



Determinants of grader agreement: an analysis of multiple short answer corpora

Ulrike Padó¹ · Sebastian Padó²

Accepted: 28 May 2021 / Published online: 12 June 2021
© The Author(s) 2021

Abstract The 'short answer' question format is a widely used tool in educational assessment, in which students write one to three sentences in response to an open question. The answers are subsequently rated by expert graders. The *agreement between these graders* is crucial for reliable analysis, both in terms of educational strategies and in terms of developing automatic models for short answer grading (SAG), an active research topic in NLP. This makes it important to understand the properties that influence grader agreement (such as question difficulty, answer length, and answer correctness). However, the twin challenges towards such an understanding are the wide range of SAG corpora in use (which differ along a number of dimensions) and the hierarchical structure of potentially relevant properties (which can be located at the corpus, answer, or question levels). This article uses generalized mixed effects models to analyze the effect of various such properties on grader agreement in six major SAG corpora for two main assessment tasks (language and content assessment). Overall, we find broad agreement among corpora, with a number of properties behaving similarly across corpora (e.g., shorter answers and correct answers are easier to grade). Some properties show more corpus-specific behavior (e.g., the question difficulty level), and some corpora are more in line with general tendencies than others. In sum, we obtain a nuanced picture of how the major short answer grading corpora are similar and dissimilar from which we derive suggestions for corpus development and analysis.

✉ Ulrike Padó
ulrike.pado@hft-stuttgart.de

Sebastian Padó
sebastian.pado@ims.uni-stuttgart.de

¹ Hochschule für Technik, Stuttgart, Germany

² Universität Stuttgart, Stuttgart, Germany

Keywords Educational applications of NLP · Short answer grading · Annotation reliability · Hierarchical generalized linear models

1 Introduction

One application of Natural Language Processing (NLP) in education is automated Short-Answer Grading (SAG), the task of assigning grades to short free-text answers (typically a few sentences long) given by students in response to open-ended questions (see, e.g., the SAG task in the 2013 SemEval competition, Dzikovska et al., 2013).

Short-answer questions are interesting from a didactic point of view because they allow language teachers to assess their students' language production skills. In both language and content instruction, reading the answers constructed by the students also yields better insights into students' thought processes and reasoning than closed question formats like multiple choice or highly-constrained ones like fill-in-the-blank tasks, where there is little chance for the teacher to understand why an answer was given. The short-answer question type has sufficient flexibility to accommodate all subject areas as well as different cognitive tasks for the student, from reproduction to application and even analysis (cf. Bloom's Taxonomy, revised by Anderson & Krathwohl, 2014). Table 1 shows some examples of short-answer questions and corresponding student answers, illustrating this flexibility: The first example is from content-matter assessment in a science setting and probes data interpretation skills. The second is a reading comprehension question in language instruction, referring to a text titled "Reporting the Facts".

A drawback of short-answer questions is the time consuming nature of grading, since each answer has to be read and considered individually by the teacher. Generating objective and reliable grades requires care and attention (cf. the significant efforts of standardized test providers invested into grading, Way et al., 2008; Baldwin et al., 2008). This is a major motivation for the use of automated methods, which help save human effort (see Sect. 2).

This strand of research has profited from a wide variety of corpora freely available for research purposes which reflect the whole breadth of educational use of short-answer questions: There are corpora of content assessment and language assessment questions, collected from native or near-native speakers or from language learners, and in different target languages (see Sect. 3). In addition, grading schemes and question difficulty vary across corpora.

This variability comes at a cost, specifically for the general applicability of methods, and the general understanding of effects. On the method level, Horbach and Zesch (2019) show that the differences between corpora have a substantial impact on the performance of machine learning systems for automated SAG. They find the clearest patterns of influence from answer length and the number of grading levels available, as well as large variation between individual questions.

At an even more fundamental level, the reliability of the *human grades* assigned to the student answers has a substantial impact. Due to the different collection and

Table 1 Short Answer Questions from two corpora (ASAP: content assessment, CREE: language assessment) with two student answers each (given verbatim) and manual graders' assessments

Corpus	Question	Answers	Rating
ASAP	A student performed the following investigation to test four different polymer plastics for stretchability. [...] Draw a conclusion from the student's data	Plastic type B can stretch the most, while A has the most tensile strength because it stretched [sic] the least	Correct
		My conclusion is that the second trial went better than the first because the numbers are much larger than the first	Incorrect
CREE	If quotes in the newspaper are not really quotes, then what are they?	It is actually a summary of what was said.	Correct
		Misquotes	Incorrect

grading conditions across the range of corpora in frequent use in the research community, we see differences in grader agreement (see Table 2 below). Therefore, any machine learning effort on the data that uses the resulting grades as target annotation is directly affected. While machine learning algorithms can tolerate a certain amount of unreliability in the training data as long as it is randomly distributed, any patterns in the unreliable annotations may be learned and reproduced by the algorithm alongside the intended characteristics of the target annotation (Reidsma & Carletta, 2008). In addition, Loukina et al. (2020) demonstrate that unreliable grading has a large effect on evaluation. The more unreliable the grading of the test data, the more distorted and possibly misleading the evaluation results will be. This directly affects the comparability of systems across different data sets that differ in the reliability of their grade annotations.

