Zehner *et al. Large-scale Assess Educ (2018) 6:7*
https://doi.org/10.1186/s40536-018-0060-3

Large-scale Assessments in Education

RESEARCH Open Access

# Automatically analyzing text responses for exploring gender-specific cognitions in PISA reading

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

*Correspondence:
fabian.zehner@dipf.de
[3] Centre for International Student Assessment (ZIB) e.V., Munich, Germany
Full list of author information is available at the end of the article
Fabian Zehner is now at the German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany. Christine Sälzer is now at the University of Stuttgart, Institute of Educational Science and Psychology (IfE), Germany.

## Abstract

**Background:** The gender gap in reading literacy is repeatedly found in large-scale assessments. This study compared girls' and boys' text responses in a reading test applying natural language processing. For this, a theoretical framework was compiled that allows mapping of response features to the preceding cognitive components such as micro- and macropropositions from the situation model.

**Methods:** In total, $n = 33,604$ responses from the German sample of the *Programme for International Student Assessment* (PISA) 2012 reading test have been analyzed for characterizing the genders' typical cognitive approaches. The analyses mainly explored the gender gap by contrasting groups of responses typical for either gender. These gender-specific responses characterize the typical responding of the genders to PISA reading questions.

**Results:** Responses typical for girls contained three to five more proposition entities from the situation model, irrespective of the response correctness. They integrated more relevant propositions and constituted better fits to the question focus. That means, in answering questions which ask for explicit information from the stimulus text, the typical girl responses appropriately encompassed more micropropositions, and typical boy responses tended to include more macropropositions—vice versa for questions requesting implicit information.

**Conclusion:** It appears that typical boy responses to PISA reading questions are characterized by struggling with retrieving and integrating propositions from the situation model. The typical girl liberally juggles these to formulate the responses. The results demonstrate that text responses are a neglected but informative source for educational large-scale assessments made accessible through natural language processing.

**Keywords:** Gender differences, Reading literacy, Automatic coding, Text responses

## Background

Reading literacy is characterized by remarkably consistent gender differences of relevant magnitudes (e.g., Mullis et al. 2012; NCES 2015; OECD 2015). While the causes for the differences remain unclear, literature suggests that open-ended responses in reading tests carry relevant information on successes and failures during the cognitive processes (Graesser and Franklin 1990; Kintsch 1998; Tourangeau et al. 2009). Thus, the present

study explores students' text responses in regard to gender differences. The literature suggests indicators that distinguish strong from weak readers and that these indicators could be observed in text responses.

A student's text response, such as *"The son wanted to help his fellow students"*, to a reading test question provides interesting insights about the respondent: Which information from the previously read text does the answer treat? Does the answer repeat the text or does it add information? And is this information relevant to correctly answer the question? Particularly in large-scale assessments, such linguistic features in responses have not been considered hitherto due to the mass of data and the resulting obstacles for reliable and objective coding. To overcome these, the present study analyzed the text responses to reading questions automatically by computer software adapted from Zehner et al. (2016). Accordingly, it also adds to the understanding of the gender gap in reading literacy. It identifies responses typical for either gender and contrasts their features. This way, the typical girl and boy in responding to reading questions are characterized.

The study compiles a theoretical framework about which linguistic features in the responses can be mapped to preceding cognitive processes determining success or failure in reading. That is, we regard responses as the outcomes of reading-associated cognitive processes and depict which indicators obtained from responses may reflect differences in the underlying cognitions having produced the response.

The *Programme for International Student Assessment* (PISA; OECD 2016) assesses, among others, the reading literacy of adolescents. 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 more than eighty countries and economies, to date. At the same time, the millions of text responses in many languages have not been an accessible source of information beyond scoring so far. Thus, recent innovations in natural language processing are an appealing methodological innovation for educational large-scale assessments.

The gender gap, as the use case of the methodological development presented here, is at the core of this study and refers to the comparison of means. Yet, a mean score is not representative of its entire distribution, and, like Stanat and Kunter (2002) observed, the subgroup distributions largely overlap. Hence, this large overlap shows that there is actually no gap between the genders, only between their means—which in turn is statistically significant and remarkably wide. That is why we, in this study, mainly conducted analyses at the level of typical responses for the genders—these typical, gender-specific responses are defined as groups of semantically similar responses significantly dominated by boys or girls, respectively. Obviously, this definition hands the decision on what is typical down to empirics. We use the term *gender* when an analysis splits the data by gender, and we use the term *gender-specific responses* and *typical boy or girl* when the analysis contrasts the genders' particularly typical responses characterizing the genders' distinct responding to PISA reading questions. Finally note that a student's gender is a distal predictor for reading literacy, whereas the cognitive approach is a proximal one. That is, the way the student attempts to solve the reading task is directly impacting the task response as well as the overall measure of their literacy, whereas the student's gender influences these only indirectly.

## Response features that are indicative for reading proficiency

In this section, we first illustrate exemplary PISA materials for contextualizing the subsequent subsections, which depict the cognitive processes in action when a student responds to a reading question. With this, we identify features that distinguish weak and strong readers.

### Sample PISA stimulus and question

For easier reading of the following subsections, Fig. 1 illustrates a prototypical PISA stimulus and question. They will serve as examples in the following subsections.

**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 referring to the text BULLYING**

Why does the article mention the death of Kiyoteru Okouchi?

**Fig. 1** Sample PISA Stimulus and Question (OECD 2006, p. 59–60). Note that this sample stimulus and this question were not part of the present study

Generally, the exemplary 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). The present study analyzed responses similar to this one to eight items referring to seven texts. Both, the sample text and question are representative for the included texts and items.

#### Theoretical framework: from reading to responding

This study endeavors to shed more light on the reasons for the observed gender difference. We, as researchers, observe it at the end of a long chain of events and conditions in the form of deviating means. For exploring the sources of the difference, the whole conglomerate—from reading the stimulus to responding to questions about it—needs to be contemplated. (I) The students come with their individual abilities and states into the testing situation. 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) 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 the students' cognitions are the original source for subgroup differences or only a mediator of other influences, the difference would always leave its imprint on the cognitions (III). Thus, the cognitions constitute a suitable level for investigating 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 and Potter 2005; Tourangeau et al. 2009; Wetherell 2007). Similarly, educational or psychological assessments typically address latent cognitive constructs. In most cases, the test takers' levels on the latent construct are extrapolated from their responses (V.c); for example, from their raw score (V.a). Thus, the connection between the responses and preceding cognitions is similarly apparent in the state of the art of assessment.

