1177/014662168701100404

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis, 54*, 2976-2989. doi:10.1016/j.csda.2010.03.004

Bridgeman, B. (1991). A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examinations General Test. *ETS Research Report Series*, 1991(*2*), 1-25. doi:10.1002/j.2333-8504.1991.tb01402.x

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. doi:10.1080/08957347.2012.635502

Burrows, S., Gurevych, I., & Stein, B. (2014). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60-117.

Carlson, S. S., & Ward, W. C. (1988). *A new look at formulating hypotheses items*. Princeton, NJ: Educational Testing Service. doi:10.1002/j.2330-8516.1988.tb00268.x

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO; 2-9

Dennett, D. C. (1991). *Consciousness explained* (1st ed.). Boston, MA: Little, Brown.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers, 23*, 229-236. doi:10.3758/BF03203370

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, 10*, 327-348. doi:10.1017/S1351324904003523

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10*, 507-521. doi:10.1093/biomet/10.4.507

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3-32.

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss, B. Levin, & M. C. Paik (Eds.), *Statistical methods for rates and proportions* (pp. 598-626). New York, NY: John Wiley.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (pp. 1606-1611).

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research, 34*, 443-498.

Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., & Zesch, T. (2007). Darmstadt Knowledge Processing Repository based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*.

Jurgens, D., & Stevens, K. (2010). The s-space package: An open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 30-35). Uppsala, Sweden: Association for Computational Linguistics.

Landauer, T. K. (2011). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 3-34). Mahwah, NJ: Erlbaum.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2011). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Martin, D. I., & Berry, M. W. (2011). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35-55). Mahwah, NJ: Erlbaum.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., & McNamara, D. S. (2011). Assessing and improving comprehension with latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 207-225). Mahwah, NJ: Erlbaum.

Organisation for Economic Cooperation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. doi:10.1787/9789264190511-en.

Organisation for Economic Cooperation and Development. (2014). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science* (Vol. *1*, Rev. ed.). doi:10.1787/9789264201118-en.

Porter, M. (2001). *Snowball: A language for stemming algorithms*. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster, Germany: Waxmann.

R Core Team. (2014). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011). *Statistics in psychology using R and SPSS*. Chichester, England: John Wiley. doi:10.1002/9781119979630

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*, 354-379.

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. doi:10.1191/0265532206lt337oa

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*, 288-297. doi:10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8>3.0.CO;2-H

Szarvas, G., Zesch, T., & Gurevych, I. (2011). Combining heterogeneous knowledge resources for improved distributional semantic models. *Computational Linguistics and Intelligent Text Processing, 6608*, 289-303. doi:10.1007/978-3-642-19400-9\textunderscore23

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). Burlington, MA: Academic Press. doi:10.1016/B978-1-59749-272-0.50003-7

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *American Statistical Association Journal, 58*, 236-244. doi:10.1080/01621459.1963.10500845

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann. doi:10.1016/B978-0-12-374856-0.00023-7

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 1646-1652).

Using and Improving Coding Guides For and By Automatic Coding of PISA Short Text Responses

Fabian Zehner[1,3]
[1]Technische Universität München
TUM School of Education
Arcisstr. 21, 80333 Munich, Germany
Email: fabian.zehner@tum.de

Frank Goldhammer[2,3]
[2]German Institute for
International Educational Research
Frankfurt am Main, Germany

Christine Sälzer[1,3]
[3]Centre for International
Student Assessment (ZIB) e.V.
Munich, Frankfurt am Main,
Kiel, Germany

***Abstract*—We propose and empirically evaluate a theoretical framework of how to use coding guides for automatic coding (scoring) and how, in turn, automatic coding can enhance the use of coding guides. We adopted a recently described baseline approach to automatically classify responses. Well-established coding guides from PISA, comprising reference responses, and its German sample from 2012 were used for evaluation. Ten items with 41,990 responses at total were analyzed. Results showed that (1) responses close to the cluster centroid constitute prototypes, (2) automatic coding can improve coding guides, while (3) the proposed procedure leads to unreliable accuracy for small numbers of clusters but promising agreement to human coding for higher numbers. Further analyses are still to be done to find the optimal balance of the implied coding effort and model accuracy.**

***Index Terms*—Automatic Coding, Automatic Scoring, Clustering, Coding Guides, Reference Texts**

I. Introduction

Short text responses play a crucial part in educational assessment. Having evolved from the field of automatic essay scoring [1], technologies for automatically evaluating short text responses have made vital progress in the last two decades. In this study, we adapt a recently described baseline approach [2] by reducing human involvement for model training. To satisfy requirements of machine learning procedures, most systems rely on relatively large amounts of manually coded training sets (often referred to as annotated data). These are not only expensive but might also contain incorrect codes, mainly due to the required mass. Hence recently, different research groups have strived to find appropriate procedures to train models with less but most informative data (cf. [3], [4], [5], [6]).

The reader will notice we use the psychometric term *coding* instead of the common *scoring*, *grading*, or *marking*. To us, the latter are a special case of coding which means to assign an entity to a category, whereas scoring means to additionally order these categories. The scope of automatic systems should not be limited to scoring only, although it is an important field of application. Analogously, with the term *coding guides* we refer to documents often used in social science studies that specify which class a response should be assigned to.

We argue that, at least, established assessments such as the *Programme for International Student Assessment* (PISA) already use documents for their manual coding that can be used to start training automatic systems. The so-called *coding guides* comprise item-specific *reference responses*: exemplary responses that are intended to constitute prototypes for their respective codes (i.e., full, partial, or no credit). Vice versa, experience teaches that empirical data will always force coding guide writers to update the coding guides during coding by adding new reference responses for two reasons—first, when a response type had not been considered and, second, to clarify the border between similar responses with different codes. Both can be supported by automatic systems identifying response types in the empirical data. This way, the coding guides help the automatic system to become trained, and the automatic system helps the coding guides to improve. Newly added reference responses then only need to be assigned to the intended code by the coding guide writers according to the assessed construct. Essentially, the number of new reference responses needs to be manageable, enabled by proper automatic identification of response types. Thus, the human is used for what the human is really good at—namely assigning a few single responses to codes—and the computer is used for what the computer is really good at—namely first sampling a few informative responses out of a mass and later on applying the learned rules to a mass. This procedure can also be used to simply systematically create coding guides from scratch.

The present study demonstrates (1) the use of coding guides as a source of training data compared to training with completely manually coded data and (2) how automatic coding via clustering improves coding guides. It also (3) examines the use of cluster centroids, in that, we show (a) how to sample prototypical responses and (b) that cluster centroids constitute representative prototypes for this. Finally (4), it shows the accuracy development in relation to the number of clusters. In this paper, accuracy is operationalized as human–computer agreement.

II. Procedures for Automatic Coding

This section first describes the basic approach taken to automatic coding and then, second, proposes adaptions. Third, the problem of sampling prototypes is elaborated. Finally, the employed system is compared to existing ones.

A. Semantic Clustering as Automatic Coding

The automatic coding system described in [2] was used. Briefly outlined, it first builds a vector space model using Latent Semantic Analysis [7] on the basis of a text corpus that is especially sampled for the respective item semantics (resulting in one corpus per item). Next, the empirical responses are preprocessed using common techniques such as stemming and spelling correction. The *bag of words*–paradigm is applied. Each response is then represented by the semantic centroid vector of all its tokens in the semantic space. In a next step, these response vectors are clustered by a hierarchical, agglomerative cluster analysis. The required number of extracted clusters can either be determined on the basis of manually assigned codes using stratified, repeated tenfold cross-validation (supervised) or with regard to the development of the clustering rest criterion (unsupervised). Once built, this cluster model serves to automatically code unseen responses by finding the most similar cluster centroid and assigning the cluster code. The cluster code is computed on the basis of manually assigned codes. This last step is where our proposed adaptation comes in.

B. The Use of Coding Guides for Automatic Coding

Requiring completely manually coded data to determine the cluster codes bears the disadvantages already described. Therefore, we propose the following to minimize the manual coding effort. First, the unsupervised variant is chosen to determine the appropriate number of clusters. These represent different response types. Second, the reference responses from the coding guides are processed in the same way as the empirical responses. Third, the reference responses are projected into the semantic space and each is assigned to the most similar cluster. Ideally, all clusters now have unambiguously coded reference responses assigned. In such a case, the final model is attained and can serve for automatic coding. But in most real-world cases, at least some of the conflicts described below are likely to appear. Once the conflicts have been solved, again a final model is attained offering the possibility for automatic coding.

The following conflicts are worth examining. They might either indicate difficulties for proper automatic coding or insufficient coding guides. For some item types, the approach to automatic coding presented here is simply not appropriate. But in principle, both can be improved by looking at two diagnostics: the frequency distribution of reference responses across clusters and the distribution of response distances to their cluster centroid within clusters. With regard to the former, the perfect coding guide would assign one reference response to one empirically evolving type. But for clusters lacking a reference response (Conflict I), a new reference response needs to be sampled from the empirical responses. Next, this new reference response is manually coded by the coding guide writers.

In other cases, the reference responses from coding guides concentrate on a few or even a single response type and, hence, a single cluster. This is not ideal but not necessarily a problem. Such a case only presents a conflict if these reference responses are intended to capture different response types. This might reveal an inappropriate semantic space and should raise awareness. Even worse, reference responses assigned to the same cluster could belong to different codes (Conflict II). This would reveal an insufficient semantic space or cluster model.

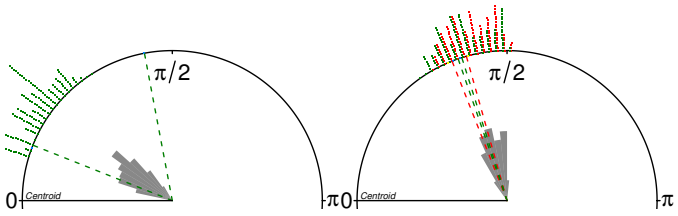

Figure 1. Exemplary Rose Diagrams; visualization of one accurate and one inaccurate item #6 cluster. The horizontal lines constitute the cluster centroids and the points represent responses within the clusters and their distance to the centroid (clockwise). Green points are manually coded as full credit, red ones as no credit. The dashed lines depict coding guide reference responses.

Moreover, the cluster-wise distributions of all response distances to their cluster centroid are informative diagnostics. Distances between two vectors x and y are operationalized as arccosine, $\Delta_{\vec{x},\vec{y}} = arccos(\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|*|\vec{y}|})$, and, hence, the within-cluster distance of a response $\vec{r}$ assigned to cluster i with centroid $\vec{c_i}$ is given by $\Delta_{\vec{c_i},\vec{r}}$. Half rose diagrams [8] can help to visualize the distribution and the reference response's relation to it (cf. Figure 1). These diagrams are histograms for circular data with the adaptation that the coefficient's range is $[0, \pi]$, thus, only a semicircle. When considering a reference response's position within this distribution, it is considered a prototype for this cluster if it is close to the centroid, indicated by a relatively small distance; this would show a good fit between the reference and the empirical responses. Other reference responses might represent response types not occurring in the empirical data at all if they are at the right tail's end of the distance distribution (Conflict III). These clusters then need a new reference response analogous to Conflict I if they have no prototypical reference responses assigned. The next subsection describes this sampling process.

C. Sampling Prototypes

One crucial step in the new procedure is to determine which empirical responses should be sampled as additional reference responses. When a regression is carried out on the sampled responses, optimal design algorithms, such as the Fedorov exchange algorithm [9], are highly effective [4]. That is because they select the most informative responses for the regression, those at the distribution's periphery. In the case of an unsupervised clustering has already been carried out, the researcher does not need to know analogously about the informative border responses. Instead, the researcher is in the fortunate situation of knowing about the data's underlying structure and, hence, knows about response types—basically, any response from each cluster could be sampled. Unfortunately, the number of clusters is not deterministic but ambiguous and is desired to be relatively small. This often

results in clusters comprising more than one response type in a strict sense. For clustering approaches, the overall goal is to identify one response that is most prototypical for the whole cluster. Often, responses close to their centroid are simply assumed to be the most prototypical (cf. [2], [3]). In [5], a list heuristic is used to find the response with the highest similarity *and* the most connections. In other terms, the heuristic seeks for the densest area within a cluster. This appears to be the most evidence-based reasoning. In the present study, we show empirical evidence of responses that should be sampled as prototypes for their cluster.

When striving to find a cluster's densest area, the most elegant way might be kernel density estimates. However, in most applications of automatic coding, these cannot be used because too many dimensions are employed resulting in a space that is too sparse. To examine dense areas in clusters, we here adopt an approximation similar to the list heuristic used in [5], making use of the fact that dense areas comprise many responses with relatively low pairwise distances. This is analogous to the definition of dense areas in kernel density estimation searching for the smallest area with the highest density (cf. [10]). Thus, we sort responses' pairwise distances to other responses within the cluster increasingly. These are then plotted per response. The responses with the most relatively low distances belong to the densest area. This approach guarantees to find local dense areas, but it does not guarantee that these responses do not just constitute a dense area within the cluster's periphery.

D. Related Work

The basic approach of the automatic coding employed in this study is to cluster vectors that represent the semantics of the responses [2]. The underlying concept is that the response type is the result of the respondent's cognitive processing and reaction to the question. Both the heterogeneity of respondents and the characteristics of the question determine evolving response types. Hence, responses can be grouped into question-specific types and most questions allow more than one type of correct responses (i.e., different lines of reasoning or different wordings).

A comparable system called *Powergrading* has recently been developed and partly attained remarkable performance [11]. It applies two-level clustering and learns a distance metric by a supervised method. Despite the powerful methodological approach taken, one reason for the high accuracy might be that the analyzed questions evoked a similarly low language diversity in the responses as one of the items analyzed in [2], which was also automatically coded excellently. *Powergrading* uses a fixed number of clusters, which is not plausible for typical assessment needs where different questions naturally tend to evoke different numbers of response types.

The unsupervised system described in [12] also makes use of similarities between responses based on Latent Semantic Analysis but does not group them by clustering. The authors propose to enrich the original response key with the words from the empirical responses that have the highest similarity to it. Whereas this procedure is comparable to our mechanism which adds empirical responses as new reference responses to the coding guide, it conceptually appears not to be optimal to us. Instead of allowing for different lines of reasoning, this might only add synonyms to the response key which already should have been considered similar by the vector space model.

A relatively large body of works dealing with the automatic grading of assignments recently has been evolving around massive open online courses (MOOCs). With some having tens of thousands of student assignments, they are a natural and important field of application for automatic coding. Which automatic systems are applicable highly depends on the kind of responses that are analyzed. For example, some systems were developed to automatically grade programming source texts. Some of these systems take comparable approaches to ours as similarities between students' and key responses are in the center of interest (e.g., [13], [14]). Some also apply clustering methods on these similarities (e.g., [15]). But the essence of how similarity is operationalized differs crucially. These systems need to deal with characteristics of formal language. When responses predominantly are natural language, as is the case for our needs, the systems mainly require information about the semantics of words (such as in [2], [11], [12], [16]).

Since the processing of natural language is relevant to various different domains—for example, dialog systems—a vast diversity of implementations with different adaptations is available in literature. For instance, one interesting concept of weighting single words in vectors of a vector space model are extrema, proposed in [17]. Yet, the authors found that extrema do not outperform the baseline in the domain of questions.

Furthermore, one important objective of our research is to implement automatic coding in the assessment area of social sciences. Thus, we decided to design our system in a way that is accessible and transparent in terms of a clear understanding of the underlying procedures. As the focus of social sciences is mainly content-related, it is worthwhile to be able to incorporate the outputs into the research's theory, such as response types represented by clusters. This is often hardly feasible when applying more powerful machine learning methods such as support vector machines or neural networks where the outputs are, besides the desired classifier, feature weights and neuron thresholds that are difficult to interpret. Hence, we selected clustering and Latent Semantic Analysis, which is conceptually related to factor analysis, since both these techniques are close to common methods in social sciences. Moreover, utilized unsupervised clustering is the most natural method of applying our main concept described above—questions evoke different response types that correlate with the respondent's cognitive reaction to the question.