This leads us to believe that grader agreement and its variance across corpora is a relevant topic of consideration and informative for any researcher working with short answer grading corpora. Arguably, the significant variation of characteristics between the available corpora affects the ease with which two independent annotators agree on the same grade for a given answer. The challenge here is to draw insights on this effect from a range of corpora of different sizes and a range of properties nested in a hierarchical structure, since the relevant properties vary not only on the corpus level (like target language or native/non-native speakers) but also on the question level (some corpora mix content and assessment questions and most vary question difficulty).

In this article, we propose a principled, quantitative approach that addresses these challenges with the use of *generalized mixed effects models* to compare and contrast the properties of the most common Short Answer Grading corpora (Gelman & Hill, 2006). They are a common tool in psycholinguistics and other areas, but are rarely used in (applied) NLP. Mixed effects models enable us to analyze a target variable in terms of a hierarchically structured combination of predictors, which is crucial for the SAG data. They offer a principled way to identify the effect of these properties on the outcome and to assess whether these effects are statistically significant.

Most previous studies on grading consistency focus on the corpus level, typically comparing the overall consistency of judgments between different annotation settings by measuring inter-rater correlation (for interval and ratio scales) or inter-rater agreement (for nominal scales). For inter-rater agreement, a standard choice is the kappa (κ) family of agreement measures (Artstein & Poesio, 2008; Reidsma & Carletta, 2008) which correct the observed agreement by an estimate of agreement expected by chance. In contrast, our study considers the agreement between two annotators' grades *at the level of individual answers*. This permits us to carry out a fine-grained analysis of which properties at the corpus level *and* at the question and answer levels, that systematically influence grader agreement within and across different short answer corpora.

More specifically, we ask about the influence of *answer length*, *question difficulty level* and *answer correctness* (the ground truth assessment whether the answer is in fact correct or not) on grader agreement. Answer length may influence ease of grading since longer answers take longer to read and possibly open up more margin for interpretation. Question difficulty is included because intuitively, a question

testing recall of facts is easier to grade consistently than one that asks the student to include a justification or make an argument. Previous work implies that this is in fact also the case for machine learning systems which need less information to grade easier questions than harder questions (Padó, 2017). Finally, answer correctness is included because the grading task is likely asymmetric—we hypothesize that it is easier to identify correct answers consistently since comparison against a reference answer or a positive statement in a scoring rubric is probably simpler than assessing ‘creative’ answers.

Our analysis yields a better understanding of short answer corpora. We find that some simple intuitions hold true (e.g., that longer answers are harder to grade), while others do not (e.g., that sets of similar answers are easier to grade). In addition, we obtain a nuanced picture of consistency across SAG corpora: some properties behave consistently, while others show idiosyncratic behavior. Similarly, some corpora are in line with the general tendencies, while others form outliers, often due to unbalanced distributions with regard to certain properties.

This paper is structured as follows: in Sect. 2 we discuss previous work on analyzing short answer grading corpora. Section 3 provides an overview of SAG corpora and their properties. Section 4 introduces mixed-effects logistic regression models and our experimental setup. Sections 5 and 6 describe our two experiments—one builds general models for corpus aggregates at the level of assessment tasks (language and content assessment), one relates the general models to models for individual corpora. Section 7 concludes with a discussion.

2 Background

The corpora our study is concerned with support the task of Short-Answer Grading (SAG). Similar to many NLP tasks, SAG was first considered in the 1960s (Page, 1966). The first corpora as a resource for machine learning approaches became available in 2011 (Meurers et al., 2011a, b; Mohler et al., 2011). The visibility of the field was raised through public challenges, notably the Automated Short Answer Prize¹ (ASAP) run by the Hewlett Foundation in 2012 and the SAG shared task at SemEval-2013 (Dzikovska et al., 2013). Both competitions introduced new standard data sets for comparable evaluations. New data sets are continuously being created, and expand the range of existing corpora. Padó and Kiefer (2015) assemble content-assessment questions for German to supplement the existing German language-assessment data; Horbach et al. (2018) even create a German version of part of the ASAP data set in order to vary a specific corpus property (namely, language) in a controlled way. The size, range and quality of the available SAG corpora is crucial for the development of the NLP models for SAG.