Equally to aligning instruments (II.b) and the interpretation of their outcomes to the theoretical framework of the construct (V.c), the respective operationalizations for mapping response features to cognitions need to be closely led by theoretical models, which are depicted in the following subsections. According to the *situation model* (Kintsch and 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 and 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 context.

#### *Reading cognitions—the Situation Model*

Kintsch and van Dijk (1978) introduced a holistic model to describe the cognitive processes involved while readers attempt to understand a text.[1] 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]). Thus, propositions are "a representation that more directly reflects the semantic relations that are crucial for how people understand, remember, and think with language" (Kintsch 1998, p. 49).

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 not explicitly stated but implied by the text. They can be inferred by the reader from micropropositions (bottom-up) as well as inferred or retrieved from the reader's knowledge (e.g., descriptive knowledge, schemata; top–down).

Good and bad readers differ in how easily they can access propositions from their memory and in how validly they reconstruct inaccessible information (Kintsch and 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. The usage of some type of proposition count, for diverse purposes, had been commonplace in the already referenced studies by Kintsch and van Dijk and has been established since then (e.g., Martín-Loeches et al. 2008; Olson et al. 1985; Walker and Kintsch 1985). As the second operationalization of crucial features in the situation model, the present study assessed whether a proposition comes from the micro- or macrostructures.

#### *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 requires 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. Second (cf. IV), the student needs to articulate the chosen solution. These phases interact with each other and can iterate several times.

[1] Subsequent works refined the situation model and culminated in Kintsch (1998), describing the *construction-integration model*. As the model's main features remained untouched and can serve as a fertile basis for the present study, the theoretical depiction focuses on the original work by Kintsch and van Dijk (1978).

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 and Clark 1985; Graesser and 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 and Franklin 1990). For example, for answering the sample item (cf. Fig. 1), the student needs to focus on the semantics *article mentions death of Kiyoteru*.

After having identified category and focus, the student's cognitions query corresponding knowledge structures in the memory relevant for answering (Graesser and Franklin 1990). The student's situation model of the stimulus serves as one such memory structure. 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 the situation model and QUEST come together. Briefly summarized, the knowledge structures comprise propositions, called statement nodes, that are winnowed down to those that are relevant and compatible to the question focus and category (Graesser and Franklin 1990). These final, semantically relevant propositions form the set of legitimate solutions for formulating the response. For the present study, accordingly, the third measure captured the semantic relevance of the expressed propositions for the question focus and category. This is in line with the definition of text passage relevance, which is "the degree to which a segment is germane to a specific task" (McCrudden and Schraw 2007, p. 114).

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 and Franklin 1990; McCrudden and Schraw 2007; van Dijk and Kintsch 1983). In the assessment context, the students ideally do their best to correctly answer the question. The entire process is not only influenced by the students' volition but, according to the MD-TRACE model (Rouet and Britt 2011), they also make decisions during reading—some *im*plicit, some *ex*plicit—determined by their task understanding. 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 and Clark 1985, p. 269). 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, syntax use, and engagement.

## Automatically extracting features from text responses

The preceding section 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. This section describes how these features can be automatically extracted at a conceptual level and depicts related work. The three features, picked from this complex process, meet two conditions—throughout, they are a crucial component of the process, and they are directly observable in the response. For example, relevance affects the process through the question focus when related knowledge structures are queried and also significantly aids to narrow down the information to the final set of legitimate solution propositions. At the same time, relevance is inherent to the decision whether a response is correct or not and, thus, is apparently observable.

### The basic processing: semantics and parts of speech

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 et al. (2016). As the first staging post, the software builds groups of semantically homogenous responses. For this, it 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 thus be reduced to 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 et al. 1990) to a text corpus. This step constitutes the core of the methodology proposed here. The result is 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 ("average") vector of all tokens for each response and clusters these in order to group responses into response types. Besides this procedure, 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 (POS; e.g., `NN` for *Japan*: *Normal Noun*). The Stuttgart–Tübingen Tagset (STTS; Schiller et al. 1999) was applied for analyzing German responses; see Table S1 in Additional file 1 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 ./$ .[2]

#### Approximating the extraction of propositions: Proposition Entities

For the following depiction, it is important to bear the following in mind. Many students who are assessed in low-stakes studies respond in a largely sloppy style. Such improper language entails severe difficulties for natural language processing (NLP). This is why we stick to rather coarse-grained methodological approaches, such as *stemming*—cutting of affixes instead of resolving the actual word stem—or the *bag of words*—a paradigm investigating words as separate units and, thus, neglecting word order and syntax. As high-stakes assessments bring about language in responses that complies more to the standard, more sophisticated NLP technologies can be used there. But the employed operationalizations in the present study needed to tolerate language deviating from the standard regularly. Hence, they are regarded as approximations of the indicators described in the theoretical framework.

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) in order 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 (predicates with arguments), 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 et al. 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 et al. 2014; Mohammad et al. 2014; Shrestha et al. 2015). Powerful approaches such as PropBank (Palmer et al. 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. For just this reason, a recent development in proposition extraction, so-called *Minimal Meaningful Propositions*, conceptually disregards "any implications, presuppositions, or entailments" (Godea et al. 2016, p. 3229). (3) Out of these numerous propositions, whether the propositions are relevant or not is furthermore dependent on the question focus, category, and stimulus context. In some contexts, the matter of relevance is also a subjective one (cf. the Error Analysis in Godea et al. 2016, p. 3231). 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,

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

*"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"* (was extreme). 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 S1 in Additional file 1 for the tag 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 and adverbs, which often describe subjects or actions.
- 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.

The approach of approximating single, specific words as propositions' elements is not entirely new. CPIDR is a system that similarly annotates texts with their parts of speech and interprets specific parts of speech as propositional ideas (Brown et al. 2008). Since CPIDR is tailored to English texts and centers around propositional quantity, which is only part of our study, the program was not suitable for our purposes.

In contrast to PEs, the potent concept of *Minimal Meaningful Propositions* (MMPs; Godea et al. 2016) refers to (intentionally) minimal but complete statements comprising several words. The MMP concept is implemented through, among others, syntactic parsing and aims at using further fine-grained technologies in the future such as identifying semantic roles of constituents and coreference resolution. All these are highly language dependent and reliant on proper language. This condition is not met by the improper language often prominent in low-stakes responses. The—in these terms—rather coarse-grained concept of PEs, requiring only tokenizing and POS tagging, is scalable to more languages and conditions, such as incorrect language. Note that we critically discuss the concept of PEs in "Limitations and directions".