III. Methods

This section briefly describes the materials used and data collected as well as the employment of the system and the analyses carried out.

Table I
ITEM CHARACTERISTICS

Item	Domain[a]	Aspect[b]	Correct	n	Words[c]
#1	read	B	83%	4,152	12.3 (4.6)
#2	read	C	43%	4,234	15.6 (9.0)
#3	read	B	10%	4,234	12.5 (6.3)
#4	read	A	59%	4,223	5.6 (3.0)
#5	read	C	56%	4,234	14.7 (6.2)
#6	read	B	80%	4,152	12.4 (6.9)
#7	read	C	68%	4,152	13.6 (7.0)
#8	read	B	69%	4,223	14.4 (5.5)
#9	math	M	35%	4,205	14.0 (6.8)
#10	scie	S	58%	4,181	11.1 (5.2)
Total			56%	41,990	12.6 (6.1)

[a] one of *read*ing, *math*ematics, *scie*nce
[b] A = Access & Retrieve, B = Integrate & Interpret, C = Reflect & Evaluate, M = Uncertainty & Data, S = Explain Phenomena Scientifically (according to PISA framework [20])
[c] average word count in nonempty responses (SD)

Table II
EXEMPLARY PISA ITEMS

Question	Full Credit Response	No Credit Response
One part of the article says, "A good sports shoe should meet four criteria." What are these criteria?	It must provide exterior protection, support the foot, provide the player with good stability and must absorb shocks.	Protect against knocks from the ball or feet. 2. Cope with unevenness in the ground. 3. Keep the foot warm and dry. 4. Support the foot.
Why does the article mention the death of Kiyoteru Okouchi?	To give the background to why people are so concerned about bullying in Japan.	It's just to grab your attention.

Note. Further released items and details such as the stimulus texts can be found in [21] (cf. pages 53 and 60 for the two given items).

A. Materials and Data

The data and coding guides used in the presented analyses stem from the German PISA 2012 sample. This includes a representative sample of 15-year-old students as well as a representative sample of ninth-graders in Germany. A detailed sample description can be found in [18] and [19]. Due to a booklet design, the numbers of test takers varied for each item ($4152 \geq n \geq 4234$). In PISA 2012, reading, maths, and science were assessed paper-based. That is why not all items but only ten transcribed ones, including eight reading, one maths, and one science item, were at hand. All items were coded dichotomously, that is, responses either got full or no credit. Table I presents some more item characteristics. More details on the assessed constructs and the transcription procedure can be found in [2]. Item and response contents cannot be reported due to the items' confidentiality. An example of two typical PISA items and respective responses are given in Table II; these were not part of the analysis. With regard to the kind of responses that are evoked, the first example is representative for item #4 and the second one for all others.

B. Employment of the Theoretical Framework

In the analysis presented here, we applied the described theoretical framework as follows. In order to decide on a cluster code, the system takes into account the coding guide reference responses assigned to this cluster. Then, those responses with a distance to the cluster centroid of at least the cluster-specific distribution's mean plus 1.6 standard deviations are omitted due to them being insufficiently prototypical. This way, the reference responses farther away from the centroid than 95 percent of all responses in the cluster are not used as they do not have empirical equivalents. If the remaining prototypical reference responses all have the same code, this code is used as the cluster code. If there are contradicting codes, however, the cluster is flagged for manual inspection and the code with the highest frequency within these prototypes is used as the cluster code. In case of a tie between codes (i.e., none of the codes reaches a majority), the reference responses are not used at all but a new empirical reference response has to be sampled as a prototype. This is also true for cases in which no reference response is assigned to the cluster at all.

In cases in which a new prototype had to be sampled, we selected the k responses closest to the respective cluster centroid. To analyze how the procedure works out with minimal coding effort, we set k to 1. In case the resulting codes differ, the semantic space needs to be analyzed manually. As the data we used in the analysis already had been completely manually coded by humans, we did not need to code the sampled responses but just used the manual code.

C. Analyses

All analyses used the default parameter setup suggested in [2] including stemming, spelling correction, 300 dimensions in the vector space model, cosine as the distance metric, and Ward's method for agglomeration. Analysis I investigates the required changes for the coding guides to cover all the empirical response types. The conflicts as described above that arose are depicted. In Analysis II, we show empirical evidence how prototypes should be sampled from clusters in case no reference response is available.

Analysis III follows two interests. First, it studies the system's accuracy when trained by the proposed procedure using coding guides (`cg`) compared to the accuracy when trained by the completely manually coded data (`man`). Cohen's kappa ($\kappa_{h:c}$) and the coefficient $\lambda_{h:c}$ introduced in [2] are reported for each condition. The latter is a corrected coefficient of percentage of agreement giving the proportion of the actual human–computer agreement's increase to the highest attainable increase: $\lambda_{h:c} = \frac{\%_{h:c_i} - \%_{h:c_1}}{100 - \%_{h:c_1}}$. For this, the percentage of agreement for the model with only one cluster ($\%_{h:c_1}$) is subtracted from the percentage of agreement for the final model with i clusters ($\%_{h:c_i}$). This difference is then divided by 100 minus the percentage of agreement for the model with one cluster, constituting the highest attainable increase. Taking a conservative approach, we overestimated the accuracy in the `man`-condition because all data were used for training and testing simultaneously and, hence, constitute a difficult benchmark to reach. This was necessary because the `cg`-

Table III
CODING GUIDE CONFLICTS

Item	Number of Clusters	Number of Ref. Resp.	Conflict I	Conflict II	Conflict III
#1	52	17	39 (75%)	1	7 (41%)
#2	48	21	34 (71%)	1	5 (24%)
#3	70	31	50 (71%)	1	7 (23%)
#4	7	17	1 (14%)	2	3 (18%)
#5	53	32	32 (60%)	2	7 (22%)
#6	53	21	39 (74%)	0	4 (19%)
#7	46	15	35 (76%)	0	3 (20%)
#8	31	17	22 (71%)	1	4 (24%)
#9	46	12	37 (80%)	1	1 (8%)
#10	55	15	44 (80%)	2	1 (7%)
Total	461	198	333 (72%)	11	42 (21%)

Note. Conflict I: clusters without coding guide reference response (percentage relative to number of clusters), II: cases in which reference responses with contradicting codes were assigned to the same cluster, III: reference responses without empirical correspondence (percentage relative to number of reference responses)

condition used the whole data for training and applying a cross-validation here might have introduced artificial effects.

Second, Analysis III examines the accuracy in relation to the number of clusters, comparable to the learning curve analysis in [3]. The number of clusters directly implies the coding effort. Particularly in this second part of the analysis, $\lambda_{h:c}$ is the optimal measure because it primarily indicates accuracy increase being corrected for a stable overall agreement by chance as well as empty responses and, thus, sensitively shows up accuracy changes. In comparison, $\kappa_{h:c}$ is an unstable measure in this context as its range depends crucially on marginal totals, which vary for each run. Therefore, it should be interpreted with awareness; nevertheless, we additionally report on this scale as most readers will be familiar with it.

IV. RESULTS

This section is structured by the different analyses described in the previous section.

A. Analysis I: Improvement of Coding Guides by Automatic Coding

The numbers of clusters were chosen with regard to the development of the rest-criterion. Table III depicts the number of clusters by item, the number of reference responses extracted from the coding guides, and how many conflicts occurred. In rare cases (II: 11 of 198), reference responses with different codes were assigned to the same cluster indicating an insufficient model. Often, this was due to generally improper automatic coding within the items; that is, when the system neglects a relevant linguistic information impacting a response's code, such as negation might. Another reason for these conflicts are the relatively small numbers of clusters as forcing clusters to join might result in the mixing of reference responses with different codes.

Partially, Conflict III cases can similarly stem from imperfect automatic coding; indeed, the three items #1, #3, and #5 performing poorest according to [2] show up with 22–41 percent of reference responses that are discarded because they are very remote from their cluster centroids. But generally across items, with the exception of the math and science item, there appears a relatively high tendency of 21 percent of reference responses that cannot be mapped to empirical response types. For an exemplary visualization, refer back to the left part of Figure 1 where there is one very prototypical reference response, represented by the dashed line on the left. Also, there is a reference response assigned to this cluster that is way apart from the cluster's distribution. Furthermore, in this figure, the empirical grounding for the procedure described in Section II-B, to omit reference responses if they are not prototypical, can be found. Although one would assume the distribution to be half of a very steep normal distribution around the centroid, the distributions very much behave like normal distributions with their mean shifted at the range of $[\frac{\pi}{4}, \frac{\pi}{2}]$. This is, amongst others, due to the vectors' hyperdimensionality and is additionally dependent on the number of responses in the cluster. Moreover, the distance distribution can be used, for example, as an indicator for the cluster's homogeneity or often even purity—obviously the distribution of the inaccurate cluster (mixing codes; cf. right part of Figure 1) is more shifted away from the centroid than the pure cluster (cf. left part of the figure).

High rates of clusters do not have reference responses assigned to them at all (I: 72% on average). Considering the discrepancy between the number of reference responses available in the coding guides and numbers of clusters extracted, this might not be surprising. Also, the automatic coding generally distinguishes more response types than humans do because its approximation of language comprehension is highly superficial. Nevertheless, different clusters exist due to concrete differences on the language level, which might influence human coders, so the values of up to 80 percent show a high potential for coding guide improvements.

B. Analysis II: Sampling Prototypes

For all items and clusters, the procedure described in Section II-C was conducted. Patterns of pairwise distances of responses were analyzed as shown in Figure 2 to identify dense regions within clusters. Two exemplary clusters are given, a bigger and a smaller one. Each line represents one response within one cluster of a specific item. As the increasingly ordered distances are plotted, those curves of responses correspond to dense areas that are relatively low as long as possible with regard to the x-axis. Such responses have many low distances to other responses and, thus, are member of a relatively dense area within the cluster. Additionally, the five responses that are closest to the cluster centroid stand out in black, dashed lines.

Obviously, the responses close to the cluster centroid are located in the relatively densest areas. This finding is exceptionally consistent across all items and clusters. The list heuristic, on the other hand, ensures to find the densest areas but not necessarily one that is prototypical for all other responses in the cluster. It is conceivable to find a small dense area at the cluster's periphery that is not representative for the rest of the

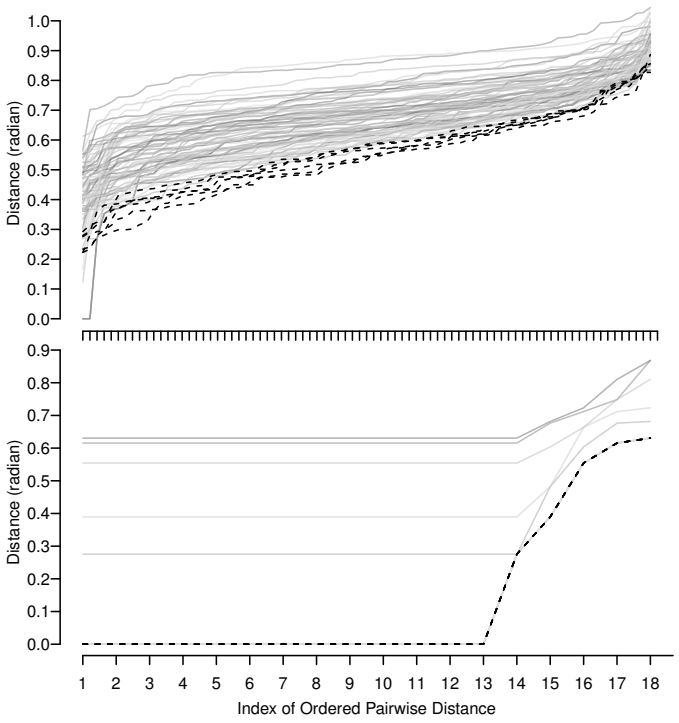

Figure 2. Identifying Prototypes. Two exemplary figures of increasingly ordered pairwise distances of responses within one cluster (item #10, clusters 3 [top] and 19 [bottom]). Each line constitutes one response, the black, dashed ones are the five responses closest to the cluster centroid.

cluster. Hence, in rare cases, such a list heuristic as used here and in [5] can be misleading; an example can be found at the upper part of Figure 2 where the lowest lines are two similar responses of which the following distances are relatively high. Therefore, we recommend to use the responses close to the centroid for sampling of prototypes. These constitute the optimal prototypes as they are always in dense areas and at the same time at the cluster's center guaranteeing to be the best representatives of all cluster members.

C. Analysis III: Comparison of `cg` and `man` and Learning Curve

The first part of this analysis focuses on the accuracy of the `cg`-condition compared to the `man`-condition. The results can be found in Figure 3 by comparing the top row (`man`) with the bottom row (`cg`). Basically, the system's accuracy drops remarkably by using the proposed approach, condition `cg`. The accuracy seems to randomly jump instead of steadily improving along increasing cluster numbers. We assume that our goal to minimize the simulated coding effort, using $k = 1$ for the prototype sampling, gave too much weight to single response codes. These could influence many other responses in cases of large clusters. This is supported by Figure 3 showing the curves' stabilization at $\geq$100 clusters. These larger numbers of clusters result in smaller clusters, and thus, in a reduced impact of single responses on other responses. The accuracy values of the two conditions in these higher ranges of numbers of clusters do not differ greatly but more analyses with more reliable, and thus more comparable, setups are necessary.

In the second part of the analysis, we concentrated on the relationship between accuracy and number of clusters. Again, the results can be found in Figure 3. Contrary to our intention, the results of the `man`-condition should be interpreted in the first place because the `cg`-condition's results seem partially imprecise. Nevertheless, the findings can also be mapped to the `cg`-condition—if we assume the worst case scenario, where no reference response was accessible to automatically code the empirically found response types, each cluster that is extracted entails more coding effort. Yet, every bit of information helps to build more reliable models. This trade-off is obvious in the results. Generally, it can be seen that the more clusters are extracted the higher the agreement will be. This is not surprising, particularly as the test data is identical to the training data, as explained previously. Still, this is not the crucial aspect but it is obvious that a lot of different response types can be found in about 4,200 responses. Item #4 here serves as a good showcase. It represents the least complex item type in PISA in which the test taker only needs to repeat information that is explicitly given in the stimulus, resulting in a very low language diversity in responses (cf. the item's average word count in Table I as an approximation). In this case, the test takers are asked to repeat a list of four terms. Although the automatic coding of this item already reaches a very good agreement with 7 clusters, it continues to show marginal improvements in the range of 280 clusters. Of course, the response types represented by these clusters only occur occasionally—the first cluster still carries 57 percent of nonempty responses. Yet, this case shows the language diversity that automatic coding needs to deal with in even such a simple setting. This underlines the importance of vector space models and their semantic concepts opposed to pure word-based processing.

Also interesting in Figure 3 is the steepness of the curves. A steep learning curve means that with few clusters a high gain in accuracy is attained. This can particularly be found for the items #4 and #2, which converge towards their optimum almost in the range of 10–20 clusters. Others show steady improvement, reaching their optima only at about 70–90 clusters. Still, there are items that are not properly coded and show a very low improvement.