Work on SAG in the NLP context either takes the form of fully automated grade prediction or of support for manual graders to reduce human effort. The first strand, automated grade prediction, is usually framed as a classification task with the manually annotated grade levels as target classes. Models can be trained either to

¹ <http://www.kaggle.com/c/asap-sas>.

score unseen answers for questions with available training answers (“seen-question” setting). This approach yields high model quality, but restricts the practical application of the model to settings where a sufficient amount of sample answers exists before testing. This is usually not the case outside the standardized testing context. Alternatively, models can be trained to predict answer quality for completely novel questions (“unseen-question” setting). This generally involves comparing student answers to correct reference answers so that the models learn to recognize what types of deviation from the reference answer is acceptable (Burrows et al., 2015).

The second strand of SAG research targets support for human graders. The intention is to save human effort while preserving the opportunity for teachers to gain insights from student answers and to use human supervision in order to alleviate the known weaknesses of machine grading. One standard approach is to cluster student answers according to their similarity. Annotators then assign grades to clusters instead of individual answers, which saves grader effort while still giving annotators insight into typical student reasoning (Brooks et al., 2014; Horbach et al., 2014).

The machine learning approaches used to model SAG typically take answer text as input and score only its content correctness. Other qualities of the text, such as language correctness and complexity, argumentative and text organization are not considered, neither are covariates such as difficulty level, answer similarity, or assessment task (Burrows et al., 2015).

As the use of NLP for SAG matures and the number of corpora rises, the need to better understand the available data is becoming more pressing. Burrows et al. (2015) comprehensively list the data sets used by all the SAG approaches they survey. More recently, Horbach and Zesch (2019) present a targeted analysis of the effect of various data properties on the performance of an NLP SAG system. They investigate the influence of the properties on the amount of variance present in the data, where less variance means better system performance. They find effects of answer length and grading scale as well as large performance differences between individual answers, but no large influence of target language and students’ language proficiency. We aim to address similar issues, but at the level of annotation consistency of manual grades, which forms the basis for training and evaluation of machine learning systems.

3 Data

3.1 Corpora

Each of the short-answer corpora available for research was collected to serve a specific research interest, and as a result, there are substantial differences in corpus characteristics. For example, SAG corpora may contain a single set of grade annotations or multiple grades by different annotators. The goal of our study is to consider as many major SAG corpora as possible; however, given the research question, we can only include corpora with multiple human grade annotations.

Notably, this means we exclude the popular SemEval-2013 data (Dzikovska et al., 2013), which was created for a shared task with the intention of bringing together research on SAG with research on textual entailment (Dagan et al., 2009), but which only provides a single grade annotation for each item.

Furthermore, we also disregard the pioneering data from Mohler et al. (2011). The double annotation available for this corpus has been found to be quite inconsistent (Mieskes & Padó, 2018). The reason for this is probably that no grading guidelines (in the form of rubrics or reference answers) were available (Mohler et al., 2011). This differs from the annotation process for the other SAG corpora, and is an extraneous reason for grader disagreement that we cannot hope to model.

This leaves us with the corpora shown in Table 3: The training portion of ASAP², a corpus of English-language school test items for science and English, and ASAP-de, a collection of German answers to translations of three of the ten questions in English ASAP (Horbach et al., 2018). Further, we use CREE (Meurers et al., 2011b) and CREG (Meurers et al., 2011a), an English-language and a German-language collection of language learner answers to reading comprehension questions. Since the standard CREG version (CREG-1032) by design only contains answers that were graded in agreement, we extracted additional answers to the questions documented and annotated in CREG-1032 from the larger CREG-23k corpus version (which also contains answers with disagreeing annotations).

CSSAG (Padó & Kiefer, 2015) is a collection of German answers to university-level test items for a programming class. Finally, we include the English-language Powergrading corpus, short PG, (Basu et al., 2013) containing answers to questions from the US immigration quiz collected through crowd sourcing.

3.2 Grader agreement

Our target variable is grader agreement. As discussed above, we want to capture the agreement between the graders down to the level of the individual answer. For use as the target variable, we need to model it in a way that is as comparable between corpora as possible.

One important source of incomparability between corpora is the *grading scheme* (the number of levels available to the annotators), which can vary by question. In our data, only CREE, CREG and PG use the same grading scheme for all answers (namely binary grades, i.e., correct/incorrect). The two ASAP variants and CSSAG vary the grading scheme by question. ASAP uses either a three-point or a four-point scale, while the majority of CSSAG is graded on a three- or five point scale. More grading levels potentially lead to more disagreement between the annotators, but when they choose adjacent categories, the disagreement is less serious than for a grading scale with fewer levels. Therefore using the absolute difference in assigned grades is misleading.