### How to extract the three measures conceptually

As introduced before, the study reports three measures. The first one is the count of PEs for approximating 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 and 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 and 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 which exclusively comprise propositions relevant to the solution.

## Selected work on the reading literacy gender gap

As the proposed methodology's use case in this study is the investigation of the gender literacy gap, this section briefly reports a few central findings centering around the matter.

Particularly over the last three decades, research has been producing a huge body of findings about gender differences in reading literacy. In PISA, girls have been consistently outperforming boys (OECD 2002, 2004, 2007, 2010b, 2013, 2016). That is the case across all countries—with only two non-significant exceptions in 2000 and one in 2003—and across all cycles, constituting 348 comparisons in total. With the scale's standard deviation of 100 points, the gender gap average in the OECD ranged from 27 points in 2015, 31 points in 2000, 34 points in 2003, 38 points in 2006 and 2012, to 39 points in 2009.

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 and Linn 1988; Thorndike 1973; White 2007; Wolters et al. 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).[3] In the PISA 2009 data, when reading strategies and engagement are used as predictors, gender does not significantly account for any more variance in reading literacy (Artelt et al. 2010).

Because this study exclusively analyzed 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 until 2012 but lower in 2015: 34 points in 2000 (OECD 2002; Stanat and Kunter 2002), 42 points in 2003 and 2006 (Drechsel and Artelt 2007; OECD 2004, 2007; Schaffner et al. 2004), 40 points in 2009 (Naumann et al. 2010; OECD 2010b), 44 points in 2012 (Hohn et al. 2013; OECD 2013), and 21 points in 2015 (OECD 2016; Weis et al. 2016).

## Research questions

The preceding sections describe how good readers differ from poor readers. The presented analyses demonstrate the feasibility and usefulness of the proposed methodology by exploring the reading gender gap and answering two research questions along the lines of the proposed indicative features. (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- versus macropropositions and the extent to which these are relevant?

## Methods

### 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 et al. (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 1). In PISA 2012, *reading*, *math*, and *science* were assessed paper-based. Parts of the German text data had been transcribed for a previous study, thus, they constituted a unique source of information (for the detailed transcription procedure cf. Zehner et al. 2016). Also, the large oversample of ninth-graders (see above), in addition to the regular PISA sample, provided the analysis with additional information and power.

### 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 for which the transcribed response data were available (cf. the reading items in Zehner et al.

[3] 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.

**Table 1 Item characteristics**

| Item | Cat[a] | Aspect[b] | Correct (%) | *n* | ♀[c] (%) | Words[d] |
|---|---|---|---|---|---|---|
| 1. EXPLAIN PROTAGONIST'S FEELING | WHY<s> | B | 83 | 4152 | 49 | 12.3 (4.6) |
| 2. EVALUATE STATEMENT | ENABLE<a> | C | 43 | 4234 | 48 | 15.6 (9.0) |
| 3. INTERPRET THE AUTHOR'S INTENTION | SIG<a> | B | 10 | 4234 | 48 | 12.5 (6.3) |
| 4. LIST RECALL | CON<s> | A | 59 | 4223 | 50 | 5.6 (3.0) |
| 5. EVALUATE STYLISTIC ELEMENT | HOW<s> | C | 56 | 4234 | 48 | 14.7 (6.2) |
| 6. VERBAL PRODUCTION | WHN<a> | B | 80 | 4152 | 49 | 12.4 (6.9) |
| 7. SELECT AND JUDGE | ENABLE<s> | C | 68 | 4152 | 49 | 13.6 (7.0) |
| 8. EXPLAIN STORY ELEMENT | CON<e> | B | 69 | 4223 | 50 | 14.4 (5.5) |
| Total | | | 59 | 33,604 | 49 | 12.6 (6.1) |

The items refer to seven different stimulus texts. Only items 1 and 6 ask students about the same text

[a] Question category according to QUEST (Graesser and Clark 1985; Graesser and Murachver 1985); in the form FUNCTION<T> with FUNCTION∈{WHY, HOW, ENABLE, CONS, WHEN, WHERE, SIG, WHN} and T∈{*s*tate, *a*ction, *e*vent}

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

2016). These eight items refer to seven different stimulus texts and only the items 1. EXPLAIN PROTAGONIST'S FEELING and 6. VERBAL PRODUCTION ask about the same text. Due to repeated measurements in PISA, the item contents are confidential and cannot be described here. For the same reason, example responses cannot be published as they would disclose the item contents. Instead, the analyses characterize the responses using the linguistic measures. The eight items are listed in Table 1, each given a name for better traceability in the following. A sample stimulus text and question is presented in the previous section Sample PISA Stimulus and Question.

The items ranged from difficult (10%) to easy (83%; cf. Table 1). For estimating person ability, empty responses were coded as incorrect if reached in the test administration. The question category (sensu Graesser and Clark 1985) varied equally across items. According to the PISA framework (OECD 2013), the assessed cognitive aspect mainly varied between the two of three aspects *Integrate and Interpret* and *Reflect and Evaluate*. The third aspect, *Access and Retrieve*, was only represented by one item. Access and Retrieve items require the student to find a specific information explicitly stated in the text. Thus, the responses should basically include micropropositions. For the other two categories, micro- as well as macropropositions can be necessary for correct answering. The available data used in the present study only contained one Access and Retrieve item, which is not representative for the PISA assessment. Thus, the reporting only relates to the PISA framework categories when their characteristics are relevant to the matter.

Finally, 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. This measure was used to control for test motivation.

### Measures and analyses

In summary, the analyses used three measures to distinguish responses (cf. the section "Automatically extracting features from text responses Processing of Short Text Responses").

1. *Proposition Entity Count* (PEC): number of words annotated as a PE—i.e., nouns, pronouns, non-auxiliary verbs, adjectives, adverbs, answer particles
2. *Micro*: a PE's degree of similarity to the stimulus text and question
3. *Relevance*: a PE's degree of relevance for solving the item

Only PEs were included in the analysis of the micro and relevance measures. For this, PEs with different degrees of micro and relevance measures were counted: relevant and irrelevant micropropositions as well as relevant and irrelevant macropropositions. Propositions were classified as micropropositions if their micro similarity value ranged within the distribution's upper 25% of all PEs, as macropropositions otherwise. Analogously, propositions with a relevance value within the distribution's lower 25% of all PEs were classified as irrelevant, as relevant otherwise. This logic applied the conventional norm definition, with the middle 50% of a distribution constituting the average. That means, propositions needed to be at least somewhat dissimilar from the stimulus text and question in order to be macropropositions and at least somewhat similar to the reference responses in the coding guides in order 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 for correctly answering 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, 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 et al. (2016), except that the entire data was used as training data without cross-validation. The numbers of clusters were determined 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 gender ratio and 95% Wilson confidence intervals (CI) were 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—that is, all boy responses would contain the same semantics as would all girl responses; at the same time, the two groups' semantics would need to be informatively different. Rather, the procedure identifies homogeneous responses typical for boys or girls and then pools these gender-specific responses.