V. Conclusion and Directions

The presented theoretical framework and empirical evaluation show that the established use of coding guides in assessments and the nascent field of automatic coding can benefit from each other. The automatic coding approach we took up from [2] is based on identifying response types that can be used as reference responses in coding guides. The first analysis showed a high potential of real coding guides employed in PISA to be replenished with further reference responses. These can be identified very efficiently by the operated automatic coding system without any manual coding

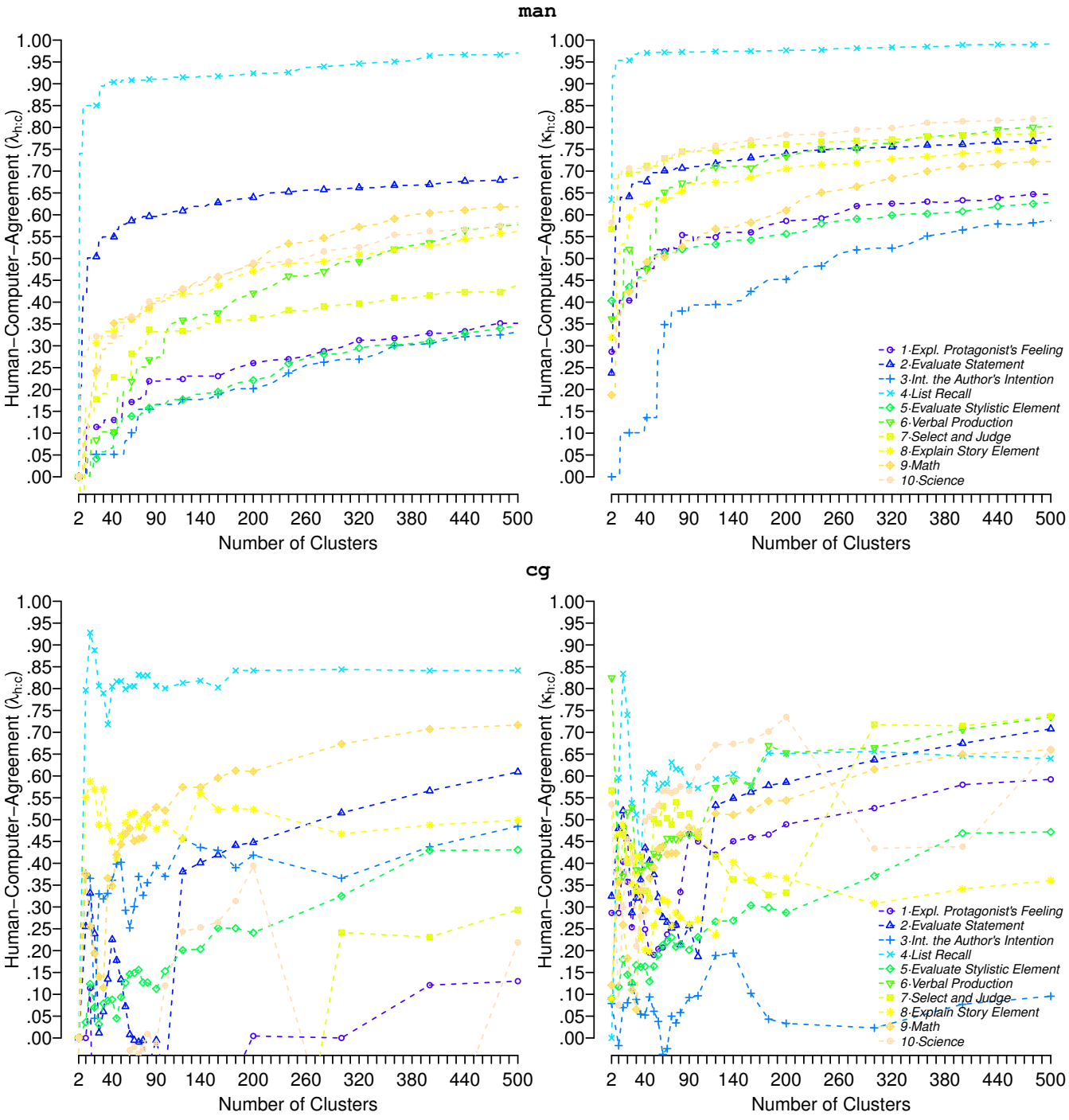

Figure 3. Accuracy by Number of Clusters and Condition. Human–computer agreement ($\lambda_{h:c}$ left, $\kappa_{h:c}$ right) for the `man`- (top) and `cg`-condition (bottom) plotted against the number of clusters being extracted for each item. The $\lambda_{h:c}$-coefficient indicates the relative accuracy increase compared to a solution with one cluster.

by humans, which not only costs time and money but is also prone to errors. The system puts only a baseline approach into practice leading to this advantage of unsupervised methods over most other, supervised automatic coding systems which require at least a minimum of such manual coding because they utilize powerful machine learning procedures such as support vector machines.

Conversely, the concept of coding guides, which are produced by the items' experts, most often the item developers themselves, is very promising to the field of automatic coding. Strongly reducing the manual coding effort allows to let experts decide on the coding of responses, or rather response types. The resulting training data would contain fewer errors due to exhaustion, inconsistency, or even misconceptions of trained coders (an elaborate, concise overview of rater cognition typically influencing manual coding can be found in [22]).

Despite the promising approach, the empirical accuracy of the automatic coding system showed unreliable variation using coding guides and sampling of new reference responses in the range up to 100 clusters. From this point, the system's performance becomes more accurate and reliable with not too much deviation from the original system. This evidence suggests that using only one newly sampled response as the prototype and as the decision on the code for the whole cluster ($k = 1$) leaves too much impact to chance. This is true, although we showed evidence that the sampled responses close to the cluster centroid are indeed most prototypical for their clusters. The combination of using only $k = 1$ responses for sampling to simulate minimal coding effort with only a few clusters, which in turn led to relatively large clusters, is likely to have produced the inaccurate performance. Hence, we suggest setting k at higher values such as 3 or 5. A first analysis showed the expected improvements but a more systematic analysis is necessary regarding how to optimally balance k and the numbers of clusters without overtaxing the manageable amount of manual coding effort. Nevertheless, the proposed approach appears promising when taking into account the relatively small loss in accuracy as opposed to the system that was trained with over 4,000 response codings in the range of ≥ 100 clusters. Even if no reference responses are usable for 100 empirical clusters, the effort to manually code 100 responses, which constitutes about 2% of the 4,200 codings that were needed otherwise, seem manageable for coding guide writers.

References

[1] E. B. Page, "The imminence of grading essays by computer," *Phi Delta Kappan*, vol. 48, pp. 238–243, 1966.

[2] F. Zehner, C. Sälzer, and F. Goldhammer, "Automatic coding of short text responses via clustering in educational assessment," *Educational and Psychological Measurement*, 2015. [Online]. Available: http://epm.sagepub.com/content/early/2015/06/06/0013164415590022

[3] T. Zesch, M. Heilman, and A. Cahill, "Reducing annotation efforts in supervised short answer scoring," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Ed., 2015, pp. 124–132.

[4] N. Dronen, P. W. Foltz, and K. Habermehl, "Effective sampling for large-scale automated writing evaluation systems," *arXiv preprint arXiv:1412.5659*, 2014.

[5] L. Ramachandran and P. Foltz, "Generating reference texts for short answer scoring using graph-based summarization," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Ed., 2015, pp. 207–212.

[6] J. Z. Sukkarieh and S. Stoyanchev, "Automating Model Building in c-rater," in *Proceedings of the 2009 Workshop on Applied Textual Inference*, 2009, pp. 61–69.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[8] S. R. Jammalamadaka and A. Sengupta, *Topics in circular statistics*, ser. Series on multivariate analysis. River Edge, N.J: World Scientific, 2001, vol. v. 5.

[9] V. V. Fedorov, *Theory of optimal experiments*, ser. Probability and mathematical statistics. New York: Academic Press, 1972.

[10] D. W. Scott, *Multivariate density estimation: Theory, practice, and visualization*, ser. A Wiley-Interscience publication. New York, NY: Wiley, 1992.

[11] S. Basu, C. Jacobs, and L. Vanderwende, "Powergrading: A clustering approach to amplify human effort for short answer grading," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 391–402, 2013.

[12] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 567–575.

[13] A. Nguyen, C. Piech, J. Huang, and L. Guibas, "Codewebs: scalable homework search for massive open online programming courses," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 491–502.

[14] S. Srikant and V. Aggarwal, "A system to grade computer programming skills using machine learning," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1887–1896.

[15] E. L. Glassman, R. Singh, and R. C. Miller, "Feature engineering for clustering student solutions," in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 171–172.

[16] S. Jing, "Automatic Grading of Short Answers for MOOC via Semi-supervised Document Clustering," in *Proceedings of the 8th International Conference on Educational Data Mining*, O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds., Madrid, 2015, pp. 554–555.

[17] G. Forgues, J. Pineau, J.-M. Larchevêque, and R. Tremblay, "Bootstrapping Dialog Systems with Word Embeddings," Montreal, 12.12.2014. [Online]. Available: http://www.cs.cmu.edu/~apparikh/nips2014ml-nlp/camera-ready/forgues_etal_mlnlp2014.pdf

[18] M. Prenzel, C. Sälzer, E. Klieme, and O. Köller, *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann, 2013.

[19] OECD, *PISA 2012 results: What students know and can do - Student performance in mathematics, reading and science*, volume 1, revised edition, february 2014 ed. Paris: OECD Publishing, 2014.

[20] ——, *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing, 2013.

[21] ——, "PISA Released items - reading," 2006. [Online]. Available: http://www.oecd.org/pisa/38709396.pdf

[22] I. I. Bejar, "Rater cognition: Implications for validity," *Educational Measurement: Issues and Practice*, vol. 31, no. 3, pp. 2–9, 2012.

Further Explaining the Reading Gender Gap: An Automatic Analysis of Student Text Responses in the PISA Assessment

Fabian Zehner[1,3], Frank Goldhammer[2,3], Christine Sälzer[1,3]

[1]Technische Universität München, [2]German Institute for International Educational Research, [3]Centre for International Student Assessment (ZIB) e.V.

Author Note

Fabian Zehner, TUM School of Education, Centre for International Student Assessment (ZIB) e.V., Technische Universität München; Christine Sälzer, TUM School of Education, Centre for International Student Assessment (ZIB) e.V., Technische Universität München; Frank Goldhammer, German Institute for International Educational Research, Centre for International Student Assessment (ZIB) e.V.

Correspondence concerning this article should be addressed to Fabian Zehner, TUM School of Education, Centre for International Student Assessment (ZIB) e.V., Arcisstr. 21, 80331 Munich, Germany. E-mail: fabian.zehner@tum.de

Abstract

The gender gap in reading literacy is repeatedly found in large-scale assessments. This study compared features in girls' and boys' responses in a reading test by applying natural language processing techniques. For this, a theoretical framework was compiled that allows mapping of features in the responses to the underlying cognitive components such as micro- and macropropositions in the situation model. In total, $n = 33,604$ responses from the German sample of the *Programme for International Student Assessment* (PISA) 2012 reading assessment have been automatically analyzed in order to find differences in the cognitive types of males and females. Results showed that girl types used more propositions (2.8 to 4.9) in the situation model, whether the response was correct or not. They integrated relatively more relevant propositions and more successfully adapted to the level of the question focus and category. That means, in answering questions which ask for explicit information from the stimulus text, the girl types appropriately used more micropropositions, and boy types tended to use more macropropositions—vice versa for questions asking for information that needed to be inferred. It appears that boy types struggle with retrieving and integrating propositions from the situation model while girl types liberally juggle these to formulate their responses.

Keywords: gender differences, reading literacy, automatic coding, raw responses

Further Explaining the Reading Gender Gap: An Automatic Analysis of Student Text Responses in the PISA Assessment

Reading literacy is characterized by remarkably consistent gender differences of relevant magnitudes (e.g., Mullis, Martin, Foy, & Drucker, 2012; NCES, 2015; OECD, 2013b). A student's raw response to a reading test question, such as *"The son wanted to help his fellow students,"* has not been accessible for studies until now but provides interesting insights about the respondent: Which information from the previously read text does the answer treat? Does the answer only repeat the text or does it add information? And is this information relevant to answer the question? Particularly in large-scale assessments, such linguistic features in responses could not be considered due to the mass of data. Therefore, this study focuses on two research matters. First, it attempts to add to the understanding of the gender gap in reading literacy by analyzing the natural language in students' short text responses. It identifies responses that are typical for either of the genders and contrasts their features. Second, it compiles a theoretical framework about how and which linguistic features in the responses can be mapped to the preceding cognitive processes. That is, we regard responses as the outcomes of reading-associated cognitive processes and depict which indicators in responses show differences in the underlying cognitions. For the automatic processing of the responses, a software with the required technology was adapted from Zehner, Sälzer, and Goldhammer (2015) and measures were derived from the framework.

The *Programme for International Student Assessment* (PISA; OECD, 2013b) assesses reading literacy in 15-year-old students along with mathematics and science. Its data provide an attractive opportunity to analyze the gender differences. Since its first round in 2000, PISA found notably large deviations between male and female adolescents in eighty countries and economies, to date. The present study used the raw text responses of German PISA 2012 students.

The gender gap is at the core of this study, and talking about the gap means talking about the comparison of means. Yet, a mean score is not representative of its entire distribution, and, like Stanat and Kunter (2002) observed, even with PISA's large gender difference, the subgroup distributions largely overlap (cf. Figure 1). Hence, there is actually no gap between the genders,

only between their means, which in turn is remarkably wide. That is why we mainly conducted analyses at the level of gender types in this study (i.e., responses particularly typical for boys or girls, respectively).

Theory

Related Work: The Reading Literacy Gender Gap

Particularly over the last twenty-six years, research has been producing a huge body of findings about gender differences in reading literacy. In PISA, girls consistently outperformed boys (OECD, 2002, 2004, 2007, 2010b, 2013b). That is the case across all countries—with only two non-significant exceptions in 2000 and one in 2003—and across all cycles, constituting 268 comparisons in total. With the scale's standard deviation of 100 points, the gender gap average in the OECD ranged from 31 points in 2000,[1] to 34 points in 2003, 38 points in 2006 and 2012, to 39 points in 2009. The countries' minimum was reached in 2000 by Peru with 7 points in favor of girls, and the maximum was reached by Jordan in 2012 with 75 points.

These numbers show an astonishingly stable figure that appears even more remarkable when considering that the effect was not always replicated in other studies (Elley, 1994; Hyde & Linn, 1988; Thorndike, 1973; White, 2007; Wolters, Denton, York, & Francis, 2014). Often, the effect sizes appear somewhat smaller, such as in the *Progress in International Reading Literacy Study* (PIRLS; international average: 16 points difference, scale's $SD = 100$; Mullis et al., 2012) or in the *National Assessment of Educational Progress* (National NAEP; $d = 0.06$ for 4th graders, $d = 0.14$ for 8th graders; NCES, 2015).[2] A stable gender gap from kindergarten through eighth grade with a growing gap for low performers was found by Robinson and Lubienski (2011, $0.10 \leq d \leq 0.24$). Sometimes, even larger effect sizes than those in PISA occur ($d = 0.60$; Gambell & Hunter, 1999). For adults, the *Programme for the International Assessment*

[1] Note that the OECD reported the average of all participating countries for 2000, which was 32 points, instead of the OECD average.

[2] Because NCES (2015) does not report effect sizes, we computed the NAEP effect sizes on base of the data given at NCES (2015), considering the sample size given at https://nces.ed.gov/nationsreportcard/reading/moreabout.aspx#students. Due to the lack of sampling weights, the values constitute only the sample effect sizes and cannot be directly compared to the other studies. Within the study, the gender differences in reading are very stable since its beginning in 1992.

of Adult Competencies (PIAAC) did not find significant gender differences in reading literacy for most participating countries (including Germany; OECD, 2013a).

All these variations across studies as well as the consistency within but not across PISA, PIRLS, and NAEP support the view of Lafontaine and Monseur (2009). They attributed the varying gender differences across studies to methodological decisions such as age- versus grade-based populations, and they emphasized the effect of whether the response format was open-ended or closed. In the light of the trend to include more open-ended questions in the last two decades, this adds to the analysis of Lietz (2006), who found an increase in the reading gender gap over time in large-scale assessments since 1992. Similar to Lafontaine and Monseur, Schwabe, McElvany, and Trendtel (2015) argued that part of the gender gap is found due to response formats and intrinsic motivation induced by the stimuli. The authors showed that female students outperformed equally skilled male students at open-ended items and that reading engagement moderated this effect, whereas only very low effect sizes were reported for these. Furthermore, male and female students can be differently engaged by specific stimuli, because they typically report different text types they favor to read for their enjoyment. Fewer girls reported reading newspapers and comics than boys do (-7% and -10%), while more girls reported reading fiction and magazines ($+19\%$ and $+14\%$; OECD, 2015). When reading strategies, along with reading engagement, are used as predictors, gender does not significantly account for any more variance in reading literacy (Artelt, Naumann, & Schneider, 2010).

In addition to the response format, the complexity of the task and of the text are sources of variance. In PISA 2000, the gender gap widened with increasingly demanding tasks (locating information vs. interpreting vs. reflecting: $d = 0.2$, 0.3, and 0.4; Stanat & Kunter, 2002). For non-continuous texts, boys almost closed the gap ($d = 0.1$), but for continuous texts, they showed particular difficulties ($d = 0.4$). The same figure is reported by Lafontaine and Monseur (2009) based on the same data. However, despite these substantial effect sizes, the subscales' differential effects vanished in PISA 2009 for the most part (Naumann, Artelt, Schneider, & Stanat, 2010; OECD, 2010b).

Because this study exclusively analyzes the German PISA 2012 data, it is relevant to state that the gender differences in Germany have been substantial across all cycles—always marginally stronger than the OECD average: 34 points in 2000 (OECD, 2002; Stanat & Kunter, 2002), 42 points in 2003 and 2006 (Drechsel & Artelt, 2007; OECD, 2004, 2007; Schaffner, Schiefele, Drechsel, & Artelt, 2004), 40 points in 2009 (Naumann et al., 2010; OECD, 2010b), and 44 points in 2012 (Hohn, Schiepe-Tiska, Sälzer, & Artelt, 2013; OECD, 2013b).

Sample PISA Stimulus and Question

For easier reading of the following subsections, a sample PISA stimulus and question is given in Figure 2. These will serve as examples for the processes described in the following. Generally, the stimulus text deals with a survey about bullying in schools. It ends with a note on the suicide of Kiyoteru Okouchi who had been bullied at school. The question asks the students why this suicide is referred to in the article. Correct responses "[relate] the bullying-suicide incident to public concern and / or the survey OR [refer] to the idea that the death was associated with extreme bullying" (p. 60), for example, "To give the background to why people are so concerned about bullying in Japan." (OECD, 2006, p. 60).

Theoretical Framework

This study endeavors to shed more light on the reasons for the observed gender differences. We, as researchers, observe these differences at the end of a long chain of events and conditions. (I) The students come with their individual abilities and states into the testing situation in a given setting. There, (II.a) they are exposed to a stimulus text and question (II.b) created by the researcher according to the construct framework. Both, (III.a) reading and comprehending the text as well as (III.b) finding an answer take place in the students' cognitions. Finally, (IV) most students formulate a response corresponding to the mental representation of their solution. In the end, (V.a) the researcher aggregates scores based on these responses, (V.b) observes group differences, and (V.c) interprets them as different degrees of reading literacy.

Whether they are the original source for the differences or only a mediator of external

influences, the difference would always be inherent to the students' cognitions (III). Thus, the cognitions constitute a reasonable level for interpreting where the differences come from. The cognitions' direct outcome is a response (IV) containing indicators for the preceding cognitions, confounded only by verbal production skills and test motivation. In the present study, we extrapolate specific features in the cognitive processes from the student responses. In survey and also in discourse research it is an established paradigm to infer from people's language to the preceding cognitive processes (cf. Heritage, 2005; te Molder & Potter, 2005; Tourangeau, Rips, & Rasinski, 2009; Wetherell, 2007). Similarly to (II.b) and (V.c), aligning instruments and the interpretation of their outcomes to the theoretical framework of the construct, the respective operationalizations need to be closely led by theoretical models, which are depicted in the following subsections. According to the *situation model* (Kintsch & van Dijk, 1978), we name the crucial cognitive processes taking place while the students process the text (III.a). Then, we frame the situation model's outcomes into the cognitive model *QUEST* (Graesser & Franklin, 1990) to trace how students achieve their solution (III.b). Further works by Graesser outline how the students finally craft this solution into their response (IV). The last theoretical subsection describes how we attempt to operationalize the outcomes in the responses in order to map them back to the cognitions. Only the terminology necessary for the present study is sketched, and the models are framed into the PISA assessment context.

Reading Cognitions—the Situation Model. Kintsch and van Dijk (1978) introduced a holistic model to describe the cognitive processes that are involved while readers attempt to understand a text. At the very base, the model states that readers extract *micropropositions* from the text base and store them in their memory. Each proposition either carries information about one of the text entities (i.a., persons, non-physical constructs) or about the relation between at least two of the entities (e.g., [parents] are aware of [bullying]). Observing the type of elements persons recall from texts, Kintsch and van Dijk concluded that readers build, in addition to micro-, also *macropropositions*. These contain higher-order information such as a paragraph's gist or additional propositions that are not explicitly stated but implied by the text. They can be inferred

by the reader from micropropositions (bottom-up), inferred or simply retrieved from the reader's knowledge (e.g., descriptive knowledge, schemata; top-down), or a combination of these. Applying the model, Kintsch and van Dijk (1978) showed that they were able to predict the probabilities of information recall when readers were asked to recall what they have read after a long break (1 and 3 months). Furthermore, good and bad readers differ in how easily they can access propositions from their memory and to what extent they are able to validly reconstruct information they cannot retrieve from their memory anymore (Kintsch & van Dijk, 1978). This particular finding serves as the base for our first operationalization measuring the number of propositions incorporated into a response. It can be interpreted as the number of elements in a student's mental situation model that are referenced to answer the question, irrespective of which level (micro or macro) they stem from. The usage of some type of proposition count, for diverse purposes, had been commonplace in the already referenced studies by Kintsch and van Dijk (also see Olson, Duffy, & Mack, 1985; Walker & Kintsch, 1985) and has been established since then (for a recent study see e.g., Martín-Loeches, Casado, Hernández-Tamames, & Álvarez-Linera, 2008). As the second operationalization of crucial features in the situation model, the present study also assesses whether a proposition comes from the micro- or macrostructures. Finally note, although the term *situation model* had originally concerned the macropropositions, it commonly refers to the whole mental representation of micro- and macropropositions by now (e.g., Zwaan & Radvansky, 1998).

Test Taker Responses as Approximation of their Cognitive Processes. While the previous section outlined how readers make sense of text, it is now of further interest how the created situation model serves to answer questions in tests such as the PISA literacy assessment. At the surface, answering such questions require two phases. First (cf. III.b), the student needs to cognitively query potential solutions by retrieving and/or inferring information from the memory and to decide for a subset of solutions. In the second phase (cf. IV), the student needs to articulate the chosen solution. These phases interact with each other and can iterate several times.

Elaborating on the first phase, Graesser and Franklin (1990) describe the model QUEST.

First of all, the student processes the question itself with two aims. (a) The question category needs to be identified in terms of the question function (i.a., `WHY`, `HOW`, `WHN`; Graesser & Clark, 1985, and Graesser & Murachver, 1985) combined with the type of semantic focus (i.e., `action`, `event`, `state`) resulting in categories such as `WHY<action>`. (b) Also, the question focus needs to be identified, defining the semantic aim of the question (Graesser & Franklin, 1990). For example in the sample item (cf. Figure 2), the student needs to focus on the semantics *article mentions death of Kiyoteru Okouchi* for answering.

After having identified category and focus, the student's cognitions query according information sources in the memory. For this, Graesser and Franklin (1990) distinguish *episodic knowledge structures* (EKS) that store information about a specific episode, such as yesterday evening's TV news about the success of a movie, and *generic knowledge structures* (GKS) that comprise more abstract knowledge aggregated from episodes, such as the distribution of how much is usually made by movies at the box office. GKS and EKS that are relevant with regard to the question focus are queried to gather potentially relevant information for answering. The student's situation model of the stimulus serves as one EKS. Conversely, if the question focuses on an area without propositions in the situation model, its macrostructures are enriched by the additionally queried knowledge structures. These steps are where situation model and QUEST come together.

Equivalently to propositions, every EKS and GKS comprises a set of statement nodes connected by arcs (i.a., `is consequence of`, `implies`; Graesser & Franklin, 1990). The student's query for information starts with so-called entry nodes (in the example i.a., *death of Kiyoteru Okouchi*) and ends up with several EKS, including the situation model, and GKS that can also be altered by the situation model's information. This mass of data is then reduced in the next step by selecting intersecting nodes across the knowledge structures, by omitting arcs irrelevant to the question category (e.g., different arcs are relevant for `CONS<event>` than for `WHY<action>` questions), and finally the resulting nodes and arcs are filtered in order to only end up with those that are semantically and conceptually compatible to the question focus (constraint propagation; Graesser & Franklin, 1990). These final statement nodes (propositions) serve as the

set of legitimate solutions. As another crucial feature of the student's cognitions, the semantic relevance of the expressed propositions for the question focus serves as the third measure for this study.

Along with the described processes during text comprehension as well as during crafting of the response, the students' cognitions always follow a given goal orientation (Graesser & Franklin, 1990; van Dijk & Kintsch, 1983). In the assessment context, the students ideally do their best to correctly answer the question. Any deviation from this reduces the probability of success, particularly for demanding processes. The entire process is not only influenced by the students' volition but, according to the TRACE model (Rouet & Britt, 2011), they also make decisions during reading—some implicit, some explicit—that are determined by their task understanding, called task model. For this, they bring internal resources into the process such as prior knowledge and self-regulation skills additionally influencing whether processes end up successfully or are initiated at all.

In the second phase, the student needs to express the chosen solution through text. Graesser and Clark (1985) and Graesser and Murachver (1985) describe the corresponding cognitive component to be highly dependent on the previously identified question category. That means for example, a `WHY<action>` question would lead to a concatenation of the solution propositions by category-specific terms such as *because* or *in order to* (Graesser & Clark, 1985, p. 269). A `HOW<action>` question would result in a concatenation of propositions by the term *by* and the transformation of the proposition's predicate to a gerund in case of a subordinate goal node (Graesser & Clark, 1985, p. 272). While this cognitive component is not very thoroughly defined in the model, the central idea is that the answer is produced by concatenating the required propositions by terms typical for the question category. The crux of this model specification in the second phase is twofold for the present study. First, the propositions are assumed to be directly incorporated into the response; hence, the linguistic level provides an ideal way for observing preceding cognitions. Second, the propositions of interest are linguistically enriched by category-specific terms that do not constitute proposition counterparts. It needs to be noted that

the described model does not provide much detail about the linguistic production, however, using purely psycholinguistic production models would go beyond this study's scope. We acknowledge that the production phase is further confounded by word retrieval and can also be hindered when applying the syntax.

Automatic Processing of Short Text Responses. The preceding subsections defined three measures worth capturing in student responses to distinguish features of the corresponding cognitions during reading: the response's total number of propositions, the degree of semantic closeness to the given text, and the propositions' relevance for answering the question. These features, picked from this complex process, meet two conditions—throughout, they crucially influence the process, and they are directly observable in the response. For example, relevance enters the process by the question focus when querying related knowledge structures and also significantly aids to narrow down the information to the final set of legitimate solution propositions. At the same time, relevance is inherent for deciding whether a response is correct or not, and thus, is apparently observable.

In order to show how the three features are operationalized, we first need to describe the general processing of responses. For this study, we adapted the software described by Zehner, Sälzer, and Goldhammer (2015). For the semantic classification of responses, the software tokenizes (splits) a response into words, corrects the spelling, cuts off affixes such as *-ing*, and removes semantically irrelevant words. The example response *"To give the background to why people are so concerned about bullying in Japan."* would result in the tokens `give`, `background`, `people`, `concern`, `bully`, and `japan`. Next, for gathering information about the semantics of words, the software applies a *Latent Semantic Analysis* (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) to a text corpus. This results in a dictionary in which each word is assigned to a 300-dimensional vector, building a so-called *semantic space*. In this space, the vectors of semantically similar terms point to similar directions, while vectors of distinct terms point to different directions. The software computes the centroid vector of all tokens for each response and clusters these to group responses into response types. Besides this procedure to build seman-

tically homogeneous groups of responses, the software's scope was extended for this study by an implementation of part-of-speech (POS) tagging. This means, the software annotates every word in a response with a tag for its part of speech (e.g., NN for *background*: *Normal Noun*). The Stuttgart–Tübingen Tagset (STTS; Schiller, Teufel, Stöckert, & Thielen, 1999) was applied to analyze German responses; see Table 1 in the Appendix for the tag meanings. The example response would thus read:

To/KOUI *give*/VVINF *the*/ART *background*/NN *to_why*/PWAV *people*/NN *are*/VAFIN *so*/ADV *concerned*/ADJ *about*/APPR *bullying*/NN *in*/APPR *Japan*/NN ./$.[3]

Where Kintsch and van Dijk (1978) refer to propositions as a whole, we operate at the level of what we call *proposition entities* (i.e., specific words) to approximate propositions. This is not precisely what was defined by the authors of the model since propositions most often consist of the specific combination of two or three words, such as CONCERNED(PEOPLE, BULLYING), which represents *"people are concerned about bullying."* For our operationalizations of the reading cognitions, propositions are not parsed but only approximated for four reasons. (1) NLP techniques are not (yet) able to reliably extract propositions from texts like responses in large-scale assessments that contain lots of informal, thus often improper, language (cf. Dzikovska, Nielsen, & Leacock, 2016; Higgins et al., 2014). Research dealing with automated speech recognition or informal messages similarly face the challenge to tolerate improper sentence boundaries, lots of syntactical and morphological errors, omitted or erroneous words, and so forth (Huang, Baker, & Reddy, 2014; Mohammad, Zhu, & Martin, 2014; Shrestha, Vulić, & Moens, 2015). Powerful approaches such as PropBank (Palmer, Gildea, & Kingsbury, 2005) are highly dependent on well-formed language, and recent improvements (e.g., Shrestha et al., 2015) are not available off-the-shelf. (2) Another shortcoming of these approaches is they cannot extract all propositions implied by a given number of words as it was intended by Kintsch and van Dijk (1978) because a single word can imply numerous propositions within the situation model. (3) Out of these numerous propositions, whether the propositions are relevant or not is furthermore depen-

[3]Note that for consistency purposes the showcases use the German STTS although it is not legitimate to apply German tags to English language. The tags are used in the way they would be used in the literal German translation.

dent on the question focus, category, and stimulus context. Finally, (4) a student response is a reaction to a specific question referring to a specific text and is, thus, constrained to a more or less narrow set of legitimate propositions. This leads to many formulations that imply more than they are expressing literally. For example, *"extreme"* could be a minimal response to the example question—although not included in the response, one would assume the corresponding subject to be *"Kiyoteru's bullying."* In linguistics, this phenomenon is referred to as pragmatics. These four reasons show that even single words need to be considered as proper propositions.

We name a word that is capable of constituting a proposition element a *proposition entity* (PE). A word is a PE if it genuinely adds to what is being referred to in the situation model, thus, it needs to be one of the following STTS tags: `ADJA`, `ADJD`, `ADV`, `NE`, `NN`, `PDS`, `PIS`, `PPER`, `PPOSS`, `PRELS`, `PWS`, `PWAV`, `PTKANT`, `VVFIN`, `VVIMP`, `VVINF`, `VVIZU`, `VVPP`, `VMFIN`, `VMINF`, or `VMPP` (cf. Table 1 in the appendix for the tags meanings). The bottom line is these are the following parts of speech:

- nouns and pronouns, which often refer to subjects in the situation model
- (non-auxiliary) verbs, which often refer to actions
- adjectives, which often describe subjects
- linguistic answer particles that are reactions to questions (e.g., "yes"), which often add a specific expression of attitude to a response

Therefore, all the PEs add genuine information to the response and they are not language artifacts or other word companions. The main logic about which tags are considered PEs is best described by examples. For example, auxiliary verbs (i.a., `VAFIN`) such as in *they*/`PPER` *are*/`VAFIN` *concerned*/`ADJ` are rather considered linguistic artifacts than indicators for genuine elements in the situation model. As another example, `PPOSSAT` is an attributive possessive pronoun followed by the noun it refers to—such as in *their*/`PPOSSAT` *concern*/`NN`. Instances of `PPOSSAT` are not considered PEs because the element that is genuinely introduced to the response is the *concern*.