**Table 2 Gender gap (difference in percent correct) by real gender and gender-specific response types**

| Item | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 |
|---|---|---|---|---|---|---|---|---|
| Gap by gender (%) | 14% | 10% | 3% | 9% | 7% | 11% | 1% | 10% |
| Gap by type (%) | 70% | 51% | 22% | 71% | 23% | 47% | 57% | 60% |

This way, the groups of typical girl responses could contain some boy responses and vice versa. This is in line with the observation stated in the Introduction. An implication of only including response types dominated by boys or girls is that we contrasted particularly gender-specific responses and did not study mixed response types. Also, this decreased the sample size leading to lower testing power.

The difference between splitting by gender and automatically selected gender-specific response types becomes apparent in Table 2, which illustrates the itemwise gender gap by split criterion. Please see Table S2 in Additional file 2 for the percentage of students included in gender-specific response types. If split by gender, the girls' advantage of solving the items ranged from 1% to 14%. If split by gender-specific response types, the advantage notably increases, ranging from 22% to 71%.

All analyses are reported per item because responses depend highly on the corresponding item. Most analyses controlled for the response correctness based on judgments by PISA's human coders. This way, the reported effects are not confounded by the fact that girls respond correctly more often.

All analyses excluded empty responses. Apart from reflecting differences in cognitive approaches as described in the Results, the PISA reading gender gap was also crucially influenced by the large number of boys who did not respond at all. For all analyzed items, there were significant proportions of empty responses by boys. On average, they produced 63% of the empty responses across the eight items ($SD = 7\%$).

In the presented data, PEC correlated negatively by $r = -0.18$ with micro and by $r = -0.15$ with relevance. Opposed to that, micro and relevance correlated moderately by $r = 0.38$.

### Software

The employed software implements open software, libraries, and packages. First, the software builds a database storing a Wikipedia dump using *JWPL* (Zesch et al. 2008). *DKPro Similarity* (Bär et al. 2013), which in turn primarily utilizes *S-Space* (Jurgens and 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 and Lally, 2004). For stemming, *Snowball* (Porter 2001) is used, and for POS tagging, the German Stanford NLP parser with the PCFG model (Rafferty and Manning 2008) is employed. The parser also annotates POS tags by rating the likelihood of candidate syntax trees and returning one tag for each word from the highest rated tree (Klein and Manning 2003). For statistical matters, the software evokes *R* (R Core Team 2016), which was also used for further statistical analyses, partly using the packages *binom* for computing ratios' confidence intervals (Dorai-Raj 2014) and *doSNOW* for parallelizing computations (Revolution Analytics and Weston 2014).

**Table 3 Proposition Entity Count (PEC) by gender and response correctness**

| Item | $\beta_g$ | $\beta_c$ | $\beta_{g*c}$ | *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* | 0.023 |
| 2. EVALUATE STATEMENT | *− 0.12* [± 0.05] | *0.18* [± 0.05] | *− 0.15* [± 0.15] | *F*(3,3365) = *55.42* | 0.046 |
| 3. INTERPRET AUTHOR'S ... | *− 0.20* [± 0.04] | 0.03 [± 0.05] | − 0.03 [± 0.15] | *F*(3,2989) = *43.90* | 0.041 |
| 4. LIST RECALL | *− 0.12* [± 0.09] | *0.23* [± 0.05] | 0.14 [± 0.20] | *F*(3,3718) = *97.61* | 0.072 |
| 5. EVALUATE STYLISTIC ... | *− 0.15* [± 0.05] | *0.36* [± 0.04] | − 0.02 [± 0.14] | *F*(3,3540) = *226.30* | 0.160 |
| 6. VERBAL PRODUCTION | *− 0.16* [± 0.09] | *0.14* [± 0.05] | 0.01 [± 0.21] | *F*(3,3959) = *70.56* | 0.050 |
| 7. SELECT AND JUDGE | *− 0.18* [± 0.06] | *0.23* [± 0.05] | 0.01 [± 0.15] | *F*(3,3764) = 128.90 | 0.092 |
| 8. EXPLAIN STORY ELEMENT | *− 0.17* [± 0.06] | *0.11* [± 0.05] | 0.03 [± 0.15] | *F*(3,3893) = *55.41* | 0.040 |

Italic 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

## Results

Structured by the research questions, this section first reports (A.I) how the genders differed in their PEC, (A.II) how these results changed when only gender-specific responses were included, and (B) the differences in the relevance and micro-/macrolevel of the PEs.

### Proposition Entity Count (PEC)

Prior to the substantive analyses of interest, PEC was checked to not just constitute an outcome of test motivation, mediated through the response length. This way, PEC turned out to be only marginally negatively related to test motivation, with $r$ ranging from − 0.02 for 6. VERBAL PRODUCTION to − 0.11 for 7. SELECT AND JUDGE.

#### *PEC by gender (A.I)*

Gender affected the number of PEs for every item significantly (cf. Table 3), 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 it is *not* caused by the fact that more girls gave correct responses and correct responses were associated with more 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 gave even more PEs when responding correctly to this item. For all items, a significant overall variation was found. Gender, the response correctness, and the respective interaction explained $R^2_{adj} = 2\%$ up to $R^2_{adj} = 16\%$ of PEC's variance.