On the other hand, words tagged with PPOSS are substituting possessive pronouns, which means that they are not followed by the respective noun—such as *theirs*/PPOSS. Words tagged with PPOSS, contrary to PPOSSAT, are considered PEs because they act as proper references to an entity in the situation model. Note that we critically discuss the concept of PEs in the Limitation Section. As introduced before, the study reports three measures. The first one is the count of PEs to approximate the number of incorporated propositions in the response, in line with the number of recalled propositions as an indicator for a reader's proficiency level (Kintsch & van Dijk, 1978).

The second measure distinguishes micro- and macropropositions as described by Kintsch and van Dijk (1978). Micropropositions are explicated in the text stimulus or question and the student only repeats them in the response. Macropropositions are logically inferred from the text or added from the student's knowledge by top-down processes. In order to measure how close a response's PE is to the stimulus or question, its semantics are compared to all tokens in the stimulus and question using cosine similarity in the semantic space. The maximum value of all these cosine similarities is then regarded as the PE's closeness to the text (high similarities indicate micropropositions repeating the stimulus or question, low similarities indicate macropropositions introducing new information compared to the stimulus and question). For better readability this is called the *micro* measure in the following. Some questions require micro-, others macropropositions in correct responses.

The third measure considers a PE's compatibility and importance to the question focus as defined in the QUEST model. Successful constraint propagation results in a response compatible to the question focus because the filtered knowledge structures had been selected with respect to it (Graesser & Franklin, 1990)—in other terms, the PE is relevant to the problem solution. We refer to this measure as *relevance* hereinafter. Here, relevance means the extent to which a PE contributes to the correctness of a response. The relevance measure, analogously to the micro measure, computes the maximum cosine similarity of a PE to all tokens given in correct example responses in the so-called PISA coding guides. These are documents for human coders that

include reference responses for deciding which responses should be considered correct. They constitute the best available source of entirely correct responses to a specific question which exclusively comprise propositions that are relevant to the solution.

Methods

Measures and Analyses

In summary, the analyses used three measures to distinguish responses (cf. Automatic Processing of Short Text Responses).

1. *Proposition Entity Count* (PEC): number of words that were annotated as `ADJA`, `ADJD`, `ADV`, `NE`, `NN`, `PDS`, `PIS`, `PPER`, `PPOSS`, `PRELS`, `PWS`, `PWAV`, `PTKANT`, `VVFIN`, `VVIMP`, `VVINF`, `VVIZU`, `VVPP`, `VMFIN`, `VMINF`, or `VMPP` (cf. Table 1 in the appendix)
2. *Micro*: a PE's degree of similarity to the stimulus text and question
3. *Relevance*: a PE's degree of relevance for possible solutions

Only PEs were included in the analysis of the micro and relevance measures. For this analysis, PEs with different degrees of micro and relevance measures were counted. These were relevant and irrelevant micropropositions, as well as relevant and irrelevant macropropositions. Propositions were classified as micropropositions if their micro similarity value within the distribution's upper 25 percent of all PEs, otherwise they were classified as macropropositions. Accordingly, propositions with a relevance value within the distribution's lower 25 percent of all PEs were classified as irrelevant, otherwise they were classified as relevant. This logic applied the conventional norm definition, with the middle 50 percent of a distribution constituting the average, while values below and above are below and above the average, respectively. That means, propositions needed to be at least somewhat dissimilar from the stimulus text and question to be macropropositions and at least somewhat similar to the reference responses in the coding guides to be relevant. While it is always worthwhile to include relevant propositions in a response, it depends on the question and stimulus whether micro- or macropropositions, or both, are necessary

to correctly answer the question. In order to not confound the measures with the PEC, the relevant and irrelevant micro- and macropropositions were analyzed as relative frequencies within the response.

Analyses involving the PEC were conducted at two levels. First, the real gender was used as the split criterion in order to analyze the measures' relationships with gender. Second, in accordance with the arguments presented in the Introduction, responses were additionally grouped into semantically homogeneous types by a cluster analysis. This was similar to Zehner, Sälzer, and Goldhammer (2015), except of that the entire data was used as training data without cross-validation. The numbers of clusters were chosen by the system's best performance (agreement between human and computer coding) aiming for relatively small numbers of clusters in order to attain clusters with appropriate sizes. Next, per cluster, the ratio between boys and girls and 95 percent Wilson confidence intervals was computed (Oranje, 2006; Wilson, 1927). Clusters with confidence intervals that did not overlap with the respective gender's expected value (given by the gender ratio for the item) were flagged as gender-specific. Only responses assigned to these clusters were used in the subsequent analyses. Other analyses than those involving the PEC were only conducted at this second level. This was, because the micro and relevance measures are semantic measures and it is not reasonable to assume that all responses by one gender were semantically homogeneous and pooling these was informative for investigating semantics; that is, all boy responses would contain the same semantics, and all girl responses would contain the same semantics, at the same time the two groups' semantics would be informatively different. Rather, the procedure identifies homogeneous responses typical for boys and girls, respectively, and then pools the gender-specific responses. This way, girl types can also contain some boy responses and vice versa. This is in line with the observation stated in the Introduction—the gender gap only represents a mean difference, whereas boys' and girls' distributions overlap for the most part. An implication of only including response types particularly associated with boys or girls is that we contrasted responses that are particularly gender-specific and did not study mixed response types. Also, this procedure decreased the sample size leading to lower testing power.

Because of responses' high dependency on the corresponding test item, all analyses were carried out per item. In most cases, the analyses controlled for the response correctness referring to judgments by PISA's trained human coders. For all analyses, only non-empty responses were included.

The analyses resulted from three research questions. (A) How do girls and boys differ in the number of propositions used in their responses? (B) How do gender-specific responses differ with respect to the use of micro- vs. macropropositions and the extent to which these are relevant? (C) How are the measures for response features related to further variables such as reading literacy and test motivation?

Materials

Items in PISA typically comprise a stimulus and a question referring to it. Responses used for the analyses stem from eight dichotomous items assessing reading literacy (cf. the reading items in Zehner, Sälzer, & Goldhammer, 2015). Due to repeated measurements, partly using the same items across cycles, the item contents are confidential and cannot be described here. The eight items are listed in Table 3, each given a name for better traceability in the following. Prior to item selection, a theoretical framework had been developed which item characteristics might potentially influence the automatic system's performance. According to this scheme, the selected items were intended to vary heterogeneously in their characteristics.

As presented in the table, the items ranged from difficult (10%) to easy (83%) with the majority being medium difficult and a slightly skewed distribution towards easiness. According to the PISA framework (OECD, 2013b), the assessed cognitive aspect mainly varied between the two of three aspects *Integrate and Interpret* and *Reflect and Evaluate* that both typically evoke more complex answers in linguistic terms than the third aspect does. The third aspect, *Access and Retrieve*, was only represented by one item and is assigned to items asking the test taker to find the relevant information given in the stimulus and repeat it. The question category as defined by Graesser and Clark (1985) varies for every item.

Participants and Procedure

The analyzed responses come from the German PISA 2012 sample. This includes a representative sample of 15-year-old students as well as a representative sample of ninth-graders in Germany. A detailed sample description can be found at Prenzel, Sälzer, Klieme, and Köller (2013) and OECD (2014). Due to a booklet design, the numbers of test takers varied for each item ($4152 \geq n \geq 4234$; cf. Table 3). In PISA 2012, *reading*, *maths*, and *science* were assessed paper-based. Hence, booklets needed to be scanned and responses were transcribed by six persons (cf. Zehner, Sälzer, & Goldhammer, 2015, for the detailed procedure).

Software

The used software implements open software, libraries, and packages. First, a database storing a Wikipedia dump was built by using *JWPL* (Zesch, Müller, & Gurevych, 2008), which also comes with an application programming interface to access the corpus data. *DKPro Similarity* (Bär, Zesch, & Gurevych, 2013), which in turn primarily utilizes *S-Space* (Jurgens & Stevens, 2010), is used to build a vector space model. The response processing makes use of components offered in *DKPro Core* (Gurevych et al., 2007), which fit into the *Apache UIMA Framework* (Ferrucci & Lally, 2004). For stemming, *Snowball* (Porter, 2001) is used, and for POS tagging, the German Stanford NLP parser with the PCFG model (Rafferty & Manning, 2008) is employed. For statistical matters, such as clustering, the software evokes *R* (Team, 2014), which was also used for further statistical analyses, partly using the packages *binom* for the computation of ratios' confidence intervals (Dorai-Raj, 2014) and *doSNOW* for parallel computations (Revolution Analytics & Weston, 2014).

Results

According to the research questions, this section first reports (A.I) how genders differ in their PEC and (A.II) how these results change when only gender-specific responses are included. Then, (B) the differences in the relevance and level of the PEs are depicted. At last, (C) relations of the measures with further variables are presented.

Proposition Entity Count (PEC)

PEC by Gender (A.I). Gender affected the number of PEs for every item significantly (cf. Table 4), in that girls incorporated more PEs into their responses. For 4·LIST RECALL, this means that, on average, girls used 0.4 PEs more than boys did. For 3·INTERPRET THE AUTHOR'S INTENTION, the effect corresponds to 1.5 PEs. The models controlled for the response correctness, meaning the mentioned gender effect is not caused by the fact that more girls give correct responses and correct responses are associated with more PEs. Rather, it appeared to be typical for boys to incorporate fewer PEs. For most items, the response correctness also showed a significant effect in that correct responses were associated with more PEs. There only was a significant interaction between gender and response correctness for 2·EVALUATE STATEMENT, showing that girls and boys did not use equally more PEs when responding correctly rather than incorrectly but girls gave even more PEs when responding correctly to this item. For all items, a significant overall variation across the four groups was found. Gender, the response correctness, and the respective interaction explained $R^2_{adj} = 2$ percent up to $R^2_{adj} = 16$ percent of PEC's variance.

PEC by Gender-Specific Types (A.II). Prior to this analysis, cluster analyses were carried out to identify gender-specific response types (see Table 2 in the Appendix for the resulting cluster solutions and the Methods Section for the rationale). Only responses assigned to significantly gender-specific types were included. As obvious in Table 5, the effect sizes of gender-specific types on the PEC were larger than the real gender effects in the models presented in the previous section. For 5·EVALUATE STYLISTIC ELEMENT, this means, on average, girl-specific types used 2.8 PEs more than boy-specific ones did. For 1·EXPLAIN PROTAGONIST'S FEELING, the effect corresponds to 4.9 PEs. Again remarkably, these effect sizes appeared although the models controlled for response correctness. Response correctness also turned out to be predictive for the PEC in most items. In most cases, correct responses were rather associated with more PEs. The items 1·EXPLAIN PROTAGONIST'S FEELING and 4·LIST RECALL were exceptions to this, in that, here, correct responses were associated with 1.5 and 0.4 fewer PEs, respectively, than were incorrect ones. For the latter, this finding is intuitive since students only needed to repeat

four terms from the stimulus text, whereas students answering incorrectly tended to write more than these four terms. The interaction effect between the response correctness and gender-specific type was significant for the items 1·EXPLAIN PROTAGONIST'S FEELING, 2·EVALUATE STATEMENT, and 6·VERBAL PRODUCTION. For the first of these, the interaction term partly compensated for the main gender effect, in that girl-specific types used 4.9 more PEs in general but the correct girl-specific types tended to use 1.5 fewer PEs (cf. Figure 3 for all items' group means and 95% CI). Due to the large standard error in this item for incorrect girl-specific responses, we do not give too much weight to this positive interaction. For the latter two items, the significant interactions added to the main gender effect—in these items successful female-specific responses tended to use even more PEs, while there was no difference between successful and unsuccessful male-specific responses. Overall, the models that included the gender-specific types all detected significant variation in the PEC across the four groups and explained more of the variation ($14\% \leq R^2_{adj} \leq 43\%$) than did the pure gender-based models presented in the previous section.

PE Features: Micro and Relevance Measures (B)

In this analysis, PEs were classified according to two dimensions, whether they constituted micro- or macropropositions and whether they were relevant or irrelevant to answering correctly. Again, only the gender-specific types were included. Figure 4 plots the relative frequencies of relevant micropropositions, irrelevant macropropositions, and so forth, within gender-specific response types. The lower, dashed abscissa shows the deviation of boy- and girl-specific *incorrect* responses. The deviations are sorted by their magnitude; while deviations dominated by boy-specific responses are presented on the left, those dominated by girl-specific responses are presented on the right. For example for item 1·EXPLAIN PROTAGONIST'S FEELING, the lower bar on the very right shows that incorrect girl-specific response types used 19 percent more relevant micropropositions (*Mic & Rel*) within their responses to this item than incorrect boy-specific response types did. The Venus symbol below the bar indicates that the deviation shown by the bar is dominated by girl-specific types. On the other side, the lower bar on the very left shows that incorrect boy-specific response types used 18 percent more macropropositions within their

responses (*Mac*). The Mars symbol below the bar indicates that the deviation given by the bar is dominated by boy-specific types. In addition to the magnitude of the deviation, the value of the dominating gender's incorrect responses for this micro/relevance class is shown by the dashed line. For example, the 18 percent more macropropositions in incorrect boy-specific responses is the result of boy-specific types having used 36 percent macropropositions. The analogous information for the *correct* responses is shown by the upper, solid abscissa and the solid line. Both lines are encapsulated by their 95 percent confidence intervals.

The figures for the eight items show the following. No items with very small deviations across the gender-specific types appeared overall. A few items occurred with relatively small deviations for correct responses (i.e., 1·EXPLAIN PROTAGONIST'S FEELING, 3·INTERPRET THE AUTHOR'S INTENTION, 6·VERBAL PRODUCTION). That is, for these items, boy-specific responses that were correct only differed marginally from girl-specific ones. On the other hand, for the same items, incorrect boy-specific responses differed from girl-specific ones. For example in item 6·VERBAL PRODUCTION, the lower bars and the dashed line show that girl-specific responses tended to include more *relevant* PEs than boy-specific ones did, particularly *relevant micropropositions*. Opposed to that, boy-specific responses incorporated about 12 percent more *irrelevant* PEs. The same figure can be found for item 3·INTERPRET THE AUTHOR'S INTENTION where incorrect girl-specific responses contained 17 percent more *relevant macropropositions* and boy-specific ones tended to use *irrelevant* PEs, *micropropositions*, *irrelevant macropropositions*, and *irrelevant micropropositions*. The interesting bottom line of these relations is, while *correct* boy-specific responses did not notably vary from girl-specific ones, *incorrect* girl-specific responses still contained more *relevant* PEs. Also, they still referred to the more appropriate micro/macro level of the situation model. In item 3, the author's intention needed to be reflected on. Boy-specific responses here predominantly named the text's *micropropositions*. Although having failed to pass the threshold of correctness, the *incorrect* girl-specific responses did treat the correct situation model's level to answer the question. It was vice versa in 1·EXPLAIN PROTAGONIST'S FEELING, where the question was intended to be answered using *micropropositions* from

the text. Here again, incorrect girl-specific responses retrieved their PEs from the *microstructures* (also *relevantly*), while boy-specific responses consisted of 18 percent more *macropropositions* (36% in total). Hence overall, the girl-types seem to refer to *relevant* PEs and the more appropriate level in the situation model (independently from the response's final correctness). For the other items, the same figures can be found rather consistently for incorrect responses as well as correct responses.

The described pattern was only broken in 8·EXPLAIN STORY ELEMENT. There, *correct* boy-specific responses contained more *relevant micropropositions* and generally more *relevant* PEs than girl-specific ones did. Here, the correct girl-specific responses appeared to contain predominantly *micropropositions* (80%), of which several were *irrelevant* (20%). On the contrary for *incorrect* responses, the pattern turned back again in favor of girl-specific responses.

Exploring Relations to Further Variables (C)

Despite the differences in cognitive approaches between girls and boys, as described in the previous section, the gender gap in this data was also crucially influenced by the large number of boys who did not respond at all. For all analyzed items, there are significant proportions of empty responses by boys. On average, they produced 63 percent of the empty responses across the eight items ($SD = 7\%$).

Furthermore, the PEC turned out to correlate with the students' overall reading literacy from moderately in 1·EXPLAIN PROTAGONIST'S FEELING ($r = .16$) to fairly in 7·SELECT AND JUDGE ($r = .32$). This figure hardly changed when the models were controlled for the students' general test motivation. In PISA 2012, the test motivation was measured by assessing the difference between self-reported engagement in the PISA test and self-reported expected engagement in a test relevant for the student's school grades. Overall, this measure revealed some flaws but constituted the best available indicator for the test motivation in the 2012 data for the reading items. The test motivation itself was only marginally negatively related to the PEC with r ranging from $-.02$ for 6·VERBAL PRODUCTION to $-.11$ for 7·SELECT AND JUDGE. That means the PEC is not just an outcome of test motivation, mediated through the response length.

Across all items, the number of relevant PEs used in responses was fairly correlated with the students' overall reading literacy—ranging from $r = .21$ in 1·EXPLAIN PROTAGONIST'S FEELING to $r = .39$ in 4·LIST RECALL. However, the relationship between the two variables is not exhaustively mirrored by the linear correlation coefficient (cf. right part of Figure 5 as an example). The scatter plots reveal a linear relation between the number of relevant PEs used and the students' reading literacy in that few relevant PEs can be associated with low reading literacy, and the more relevant PEs included, the higher the students' reading literacy was. But the opposite was not true. Highly literate students did not always use many relevant PEs in their responses, because they were also able to express their solution with only a few relevant PEs. At the same time, the left part of Figure 5 shows that highly literate students did not use less PEs than needed (in this item, they needed to name at least two of four terms from a list), hence, they appeared to successfully adapt to the question.

Finally, the difference between splitting by gender and automatically selected gender-specific types becomes apparent in Table 6, which illustrates the itemwise gender gap by split criterion. Note that the selection of the types was exclusively determined by significant proportions of the genders. Please see Table 2 in the Appendix for the percentage of students included in gender-specific types. When splitting by the students' gender, the girls' advantage of solving the items ranges from 1 to 14 percent, and when comparing the gender-specific types, the advantage notably increases, ranging from 22 to 71 percent.

Discussion

The present study followed two interests. First, it aimed to shed further light on the gender differences in reading literacy. Second, a theoretical framework along with according techniques were assembled in order to make raw responses in large-scale assessments accessible to studies as a new source of information. The theoretical framework sketches how mental representations of a stimulus are created by the student and how these are crafted into a response. The situation model (van Dijk & Kintsch, 1983) comprises propositions that are spread across different levels (micro- and macrostructures) varying in how close they are to the text base. Next,

according to QUEST (Graesser & Franklin, 1990), when attempting to answer a test question, the student first identifies the question's semantic focus and conceptual category. The student then uses the situation model as one source of several to gather potential propositions that would help to answer the question. In order to narrow down potential solutions, the propositions are filtered with respect to the semantic and conceptual fit to the question focus and category. Having decided on a subset of propositions as the solution, the student finally needs to express this in written language by concatenating the selected propositions using a linguistic template in line with the question category (Graesser & Murachver, 1985; Graesser & Clark, 1985). In this paper, we suggest that specific properties of the resulting raw response can be mapped back to successful or flawed processes, and the framework sketches the way the propositions take from the situation model towards the written response. This is done via automatic processing of the responses by extracting approximations of propositions (proposition entities; PEs), part-of-speech tagging, and semantic comparisons using Latent Semantic Analysis (Deerwester et al., 1990). This procedure was employed using the software first described in Zehner, Sälzer, and Goldhammer (2015).

The two PEC analyses and the gender gap's increase, induced by the gender types, illustrated that it is not reasonable to assume two distinct and homogeneous cognitive types separated by gender. Rather, there appear to be different cognitive types in the construct of reading literacy that are unevenly spread across genders. The boy types are characterized by parsimonious selection of PEs. On average, they use three to five PEs less than the girl types. And this is only the effect size of the cognitive type on the number of PEs within either correct or incorrect responses. The phenomenological effect for boys tends to be even higher because their probability to respond incorrectly is higher.

In some cases, the boy types' cognitive frugality suffices to produce correct responses; on top, this can seem very concise and worthwhile. This positive perspective can be flipped in light of evidence for the majority of the analyzed items, that the boy-specific characteristics were relatively consistent across correct as well as incorrect responses. That is, the boy-specific cognitive types seem to involve a less stable situation model and struggle with retrieving and inferring from

it. The high number of PEs integrated in girl-specific responses emphasizes, that these cognitive types liberally juggle the information in the situation model. Plus, the analyses indicate that, the PEs used in girl-specific responses are not simple recalls of any random information from the stimulus, but they nearly always use more PEs that are relevant and compatible to the question focus than the boy-specific responses do. The relatively few PEs boy-specific responses tend to use are more often irrelevant than those in girl-specific responses—and this is true regardless of the response correctness for almost all items. This figure adds to the previous notion that the boy types have difficulties accessing and searching the relevant parts of the situation model.

Given a required threshold of engagement, readers who realize a gap or inconsistency in their situation model try to reconstruct the missing information (Kintsch & van Dijk, 1978). This reconstruction phenomenon occurs throughout the data, but when this process fails, it is most often associated with the boy types. This can be observed in the high frequencies of irrelevant macropropositions. In these cases, the students try to come up with some information from their knowledge structures that fit the question focus in any way. Another strategy to cope with gaps in the situation model that are targeted by the question focus is to repeat micropropositions. Interestingly, both failing strategies are associated with the boy types. For example in 4·LIST RECALL, students need to directly copy four terms from the stimulus to their response. This task is a prototype for recalling micropropositions. While even the correct boy-specific responses consist of 10 percent of (irrelevant) macropropositions, the corresponding incorrect responses incorporate 54 percent of macropropositions. The same pattern but reversed can be found in other items. For example in 3·INTERPRET THE AUTHOR'S INTENTION, macropropositions are necessary to identify the author's subtext. In opposition, the boy-specific responses to this item consist of about 40 percent of micropropositions, of which 63 percent are classified as irrelevant. The bottom line is, the girl-specific cognitive types seem to identify the question category more easily in order to conclude, which knowledge structures contain the most relevant information for the correct answer.

Of course, all the discussed patterns constitute a reduction of this complex matter, and

single outcomes in the analysis do not support the proposed interpretation. For example in 5·EVALUATE STYLISTIC ELEMENT in correct as well as incorrect responses, boy-specific responses contain 10 percent more relevant macropropositions than girl-specific ones do. At the same time, girl-specific responses use 13 percent more relevant micropropositions. There seem to be two equivalently legitimate lines of reasoning that are predominantly used by one of the types. Notably, the gender gap for this item is the second smallest and to a large part stems from empty responses by boys.

Finally, test motivation was only marginally related to the measures. Partly, this might have come from the overall test motivation measure that was not directly concerning the reading assessment. On the other hand, this shows that the PEC is an independent measure holding important information. In this context, the findings by Artelt et al. (2010) are interesting. They showed that, when controlling for reading engagement and strategies, gender does not account for a significant amount of variance in the reading literacy anymore. The reading engagement indeed biased our analyses, because the analyzed items constitute only a subset of items from the reading assessment. This subset could not be balanced with regard to gender-specific interests. Reading strategies, on the other hand, are a crucial determining variable for the observations and interpretations in this paper. The reading strategies that were assessed in PISA 2009 by student self reports were *memorization*, *elaboration*, and *control strategies* (OECD, 2010a). These are strategies to different (meta-)cognitive approaches. They would directly result in different cognitive types as described in this paper. The memorization strategy refers to the degree of retentivity of micropropositions, while elaboration aims at integrating different propositions and referring from them, which would result in macropropositions. Since the PISA 2009 raw responses are paper-based and, hence, not accessible, and PISA 2012 did not assess the reading strategies, it will be highly interesting to replicate the analyses presented here with the data of PISA 2018. With reading being the main domain, these data are going to contain information about reading engagement and strategies.

Limitations and Directions

First of all, the analyses used only German data and need to be replicated for other languages. As previously mentioned, the available data stems from the paper-based assessment in PISA 2012, hence, the data needed transcription, resulting in only a subset of the data. With the new computer-based assessment in PISA from 2015 on, the full set of data will be available to the analysis. For the relevance measure, the PISA coding guides were used as the benchmark in lack of a better gold standard. As previously shown (Zehner, Goldhammer, & Sälzer, 2015) they are not exhaustive, in that, they do not cover the whole range of empirically occurring response types. Hence, the relevance measure reported here might be underestimating the true values. Still, they were the best available objective referential benchmark as to which PEs should appear in correct answers.

The operationalizations are only rough transformations of the respective parts in the theoretical models they are based on. This is mainly due to the performance level of contemporary natural language processing (NLP) techniques. We regard the proposed framework and employed technologies as a starting point for further works to refine and expand the measures. The NLP techniques made huge steps in the last two decades having enabled the presented analyses. But on the other hand, they did not reach a level at which, among others, propositions could be extracted reliably and validly from texts, such as the presented student responses that are teeming with ill-formed language (Higgins et al., 2014); that is not so much a problem at the morphological but mainly at the syntactical level. Unfortunately, this is not only a technical problem but it is a continuum as to what degree improper language can be decomposed into the writer's intentions at all (cf. Foltz, 2003). Furthermore, the models assembled in the theoretical framework contain a lot more detail in the features they are implying in cognitive processes. The features that were selected for operationalization in this study were those that were most informative to the matter of the reading literacy gender gap and at the same time feasible for the NLP techniques from the authors' perspective. Finally, the theoretical framework will also be considered for further elaboration in future works. As an example, the recent framework neglects the decoding processes

taking place while reading and comprehending.

Additionally, the operationalizations could gain further elaboration in future studies by using additional linguistic information. For example, the PEs are determined by their parts of speech. In some cases, the definition, what can be counted as a genuine incremental information, is a bit flawed. The general logic is described in the Theory Section. The example there is that pronouns tagged with PPOSSAT (e.g., *their [concern]*) are not considered PEs but PPOSS pronouns (e.g., *theirs*) are. Contrary, adjectives (ADJA, ADJD, ADV; e.g., *appropriate*) are considered PEs because they typically add genuine information that is not already given by the noun or verb they are referring to. There are cases in which this consistently applied logic has a flaw. That is, when for example adjectives take the same semantic role like PPOSSAT do. For example in *the*/ART *corresponding*/ADJA *concern*/NN, the adjective merely adds new quality to what is being said about the concern but rather further defines which concern is being talked about—a linguistic function that could be regarded quite similar to the one of the pronoun in *their*/PPOSSAT *concern*/NN. Accordingly, one could also regard auxiliary verbs as indicators for the mental representation of the tense in the situation model. This shows that what is being considered a PE is a question of relevance, because tenses would play an important role in a stimulus text in which the order of events is crucial. The definition of PEs in this study is fully compatible to the stimuli and items on which the analyzed responses are based on. With the definition of PEs being only approximations of the cognitions, such a slight flaw is regarded as acceptable, but there might be further developments in the next years that will allow a more reliable measurement. Also, some of the described problems could be overcome by using a constituent analysis (i.e., decomposing sentences into parts that linguistically belong together, e.g., words all acting as the noun phrase in a sentence).

Conclusion

In sum, our study shows that open-ended student responses contain several elements that help understand the frequently found gender gap in reading achievement. Despite the listed constraints, it appears worthwhile to utilize nascent innovations in natural language processing to

investigate raw responses as a new source of information in large-scale assessments. In future studies, it will be necessary to base the analysis on the entire item set. Also, international comparisons will add another interesting dimension to the research matter.

References

Artelt, C., Naumann, J., & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme et al. (Eds.), *PISA 2009* (pp. 73–112). Münster u.a.: Waxmann.

Bär, D., Zesch, T., & Gurevych, I. (2013). DKPro Similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 121–126). Sofia, Bulgaria: Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6{\textless}391::AID-ASI1{\textgreater}3.0.CO;2-9

Dorai-Raj, S. (2014). *binom: Binomial Confidence Intervals For Several Parameterizations.* Retrieved from `http://CRAN.R-project.org/package=binom`

Drechsel, B., & Artelt, C. (2007). Lesekompetenz. In M. Prenzel et al. (Eds.), *PISA 2006* (pp. 225–247). Münster u.a.: Waxmann.

Dzikovska, M. O., Nielsen, R. D., & Leacock, C. (2016). The joint student response analysis and recognizing textual entailment challenge: Making sense of student responses in educational applications. *Language Resources and Evaluation*, *50*(1), 67–93. doi: 10.1007/s10579-015-9313-8

Elley, W. B. (1994). *The IEA study of reading literacy*. Oxford: Pergamon Press.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327–348. doi: 10.1017/S1351324904003523

Foltz, P. W. (2003). Quantitative Cognitive Models of Text and Discourse Comprehension. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 487–523). Mahwah, NJ: Erlbaum.

Gambell, T. J., & Hunter, D. M. (1999). Rethinking gender differences in literacy. *Cana-*

dian Journal of Education / Revue canadienne de l'éducation, *24*(1), 1–16. doi: 10.2307/1585767

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge* (Vol. 17). Norwood, NJ: Ablex.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*(3), 279–303. doi: 10.1080/01638539009544760

Graesser, A. C., & Murachver, T. (1985). Symbolic procedures of question answering. In A. C. Graesser & J. Black (Eds.), *The psychology of questions* (pp. 15–88). Hillsdale, N. J.: Erlbaum.

Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., & Zesch, T. (2007). Darmstadt knowledge processing repository based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology.* Tübingen.

Heritage, J. (2005). Cognition in discourse. In H. te Molder & J. Potter (Eds.), *Conversation and cognition* (pp. 184–202). Cambridge, New York: Cambridge University Press.

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., ... Blackmore, J. (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *CoRR*, *abs/1403.0801*.

Hohn, K., Schiepe-Tiska, A., Sälzer, C., & Artelt, C. (2013). Lesekompetenz ins PISA 2012: Veränderungen und Perspektiven. In M. Prenzel, C. Sälzer, E. Klieme, & O. Köller (Eds.), *PISA 2012: Fortschritte und Herausforderungen in Deutschland* (pp. 217–244). Münster: Waxmann.

Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Commun. ACM*, *57*(1), 94–103. doi: 10.1145/2500887

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69. doi: 10.1037/0033-2909.104.1.53

ISOcat. (2008). *STTS Stuttgart Tübingen tag set.* Retrieved 12.01.2016, from `http://www`

.isocat.org/rest/dcs/376

Jurgens, D., & Stevens, K. (2010). The S-Space package: An open source package for word space models. In A. for Computational Linguistics (Ed.), *48th Annual Meeting of the Association for Computational Linguistics: Proceedings of System Demonstrations* (pp. 30–35).

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363. doi: 10.1037/0033-295X.85.5.363

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative Studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, *8*(1), 69–79. doi: 10.2304/eerj.2009.8.1.69

Lietz, P. (2006). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view. *International Education Journal*, *7*(2), 127–149.