#### *PEC by gender-specific response types (A.II)*

Prior to this analysis, cluster analyses were carried out for identifying gender-specific response types; see "Methods" for the rationale and Table S2 in Additional file 2 for the resulting cluster solutions. Only responses assigned to significant gender response types were included. As obvious in Table 4, the effect sizes of gender-specific response types on the PEC were larger than the pure gender effects in the models presented in the

**Table 4 Proposition Entity Count (PEC) by gender-specific response types and response correctness**

| Item | $\beta_{gt}$ | $\beta_c$ | $\beta_{gt*c}$ | $F$(df1, df2) | $R^2_{adj}$ |
|---|---|---|---|---|---|
| 1. EXPLAIN PROTAGONIST'S ... | *− 0.84* [± 0.21] | *− 0.19* [± 0.15] | *0.46* [± 0.45] | $F(3,1065) = 227.50$ | 0.389 |
| 2. EVALUATE STATEMENT | *− 0.37* [± 0.09] | *0.19* [± 0.08] | *− 0.27* [± 0.17] | $F(3,378) = 97.18$ | 0.431 |
| 3. INTERPRET AUTHOR'S ... | *− 0.42* [± 0.08] | − 0.04 [± 0.05] | 0.26 [± 0.26] | $F(3,668) = 35.76$ | 0.135 |
| 4. LIST RECALL | *− 0.51* [± 0.06] | *− 0.14* [± 0.07] | NA[a] | $F(2,2600) = 207.20$ | 0.137 |
| 5. EVALUATE STYLISTIC ... | *− 0.33* [± 0.10] | *0.35* [± 0.08] | 0.01 [± 0.19] | $F(3,735) = 89.36$ | 0.264 |
| 6. VERBAL PRODUCTION | *− 0.31* [± 0.24] | *0.27* [± 0.22] | *− 0.57* [± 0.44] | $F(3,802) = 163.60$ | 0.377 |
| 7. SELECT AND JUDGE | *− 0.37* [± 0.12] | *0.24* [± 0.11] | − 0.11 [± 0.20] | $F(3,1018) = 124.10$ | 0.266 |
| 8. EXPLAIN STORY ELEMENT | *− 0.38* [± 0.15] | 0.10 [± 0.13] | − 0.01 [± 0.30] | $F(3,952) = 69.55$ | 0.177 |

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

[a] No girl type with incorrect responses

previous section. For the smallest effect size in 5. EVALUATE STYLISTIC ELEMENT, on average, typical girl responses contained 2.8 PEs more than typical boy responses did. For 1. EXPLAIN PROTAGONIST'S FEELING, the effect corresponds to 4.9 PEs, the largest effect size. These effect sizes appeared controlling 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. The interaction effect between the response correctness and gender-specific response type was significant for the items 1. EXPLAIN PROTAGONIST'S FEELING, 2. EVALUATE STATEMENT, and 6. VERBAL PRODUCTION. Refer to Fig. 2 for all items' group means and 95% CI. Overall, the models that included the gender-specific response 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 purely 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 for answering correctly. Again, only the gender-specific response types were included. Figure 3 plots the relative frequencies of relevant and irrelevant micro- and macropropositions within the 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 contained 19% 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 types. On the other side, the lower bar on the very left shows

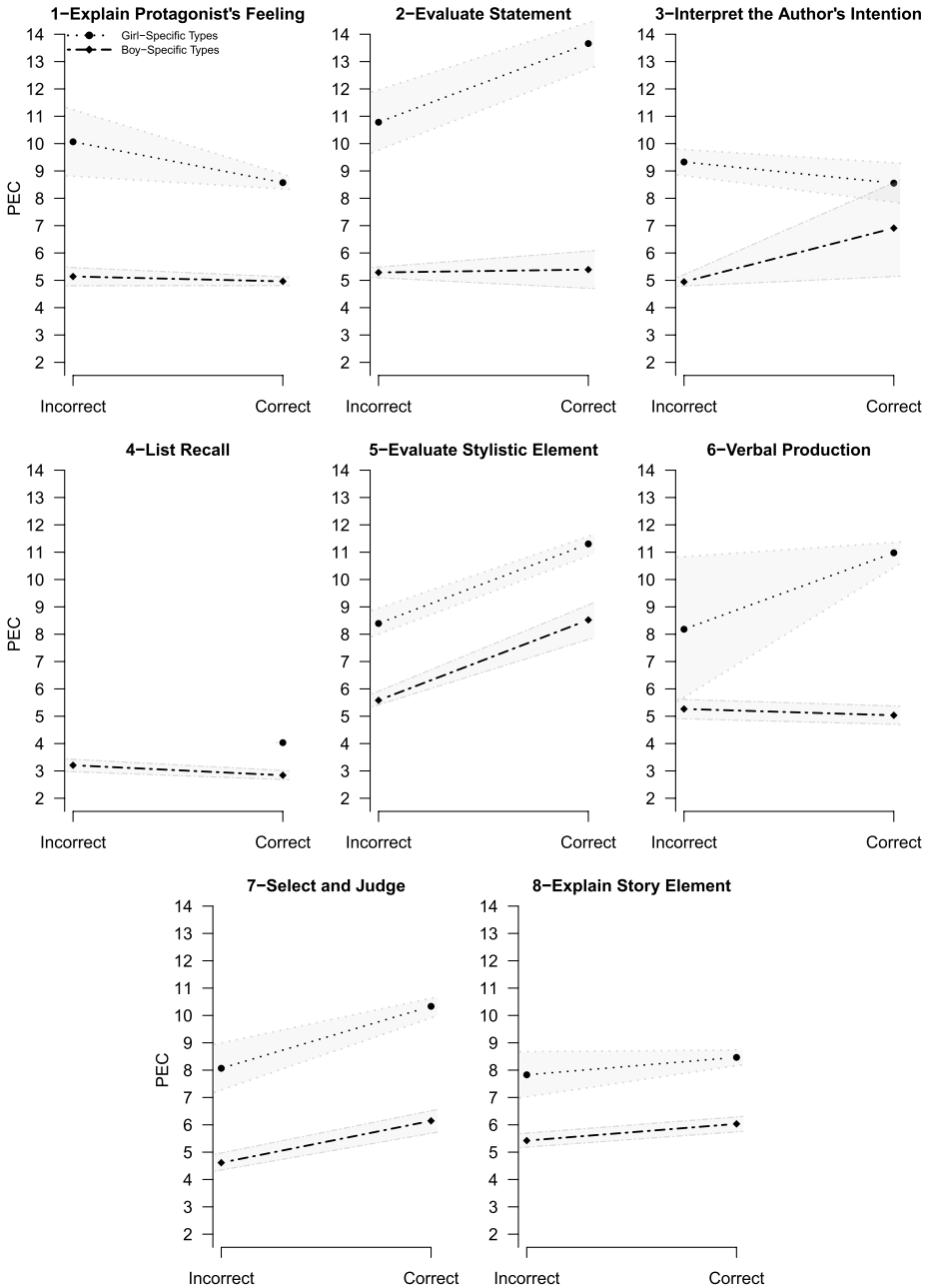


**Fig. 2** PEC by gender type and response correctness (with 95% confidence intervals)

that incorrect boy-specific response types included 18% more macropropositions within their responses (*Mac*). The Mars symbol below the bar indicates that the deviation given by the bar is dominated by typical boy responses. 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% more macropropositions in incorrect boy-specific responses are the result of these responses comprising 36% macropropositions. The analogous information for the *correct* responses is shown by the

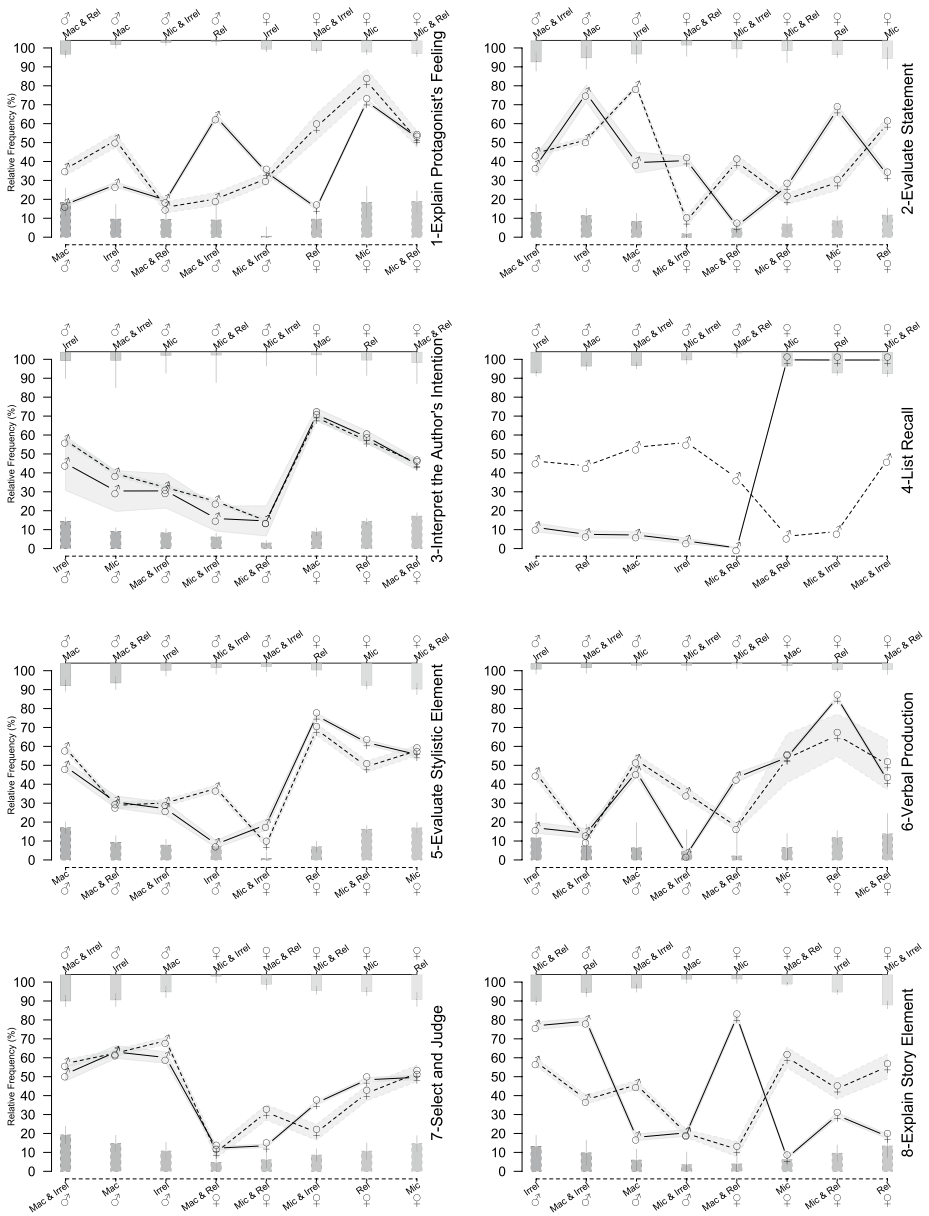


**Fig. 3** Micro and relevance measures by gender type and response correctness (with 95% confidence intervals)

upper, solid abscissa and the solid line. Both lines are encapsulated by their 95% confidence intervals.

The figures for the eight items reveal the following. All items showed marked deviations across the gender-specific response types overall—nearly in all correct as well as incorrect responses. Only the correct responses to three items differed barely (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 incorrect responses incorporated about 12% more *irrelevant* PEs. The same figure can be found for item 3. INTERPRET THE AUTHOR'S INTENTION where incorrect girl-specific responses contained 17% more *relevant macropropositions* and boy-specific ones tended to comprise *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 included PEs from the *microstructures* (also *relevantly*), while boy-specific responses consisted of 18% more *macropropositions* (36% in total). Hence overall, the typical girl responses referred to *relevant* PEs and the more appropriate level in the situation model—irrespective of 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.

## Discussion

The present study followed two interests. First, a theoretical framework along with the required technologies were assembled in order to make text responses in large-scale assessments accessible to studies as a new source of information. Second, demonstrating the proposed methodology's added value, the study sheds further light on the gender differences in reading literacy.

The theoretical framework sketches how a student builds a mental representation of a stimulus and how they craft it into a response answering a question. The situation model (van Dijk and Kintsch 1983) comprises propositions spread across micro- and macrostructures, varying in how close they are to the text base. Attempting to answer a test question, the student identifies the question focus and category (sensu QUEST; Graesser and Franklin 1990). The student uses the situation model as one source of several to gather potential propositions that would help to answer the question. Next, the student narrows down the propositions to fit 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 and Murachver 1985; Graesser and Clark

1985). In this paper, we suggest that specific properties of the resulting text response can be mapped to successful or flawed processes. This is done via automatic processing of the responses (Zehner et al. 2016) by extracting proposition entities (PEs), part-of-speech tagging, and semantic comparisons using Latent Semantic Analysis (Deerwester et al. 1990).

The PEC analyses and the gender gap's increase induced by the gender-specific response types illustrated that it is not reasonable to assume two distinct and homogeneous cognitive types for responding to PISA reading questions separated by gender. Rather, there appear to be different cognitive types in the reading literacy context that are *unevenly* spread across genders. The reported findings show how the typical responses for either gender characterize the genders in responding. Therefore, the typical boy is characterized by parsimonious selection of PEs. In the analyzed items, he used three to five PEs less on average than the typical girl. This only constitutes the effect size within either correct or incorrect responses, while the phenomenological effect is higher because boys respond more often incorrectly to the investigated items.