Martín-Loeches, M., Casado, P., Hernández-Tamames, J. A., & Álvarez-Linera, J. (2008). Brain activation in discourse comprehension: A 3t fMRI study. *Neuroimage*, *41*(2), 614–622. doi: 10.1016/j.neuroimage.2008.02.047

Mohammad, S., Zhu, X., & Martin, J. (2014). Semantic role labeling of emotions in tweets. In A. for Computational Linguistics (Ed.), *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 32–41). doi: 10.3115/v1/w14-2607

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International results in reading*. Boston College: Chestnut Hill, MA.

Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme et al. (Eds.), *PISA 2009* (pp. 23–71). Münster u.a.: Waxmann.

NCES. (2015). *The nation's report card: 2015 mathematics and reading assessments.* Retrieved 16.01.2016, from http://www.nationsreportcard.gov/reading_math_2015

OECD. (2002). *Reading for change: Performance and engagement across countries*. Paris: OECD Publishing. doi: 10.1787/9789264099289-en

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD Publishing. doi: 10.1787/9789264006416-en

OECD. (2006). *PISA released items - reading*. Retrieved 18.02.2016, from `http://www.oecd.org/pisa/38709396.pdf`

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1: Analysis*. Paris: OECD Publishing. doi: 10.1787/9789264040014-en

OECD. (2010a). *PISA 2009 results: Learning to learn - student engagement, strategies and practices (volume III)*. Paris: OECD Publishing. doi: 10.1787/9789264083943-en

OECD. (2010b). *PISA 2009 Results: What students know and can do: Student performance in reading, mathematics and science (Volume I)*. Paris: OECD Publishing. doi: 10.1787/9789264091450-en

OECD. (2013a). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. OECD Publishing. doi: 10.1787/9789264204256-en

OECD. (2013b). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. doi: 10.1787/9789264190511-en

OECD. (2014). *PISA 2012 results: What students know and can do (volume I, revised edition, February 2014)*. Paris: OECD Publishing. doi: 10.1787/9789264208780-en

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. Paris: OECD Publishing. doi: 10.1787/9789264229945-en

Olson, G. M., Duffy, S. A., & Mack, R. L. (1985). Question-asking as a component of text comprehension. In A. C. Graesser & J. Black (Eds.), *The psychology of questions* (pp. 219–226). Hillsdale, N. J.: Erlbaum.

Oranje, A. (2006). *Confidence intervals for proportion estimates in complex samples* (Vol. RR-06-21). Princeton, N.J.: ETS. doi: 10.1002/j.2333-8504.2006.tb02027.x

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71–106. doi: 10.1162/0891201053630264

Porter, M. (2001). *Snowball: A language for stemming algorithms.* Retrieved from `http://snowball.tartarus.org/texts/introduction.html`

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Rafferty, A. N., & Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German* (pp. 40–46). doi: 10.3115/1621401.1621407

Revolution Analytics, & Weston, S. (2014). *doSNOW: Foreach parallel adaptor for the snow package.* Retrieved from `http://CRAN.R-project.org/package=doSNOW`

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302. doi: 10.3102/0002831210372249

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In J. P. Magliano, M. T. McCrudden, & G. J. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Charlotte, NC: Information Age Pub.

Schaffner, E., Schiefele, U., Drechsel, B., & Artelt, C. (2004). Lesekompetenz. In M. Prenzel et al. (Eds.), *PISA 2003* (pp. 93–110). Münster: Waxmann.

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS* (Tech. Rep.). University of Stuttgart and University of Tübingen.

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, *50*(2), 219–232. doi: 10.1002/rrq.92

Shrestha, N., Vulić, I., & Moens, M.-F. (2015). Semantic role labeling of speech transcripts.

In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 9042, pp. 583–595). Springer International Publishing. doi: 10.1007/978-3-319-18117-2{\textunderscore}43

Stanat, P., & Kunter, M. (2002). Geschlechterspezifische Leistungsunterschiede bei Fünfzehnjährigen im internationalen Vergleich. *Zeitschrift für Erziehungswissenschaft*, *4*(1), 28–48. doi: 10.1007/s11618-002-0003-0

te Molder, H., & Potter, J. (2005). Mapping and making the terrain. In H. te Molder & J. Potter (Eds.), *Conversation and cognition* (pp. 8–54). Cambridge, New York: Cambridge University Press.

Team, R. C. (2014). *R: A language and environment for statistical computing*. Vienna. Retrieved from `http://www.R-project.org/`

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study. International studies in evaluation, III*. Stockholm: Almqvist and Wiksell.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2009). *The psychology of survey response* (10. print ed.). Cambridge: Cambridge Univ. Press. doi: 10.1017/cbo9780511819322

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

Walker, W. H., & Kintsch, W. (1985). Automatic and strategic aspects of knowledge retrieval. *Cognitive Science*, *9*(2), 261–283. doi: 10.1207/s15516709cog0902{\textunderscore}3

Wetherell, M. (2007). A step too far: Discursive psychology, linguistic ethnography and questions of identity. *Journal of Sociolinguistics*, *11*(5), 661–681. doi: 10.1111/j.1467-9841.2007.00345.x

White, B. (2007). Are girls better readers than boys? Which boys? Which girls? *Canadian Journal of Education/Revue canadienne de l'éducation*, 554–581. doi: 10.2307/20466650

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. doi: 10.2307/2276774

Wolters, C. A., Denton, C. A., York, M. J., & Francis, D. J. (2014). Adolescents' motivation for

reading: group differences and relation to standardized achievement. *Reading and Writing*, *27*(3), 503–533. doi: 10.1007/s11145-013-9454-3

Zehner, F., Goldhammer, F., & Sälzer, C. (2015). Using and improving coding guides for and by automatic coding of PISA short text responses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* Retrieved from http://www.aspiringminds.com/pages/assess/2015/papers/PID3896317.pdf doi: 10.1109/icdmw.2015.189

Zehner, F., Sälzer, C., & Goldhammer, F. (2015). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*. Retrieved from http://epm.sagepub.com/content/early/2015/06/06/0013164415590022 doi: 10.1177/0013164415590022

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech and Morocco.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. doi: 10.1037//0033-2909.123.2.162

Appendix A

Table 1
Stuttgart–Tübingen Tagset (Schiller et al., 1999; adapted with slight changes from ISOcat, 2008)

Tag	Description	Example (German)	Example (Literal Translation)
$(	other sentence internal punctuation	- [ ] ( ) "	
$,	comma	,	
$.	sentence final punctuation	. ? ! ; :	
*ADJA	attributive adjective	[das] große [Haus]	[the] big [house]
*ADJD	adverbial or predicative adjective	[er fährt] schnell, [er ist] schnell	[he drives] quickly, [he is] fast
*ADV	adverb	schon, bald, doch	already, soon, however
APPO	postposition	[ihm] zufolge, [der Sache] wegen	[him] according_to, [the point] because_of
APPR	preposition or left circumposition	in [der Stadt], ohne [mich]	in [the town], without [me]
APPRART	preposition with article	im [Haus], zur [Sache]	in_the [house], to_the [point]
APZR	right circumposition	[von jetzt] an	[from now] on
ART	definite or indefinite article	der, die, das, ein, eine	the *(three grammatical genders)*, a *(two grammatical genders)*
CARD	cardinal number	zwei [Männer], [im Jahre] 1994	two [men], [in the year] 1994
ITJ	interjection	mhm, ach, tja	aha, oh, well
KOKOM	particle of comparison, no clause	als, wie	as, like_a
KON	coordinative conjunction	und, oder, aber	and, or, but
KOUI	subordinating conjunction with "zu" and infinitive	um [zu leben], anstatt [zu fragen]	for [living], instead_of [asking]
KOUS	subordinating conjunction with clause	weil, dass, damit, wenn, ob	because, that, so_that, when, if
*NE	proper noun	Hans, Hamburg, HSV	Hans, Hamburg, HSV

Tag	Description	Example (German)	Example Literal Translation
*NN	noun	Tisch, Herr, [das] Reisen	table, Mr., traveling
PDAT	attributive demonstrative pronoun/alternatively: for demonstrative pronouns that occur in an NP	jener [Mensch]	that [person]
*PDS	substituting demonstrative pronoun/alternatively: for demonstrative pronouns that substitute an NP	dieser, jener	this_one, those
PIAT	attributive indefinite pronoun without determiner/alternatively: for indefinite pronouns without determiner that occur in an NP	kein [Mensch], irgendein [Glas]	no [one], any [glass]
PIDAT	attributive indefinite pronoun with determiner/alternatively: for indefinite pronouns with determiner that occur in an NP	[ein] wenig [Wasser], [die] beiden [Brüder]	[a] little [water], [the] both [brothers]
*PIS	substituting indefinite pronoun/alternatively: for indefinite pronouns that substitute an NP	keiner, viele, man, niemand	nobody, many, one, no_one
*PPER	irreflexive personal pronoun	ich, er, ihm, mich, dir	I, he, him, me, to_you
PPOSAT	attributive possessive pronoun/alternatively: for possessive pronouns that occur in an NP	mein [Buch], deine [Mutter]	my [book], your [mother]
*PPOSS	substituting possessive pronoun/alternatively: for possessive pronouns that substitute an NP	meins, deiner	mine, yours
PRELAT	attributive relative pronoun/alternatively: for relative pronouns that occur in an NP	[der Mann,] dessen [Hund]	[the man,] whose [dog]
*PRELS	substituting relative pronoun/alternatively: for relative pronouns that substitute an NP	[der Hund,] der	[the dog,] that
PRF	reflexive personal pronoun	sich, einander, dich, mir	oneself, each other, to_yourself, myself
PROP	pronominal adverb	dafür, dabei, deswegen, trotzdem	for_that, with_that, because_of_that, anyway
PTKA	particle with adjective or adverb	am [schönsten], zu [schnell]	the [most beautiful], too [fast]
*PTKANT	answer particle	ja, nein, danke, bitte	yes, no, thanks, please

Tag	Description	Example (German)	Example Literal Translation
PTKNEG	negation particle	nicht	not
PTKVZ	separated verb particle	[er kommt] an, [er fährt] rad	[he comes] in, [he rides] bike
PTKZU	"zu" in front of infinitive	zu [gehen]	to [go]
PWAT	attributive interrogative pronoun/alternatively: for interrogative pronouns that occur in an NP	welche [Farbe], wessen [Hut]	which [colour], whose [hat]
*PWAV	adverbial interrogative or relative pronoun	warum, wo, wann, worüber, wobei	why, where, when, what [...] about, what [...] when
*PWS	substituting interrogative pronoun/alternatively: for interrogative pronouns that substitute an NP	wer, was	who, what
TRUNC	truncated word – first part	An- [und Abreise]	em- [and disembark]
VAFIN	finite verb, auxiliary verb	[du] bist, [wir] werden	[you] are, [we] will
VAIMP	imperative, auxiliary verb	sei [ruhig!]	be [quiet!]
VAINF	infinitive, auxiliary verb	werden, sein	to_become, to_be
VAPP	past participle, auxiliary verb	gewesen	has_been
*VMFIN	finite verb, modal verb	[wir] dürfen	[we] may
*VMINF	infinitive, modal verb	wollen	to_want_to
*VMPP	past participle, modal verb	gekonnt	has_been_able_to
*VVFIN	finite main verb	[du] gehst, [wir] kommen [an]	[you] go, [we] arrive
*VVIMP	imperative, main verb	komm [!]	come [!]
*VVINF	infinitive, main verb	gehen, ankommen	to_go, to_arrive
*VVIZU	infinitive + "zu", main verb	anzukommen, loszulassen	[in order] to_come, [in order] to_release
*VVPP	past participle, main verb	gegangen, angekommen	gone, [have] arrived
XY	non-word containing special characters	3:7, H2O, D2XW3	3:7, H2O, D2XW3

* considered as proposition entity

Appendix B

Table 2
Cluster Solutions for Gender Type Analyses

Item	n_{cl}	$n_{sigcl_{\male}}$	$n_{sigcl_{\female}}$	$n_{\male type.0}$	$n_{\male type.1}$	$n_{\female type.0}$	$n_{\female type.1}$	$\%_{gt}$
#1	130	14	13	180 (75%)	503 (68%)	15 (73%)	472 (71%)	29%
#2	130	6	6	952 (58%)	76 (76%)	78 (67%)	141 (67%)	37%
#3	120	7	14	1379 (61%)	12 (50%)	394 (69%)	126 (69%)	64%
#4	32	8	2	706 (62%)	257 (68%)	0 (0%)	2138 (55%)	83%
#5	170	13	13	800 (65%)	136 (80%)	127 (71%)	366 (68%)	40%
#6	230	16	17	290 (76%)	253 (69%)	11 (64%)	441 (73%)	25%
#7	95	10	11	580 (68%)	160 (71%)	59 (63%)	607 (70%)	37%
#8	170	13	16	457 (67%)	326 (66%)	47 (64%)	450 (76%)	33%

Note. $n_{\male type.0}$ corresponds to the number of incorrect responses assigned to a boy type, $n_{\female type.1}$ corresponds to the number of correct responses assigned to a girl type, etc. Percentages in parentheses give the proportions of students who indeed were boys or girls, respectively. $\%_{gt}$ gives the percentage of responses included in the gender-specific analyses compared to non-empty responses in total.

Table 3
Item Characteristics

Item	Cat[a]	Aspect[b]	Correct	*n*	♀[c]	Words[d]
1·EXPLAIN PROTAGONIST'S FEELING	WHY<s>	B	83%	4,152	49%	12.3 (4.6)
2·EVALUATE STATEMENT	ENABLE<a>	C	43%	4,234	48%	15.6 (9.0)
3·INTERPRET THE AUTHOR'S INTENTION	SIG<a>	B	10%	4,234	48%	12.5 (6.3)
4·LIST RECALL	CON<s>	A	59%	4,223	50%	5.6 (3.0)
5·EVALUATE STYLISTIC ELEMENT	HOW<s>	C	56%	4,234	48%	14.7 (6.2)
6·VERBAL PRODUCTION	WHN<a>	B	80%	4,152	49%	12.4 (6.9)
7·SELECT AND JUDGE	ENABLE<s>	C	68%	4,152	49%	13.6 (7.0)
8·EXPLAIN STORY ELEMENT	CON<e>	B	69%	4,223	50%	14.4 (5.5)
Total			59%	33,604	49%	12.6 (6.1)

[a] question category according to QUEST (Graesser & Clark, 1985; Graesser & Murachver, 1985); in the form FUNCTION<T> with FUNCTION∈{WHY, HOW, ENABLE, CONS, WHEN, WHERE, SIG, WHN} and T∈{*state*, *action*, *event*}

[b] according to PISA framework (OECD, 2013b), A = Access & Retrieve, B = Integrate & Interpret, C = Reflect & Evaluate

[c] percentage of girls (note that for items 1, 4, 6, 7, and 8 there was one case with missing information about the student's gender each, referring to two students)

[d] word count in non-empty responses on average (with SD)

Table 4
Proposition Entity Count (PEC) by Gender and Response Correctness

Item	β_g	β_c	β_{g*c}	F(df1, df2)	R^2_{adj}
1·EXPLAIN PROTAGONIST'S ...	**-0.16** [±0.09]	-0.02 [±0.05]	0.01 [±0.21]	$F(3,4047)$ **= 32.10**	.023
2·EVALUATE STATEMENT	**-0.12** [±0.05]	**0.18** [±0.05]	**-0.15** [±0.15]	$F(3,3365)$ **= 55.42**	.046
3·INTERPRET AUTHOR'S ...	**-0.20** [±0.04]	0.03 [±0.05]	-0.03 [±0.15]	$F(3,2989)$ **= 43.90**	.041
4·LIST RECALL	**-0.12** [±0.09]	**0.23** [±0.05]	0.14 [±0.20]	$F(3,3718)$ **= 97.61**	.072
5·EVALUATE STYLISTIC ...	**-0.15** [±0.05]	**0.36** [±0.04]	-0.02 [±0.14]	$F(3,3540)$ **= 226.30**	.160
6·VERBAL PRODUCTION	**-0.16** [±0.09]	**0.14** [±0.05]	0.01 [±0.21]	$F(3,3959)$ **= 70.56**	.050
7·SELECT AND JUDGE	**-0.18** [±0.06]	**0.23** [±0.05]	0.01 [±0.15]	$F(3,3764)$ **= 128.90**	.092
8·EXPLAIN STORY ELEMENT	**-0.17** [±0.06]	**0.11** [±0.05]	0.03 [±0.15]	$F(3,3893)$ **= 55.41**	.040