In some cases, the typical boy's cognitive frugality suffices to produce correct responses; on top, this can seem very concise and worthwhile. However, this positive perspective can be flipped in light of evidence for the majority of the analyzed items because the boy-specific characteristics were relatively consistent across correct as well as incorrect responses. In conclusion, the typical boy seems to involve a less stable situation model and to struggle with retrieving and inferring from it. The higher number of PEs in typical girl responses emphasizes that the typical girl liberally juggles the information in the situation model. Plus, the analyses indicate that PEs used in typical girl responses are not simple recalls of random information from the stimulus, but they nearly always use more PEs that are relevant and compatible to the question focus than the typical boy responses do. The relatively few PEs typical boy responses tend to contain are more often irrelevant than those in responses typical for girls—regardless of the response correctness for almost all items. This figure adds to the previous notion that the typical boy has difficulties accessing and searching the relevant parts of the situation model when responding to PISA reading questions.

Given a required threshold of engagement, readers who realize a gap or inconsistency in their situation model try to reconstruct the missing information (Kintsch and van Dijk 1978). This reconstruction phenomenon occurs throughout the data, but when this process fails, it is most often associated with responses that are typical for boys. 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 typical boy responses. For example in 4. LIST RECALL, students need to directly copy four terms from the stimulus to their response. This task is a mock-up for evoking the recall of micropropositions. While even the correct boy-specific responses consist of 10% of (irrelevant) macropropositions, the corresponding incorrect responses incorporate 54% 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% of micropropositions, of which 63% are classified as irrelevant. The bottom line is, the typical girl seems 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% more relevant macropropositions than girl-specific ones do. At the same time, typical girl responses incorporate 13% more relevant micropropositions. There seem to be two equivalently legitimate lines of reasoning predominantly used in responses typical for one of the genders. Notably, the gender gap for this item is the second smallest and to a large part stems from empty boy responses.

In conclusion, the typical boy is weak in determining whether a question relates to the text's micro- or macrolevel and in determining the corresponding relevant information. Both could be addressed in interventions focusing on relevance instruction. With respect to the taxonomy suggested in the overview by McCrudden and Schraw (2007), *general*—opposed to *specific—relevance instruction* could constitute one way to improve reading literacy among students showing these types of responses. Automatic processing of text responses would be an attractive way for identifying such students at a large scale. In the next steps, experimental studies need to investigate the mechanisms between the cognitions of interest and the resulting text responses, for this will show which additional indicators are available in the responses and which intervention might be most appropriate for which antecedents and circumstances.

With respect to the extracted measures' intercorrelations, there was a small tendency that the more information students included, the higher the share of irrelevant information and the more they wrote semantics that were not included in the stimulus. Moreover, there was a moderate tendency that higher proportions of text and question rephrasing went together with higher proportions of relevant information. Finally, test motivation was only marginally related to PEC. Partly, this might have come from the overall test motivation measure not directly targeting the reading assessment. On the other hand, this shows the PEC to be an independent measure providing 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 assessed in PISA 2009 by student self reports were *memorization*, *elaboration*, and *control strategies* (OECD 2010a). These strategies represent different (meta-)cognitive approaches. They would directly result in corresponding distinct response types like described in this paper. The memorization strategy refers to the degree of retentivity of micropropositions, while elaboration aims at integrating different propositions and deducing relevant information from them, which would result in macropropositions. Since the PISA 2009 text 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 PISA 2018 data. 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 replication for other languages. Furthermore, large-scale studies only deliver correlative data, enabling the production of new hypotheses. But the causality of the involved mechanisms need to be tested in additional experimental studies. The substantive findings of this secondary analysis are constrained to the setting of the German PISA 2012 reading assessment, and whenever we refer to the "typical girl" or "typical boy (responses)" in this paper, this is the short—more legible—form of *particularly typical, observed response behavior for either gender in the PISA reading assessment.* For the relevance measure, the PISA coding guides were used as the benchmark in lack of a better gold standard. As previously shown (Zehner et al. 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. Although the typological approach taken in this study reveals new insights as evidenced by this study, it requires a more detailed theoretical depiction in the future. Furthermore, the *micro*/*macro* and *relevant*/*irrelevant* classification of PEs applied a norm-referenced comparison. This means, the classification is (i) dependent on the sample and, thus, (ii) not criterion-referenced, (iii) within-comparisons like gender gap analyses are legitimate, but (iv) the absolute values cannot be interpreted without further knowledge about the measures and other samples.

When PISA test takers go back to the stimulus text for retrieving information, they obviously cannot recall those from memory but need to construct the mental model again. For doing so, additional processes take place, such as skimming (Goldman and Saul 1990) and selecting the relevant information (sensu Rouet and Britt 2011). However, this case is not an exception in that the construction of the mental model, possibly an incomplete one, is a prerequisite for successfully answering the question. Thus, we consider the described basic processes to still take place then. The data's enrichment by process indicators collected in computer-based assessments will allow to identify these situations for extending the theoretical framework and the operationalizations in the future. Process data constitutes empirical information about the cognitive states and related behavior (e.g., reading strategies) mediating the construct's effect on the task product (Goldhammer and Zehner 2017).

The operationalizations in this study 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. 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 selected for operationalization in this study were the most informative to the matter of the reading literacy gender gap and at the same time feasible for the NLP techniques.

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 of 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 and adverbs (`ADJA`, `ADJD`, `ADV`; e.g., *appropriate*) are considered PEs because they typically add genuine information 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 allowing a more reliable measurement. Also, some of the described problems could be overcome by means of a constituent analysis.

In the future, it will be interesting to challenge the methodological approach presented by currently establishing developments such as *word2vec* (Mikolov et al. 2013a) or *GloVe* (Pennington et al. 2014). These are modern, but similar technologies to the Latent Semantic Analysis used here. At the core, they also use vectorial representations of words for estimating word similarities, and, among others, they achieve this by analyzing word co-occurrences and applying dimensionality reduction. However, they utilize more scalable and powerful neural networks instead of a straight-forward singular value decomposition and allow several task-specific optimizations (e.g., Mikolov et al. 2013b). Thus, it will be interesting to see whether the approach presented in this paper can be further optimized by using vectors computed by means of alternate models than Latent Semantic Analysis.

## Conclusion

In sum, our study showed that open-ended student responses contain several elements that help understand the frequently found gender gap in reading achievement. No matter being correct or incorrect, text responses to PISA reading questions that were typical for boys contained fewer bits of information, more irrelevant information, and more often addressed the wrong level of information than text responses typical for girls did.