* bold statistics are significant ($\alpha = .05$), g = gender (1 = girls, 2 = boys), c = response correct, $R^2_{adj} = R^2$ adjusted, 95% confidence intervals in brackets

Table 5
Proposition Entity Count (PEC) by Gender-Specific Types and Response Correctness

Item	β_{gt}	β_{c}	β_{gt*c}	F(df1, df2)	R^2_{adj}
1·EXPLAIN PROTAGONIST'S ...	**-0.84** [±0.21]	**-0.19** [±0.15]	**0.46** [±0.45]	$F(3,1065) =$ **227.50**	.389
2·EVALUATE STATEMENT	**-0.37** [±0.09]	**0.19** [±0.08]	**-0.27** [±0.17]	$F(3,378) =$ **97.18**	.431
3·INTERPRET AUTHOR'S ...	**-0.42** [±0.08]	-0.04 [±0.05]	0.26 [±0.26]	$F(3,668) =$ **35.76**	.135
4·LIST RECALL	**-0.51** [±0.06]	**-0.14** [±0.07]	NA[a]	$F(2,2600) =$ **207.20**	.137
5·EVALUATE STYLISTIC ...	**-0.33** [±0.10]	**0.35** [±0.08]	0.01 [±0.19]	$F(3,735) =$ **89.36**	.264
6·VERBAL PRODUCTION	**-0.31** [±0.24]	**0.27** [±0.22]	**-0.57** [±0.44]	$F(3,802) =$ **163.60**	.377
7·SELECT AND JUDGE	**-0.37** [±0.12]	**0.24** [±0.11]	-0.11 [±0.20]	$F(3,1018) =$ **124.10**	.266
8·EXPLAIN STORY ELEMENT	**-0.38** [±0.15]	0.10 [±0.13]	-0.01 [±0.30]	$F(3,952) =$ **69.55**	.177

[*] bold statistics are significant ($\alpha = .05$), *gt* = gender-specific type (1 = girls, 2 = boys), *c* = response correct, $R^2_{adj} = R^2$ adjusted, 95% confidence intervals in brackets
[a] no girl-specific type with incorrect responses

Table 6
Itemwise Gender Gap (Percent Correct) Distinguished by Real Gender and Gender-Specific Types

Item	#1	#2	#3	#4	#5	#6	#7	#8
Gap By Gender	14%	10%	3%	9%	7%	11%	1%	10%
Gap By Gender Type	70%	51%	22%	71%	23%	47%	57%	60%

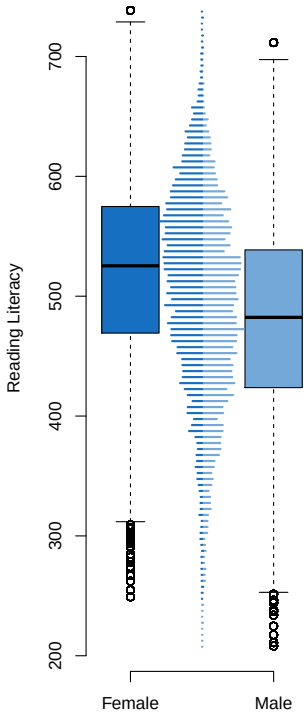

Figure 1: Reading Literacy Distributions of Males and Females

R118: Bullying

Bullying Text

PARENTS LACK AWARENESS OF BULLYING

Only one in three parents polled is aware of bullying involving their children, according to an Education Ministry survey released on Wednesday.

The survey, conducted between December 1994 and January 1995, involved some 19,000 parents, teachers and children at primary, junior and senior high schools where bullying has occurred.

The survey, the first of its kind conducted by the Ministry, covered students from the fourth grade up. According to the survey, 22 per cent of the primary school children polled said they face bullying, compared with 13 per cent of junior high school children and 4 per cent of senior high school students.

On the other hand, some 26 per cent of the primary school children said they have bullied, with the percentage decreasing to 20 per cent for junior high school children and 6 per cent for senior high school students.

Of those who replied that they have been bullies, between 39 and 65 per cent said they also have been bullied.

The survey indicated that 37 per cent of the parents of bullied primary school children were aware of bullying targeted at their children. The figure was 34 per cent for the parents of junior high school children and 18 per cent for those of the senior high school students.

Of the parents aware of the bullying, 14 per cent to 18 per cent said they had been told of bullying by teachers. Only 3 per cent to 4 per cent of the parents learned of the bullying from their children, according to the survey.

The survey also found that 42 per cent of primary school teachers are not aware of bullying aimed at their students. The portion of such teachers was 29 per cent at junior high schools and 69 per cent at senior high schools.

Asked for the reason behind bullying, about 85 per cent of the teachers cited a lack of education at home. Many parents singled out a lack of a sense of justice and compassion among children as the main reason.

An Education Ministry official said the findings suggest that parents and teachers should have closer contact with children to prevent bullying.

School bullying became a major issue in Japan after 13-year-old Kiyoteru Okouchi hanged himself in Nishio, Aichi Prefecture, in the fall of 1994, leaving a note saying that classmates had repeatedly dunked him in a nearby river and extorted money from him.

The bullying-suicide prompted the Education Ministry to issue a report on bullying in March 1995 urging teachers to order bullies not to come to school.

The article [...] appeared in a Japanese newspaper in 1996. Refer to it to answer the questions below.

Question 2: BULLYING

Why does the article mention the death of Kiyoteru Okouchi?

.....

.....

Figure 2: Sample PISA Stimulus and Question (OECD, 2006, p. 59–60)

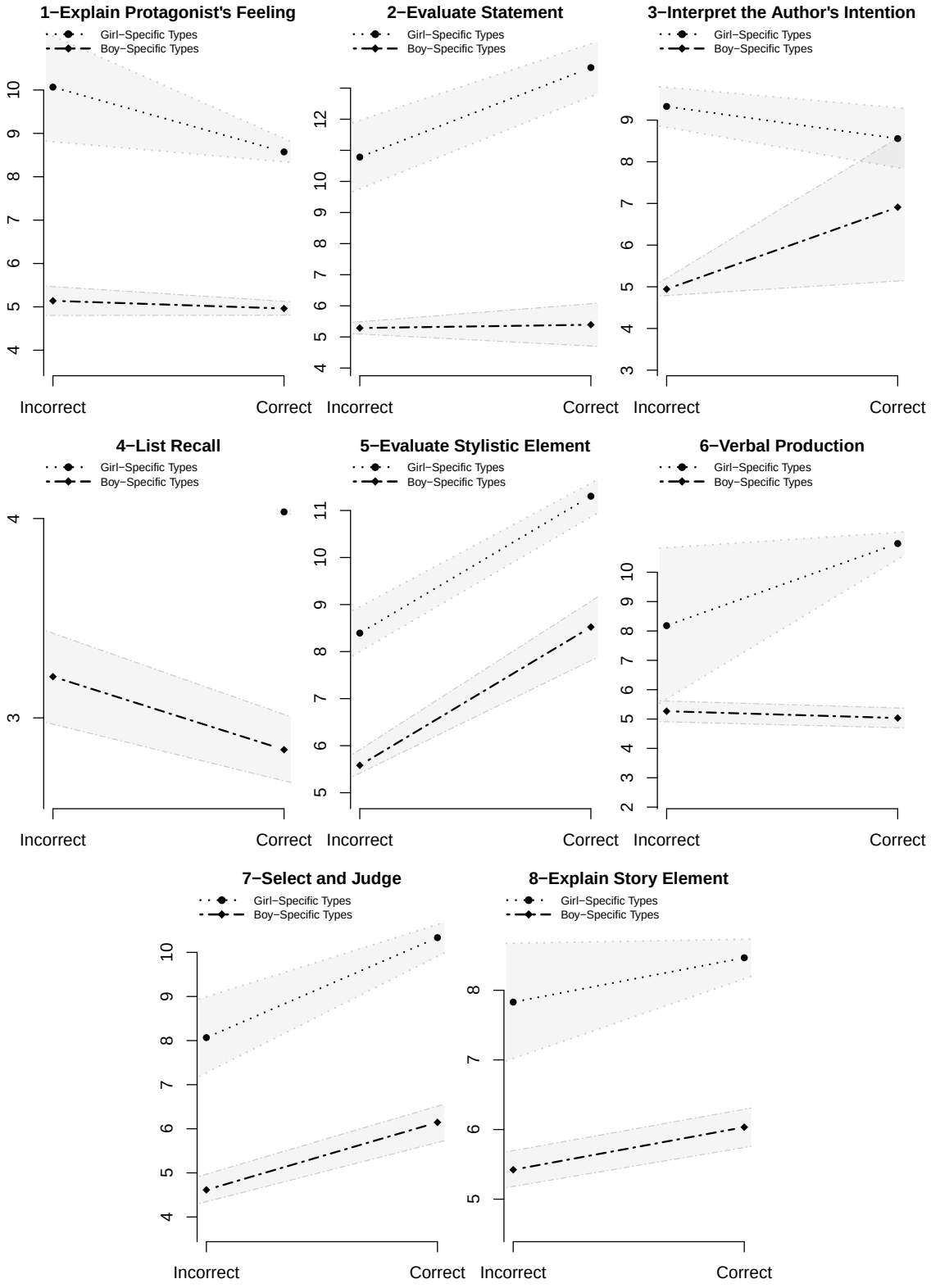

Figure 3: PEC by Gender-Specific Type and Response Correctness (with 95% confidence intervals)

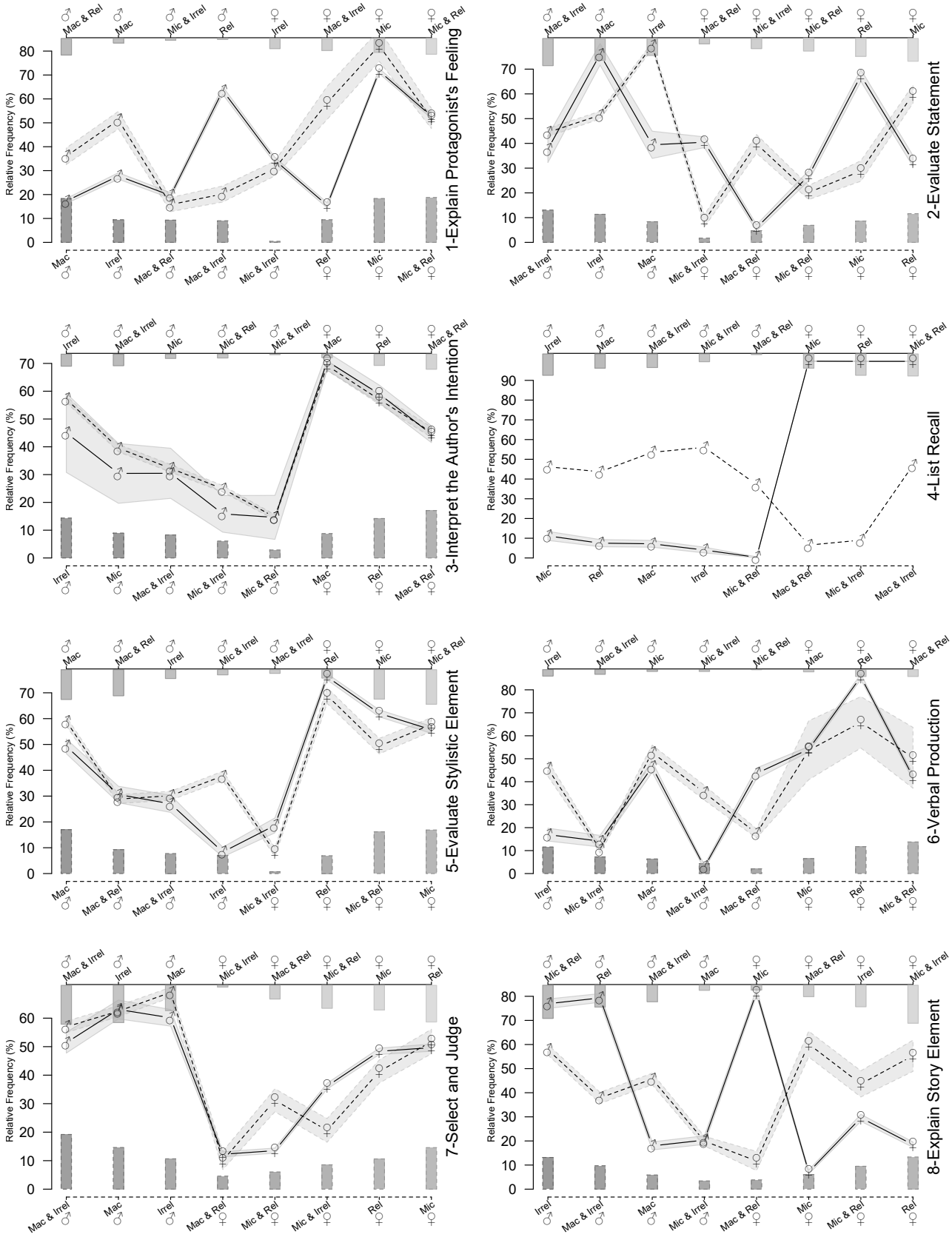

Figure 4: Micro and Relevance Measures by Gender-Specific Type and Response Correctness (with 95% confidence intervals)

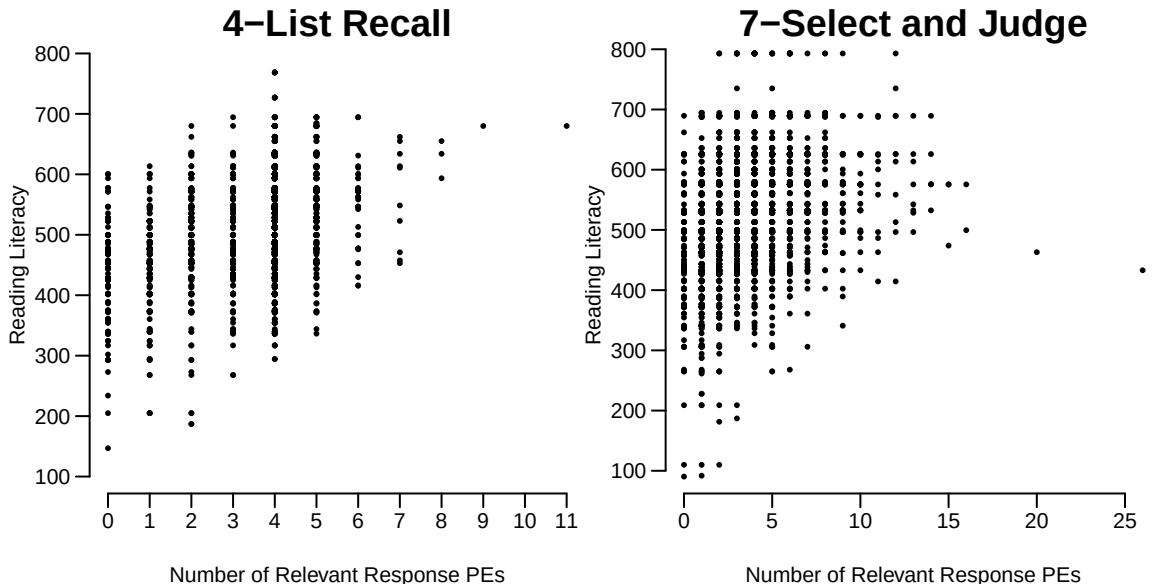

Figure 5: Relationship of Number of Relevant PE's within a Response and the Student's Overall Reading Literacy