Also, the results indicated that referring to cognitive types dominated by either of the genders is more appropriate than assuming homogeneously separated genders. Despite the listed constraints, it appears worthwhile to utilize nascent innovations in natural language processing to investigate text responses as a new source of information in educational 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.

## Additional files

**Additional file 1: Table S1.** Stuttgart–Tübingen Tagset (Schiller et al. 1999; adapted with slight changes from ISOcat 2008).
**Additional file 2: Table S2.** Cluster solutions for analyses at the level of gender-specific response types.

### Abbreviations

CI: confidence interval; NAEP: National Assessment of Educational Progress; NLP: natural language processing; OECD: Organisation for Economic Co-operation and Development; PCFG: probabilistic context free grammars; PE: proposition entity; PEC: proposition entity count; PIRLS: Progress in International Reading Literacy Study; PISA: Programme for International Student Assessment; POS tagging: part-of-speech tagging.

### Authors' contributions

FZ provided the research question and principal analysis, FG and CS contributed to how to approach the research question in analyses. FZ wrote the manuscript and FG and CS revised it. All authors read and approved the final manuscript.

### Author details

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

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note



Received: 28 October 2017 Accepted: 4 July 2018
Published online: 23 July 2018

### References

Artelt, C., Naumann, J., & Schneider, W. (2010). Lesemotivation und Lernstrategien: Bilanz nach einem Jahrzehnt. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009* (pp. 73–112). Münster u.a.: Waxmann.

Bär, D., Zesch, T., and 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: Association for Computational Linguistics.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, *40*(2), 540–545.

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.

Dorai-Raj, S. (2014). binom: Binomial confidence intervals for several parameterizations. R package, Version 1.1–1.

Drechsel, B., & Artelt, C. (2007). Lesekompetenz. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, & R. Pekrun (Eds.), *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* (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.

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.

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: Erlbaum.

Godea, A., Bulgarov, F., and Nielsen, R. (2016). Automatic generation and classification of Minimal Meaningful Propositions in educational systems. In for Computational Linguistics, A. (Ed.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (pp. 3226–3236).
Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132.
Goldman, S. R., & Saul, E. U. (1990). Flexibility in text processing: A strategy competition model. *Learning and Individual Differences*, *2*(2), 181–219.
Graesser, A . C., & Clark, L . F. (1985). *Structures and procedures of implicit knowledge, volume 17 of advances in discourse processes*. Norwood: Ablex.
Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*(3), 279–303.
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., Flor, M., Madnani, N., Tetreault, J. R., Blanchard, D., Napolitano, D., Lee, C. M., and 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 in 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. *Communications of the ACM*, *57*(1), 94–103.
Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69.
Jurgens, D. and Stevens, K. (2010). The S-Space package: An open source package for word space models. In Association for Computational Linguistics (Ed.), *Proceedings of System Demonstrations 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 30–35).
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.
Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In Hinrichs, E. W. and Roth, D. (Eds.), *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, pages 423–430, Morristown, NJ: Association for Computational Linguistics.
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.
McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, *19*(2), 113–139.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119.
Mohammad, S., Zhu, X., and Martin, J. (2014). Semantic role labeling of emotions in tweets. In for Computational Linguistics, A. (Ed.).In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41.
Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill: Boston College.
Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009* (pp. 23–71). Münster u.a.: Waxmann.
NCES (2015). The nation's report card: 2015 mathematics and reading assessments.
OECD. (2002). *Reading for change: Performance and engagement across countries*. Paris: OECD Publishing.
OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD Publishing.
OECD. (2006). *PISA released items—reading*. Paris: OECD.
OECD. (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1: Analysis*. Paris: OECD Publishing.
OECD. (2010a). *PISA 2009 results: Learning to learn - student engagement, strategies and practices (volume III). PISA 2009 Results*. Paris: OECD Publishing.
OECD. (2010b). *PISA 2009 Results: What students know and can do: Student performance in reading, mathematics and science (Volume I)*. Paris: OECD Publishing.
OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.
OECD. (2014). *PISA 2012 results: What students know and can do (volume I, revised edition, February 2014)*. Paris: OECD Publishing.
OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence. PISA*. Paris: OECD Publishing.
OECD. (2016). *PISA 2015 Results* (Vol. 1). Paris: OECD Publishing.
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: Erlbaum.
Oranje, A. (2006). *Confidence intervals for proportion estimates in complex samples, volume RR-06-21 of Research report*. Princeton: ETS.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.
Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
Porter, M. (2001). Snowball: A language for stemming algorithms.
Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.
R Core Team (2016). R: A language and environment for statistical computing.
Rafferty, A. N. and Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, (pp. 40–46).
Revolution Analytics, & Weston, S. (2014). doSNOW: Foreach parallel adaptor for the snow package. R package, Version 1.0.16.
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: Information Age Pub.
Schaffner, E., Schiefele, U., Drechsel, B., & Artelt, C. (2004). Lesekompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland; Ergebnisse des zweiten internationalen Vergleichs* (pp. 93–110). Münster: Waxmann.
Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Stuttgart: Technical report, University of Stuttgart and University of Tübingen.
Shrestha, N., Vulić, I., & Moens, M. F. (2015). Semantic role labeling of speech transcripts. Lecture notes in computer science. In A. Gelbukh (Ed.), *Lecture notes in computer science: Computational linguistics and intelligent text processing* (Vol. 9042, pp. 583–595). Berlin: Springer.
Stanat, P., & Kunter, M. (2002). Geschlechterspezifische Leistungsunterschiede bei Fünfzehnjährigen im internationalen Vergleich. *Zeitschrift für Erziehungswissenschaft*, *4*(1), 28–48.
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: Cambridge University Press.
Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study: 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* (10th ed.). Cambridge: Cambridge University Press.
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.
Weis, M., Zehner, F., Sälzer, C., Strohmaier, A., Artelt, C., & Pfost, M. (2016). Lesekompetenz in PISA 2015: Ergebnisse, Veränderungen und Perspektiven. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Eds.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (pp. 249–283). Münster: Waxmann.
Wetherell, M. (2007). A step too far: Discursive psychology, linguistic ethnography and questions of identity. *Journal of Sociolinguistics*, *11*(5), 661–681.
White, B. (2007). Are girls better readers than boys? Which boys? Which girls? *Canadian Journal of Education/Revue canadienne de l'éducation*, *30*, 554–581.
Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212.
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.
Zehner, F., Goldhammer, F., and 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)*.
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.
Zesch, T., Müller, C., and 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.