We attempt to standardize the grading scheme by investigating *binarized percentage of annotator agreement*. This relativizes the differences between grader decisions on fine-grained grading scales while preserving major disagreement and

² <https://www.kaggle.com/c/asap-sas>.

Table 2 Class distributions for inter-grader agreement for various corpora

Corpus	ASAP	ASAP-DE	CREE	CREG	CSSAG	PG
Agree	15,460	752	487	3784	1365	6693
Disagree	1747	151	79	600	403	286
Log odds	2.28	1.78	1.96	1.99	1.47	3.15

decisions on coarser scales. In all corpora, we defined agreement to be present if the annotators assigned scores that differed by less than one half of the available points. For a four-point scale, this allows graders to differ in one point, but a difference in two points counts as disagreement. On a two-point scale, the grades have to be the same to count as agreement.

The choice of one half of the available points is, in principle, a parameter of our study. However, with a more lenient definition, hardly any cases of disagreement remain, while a stricter definition throws the differences in grading granularity among corpora into sharper relief.

As with binning procedures generally, this process loses some information. However, a binary measure of agreement represents the least common denominator of the corpora we consider, since three of the corpora in our study (CREE, CREG and PG) only use a two-point grading scale. The other corpora use between three and seven points, depending on the individual question. Thus, more fine-grained measures of agreement are only applicable to substantially smaller subsets of the our data set. A second, more theoretical motivation is that our choice of binarized agreement is equivalent to Fleiss' κ when κ is computed at the level of individual data points for two annotators (viz., 1 for agreement and 0 for disagreement, Mieskes, 2009).

Table 2 lists the distribution of binarized agreements and disagreements in the different corpora. We also show the log odds between classes to quantify the prevalence of disagreement. High log odds mean that one class is strongly predominant, log odds near 1 show an equal distribution of classes. The log odds vary between 3.15 (PG, least amount of disagreement) and 1.47 (CSSAG, highest amount of disagreement), with most values around 2. The table shows no effect of the size of the grading scale on binarized consistency—the corpora that use multiple-point scales and might show inflated consistency due to the binarization are in second, fifth and sixth place ranked by consistency, while a corpus with a binary grading scale is most consistent.

3.3 Properties

This section discusses the properties of the selected SAG data collections that are presumably relevant with regard to grader agreement. We first present properties on the corpus level, then properties that vary with each question and finally properties that vary with each answer.

3.3.1 Corpus level

Table 3 shows the corpora in our study and some of their properties. On the corpus level, one obvious difference between the data sets is the *assessment task*. This describes the underlying goal of the assessment questions answered by the students. Assessment is either focused on content or on (second) language skills. With the exception of ASAP, assessment task is located on the corpus level. Content assessment (CA) corpora are ASAP-de, CSSAG and PG. CREE and CREG are pure language assessment (LA) corpora. For ASAP, half the questions are content assessment (CA) questions, and half the questions are language assessment (LA) questions (see also Sect. 3.3.2).

Assessment task co-varies with another important variable: With the exception of CREE and CREG, which are corpora of language learner answers only, all respondents are assumed to be native or near-native speakers of the target language. Therefore, while ASAP contains LA questions, these are more complex than the LA questions in CREE and CREG, since they are aimed at (near) native speakers of the target language instead of learners. Conversely, the assumption that CA respondents are native speakers does not always hold. In principle, language ability should not be a factor in SAG grading reliability, because language correctness is not taken into account during grade assignment: Mistakes by non-native respondents should not adversely influence their grades. However, Weiss et al. (2019) found a disproportionate impact of error rate (but not linguistic complexity) on final grades when investigating teacher grading of German written final exams. Therefore, language errors and lack or presence of linguistic complexity might conceivably have biased graders in our data sets, as well. In the absence of information on students' language ability, we do not investigate this further.

Another very visible difference is the *target language*. We have three German and three English corpora that are mostly unrelated; the pair ASAP and ASAP-de is currently the best attempt to control the factor of language in SAG corpora. We do not expect the language to influence grader consistency in a systematic fashion—there is no reason why speakers of English should consistently grade English answers more strictly or leniently than speakers of German grade German answers.

Table 3 Properties of short answer corpora (CA content assessment, LA language assessment, *en/de* English/German, *Stand.* Standardized testing, *Classr.* Classroom testing)

	ASAP	ASAP-de	CREE	CREG	CSSAG	PG
Assessment task	CA/LA	CA	LA	LA	CA	CA
Language	en	de	en	de	de	en
Collection	Stand.	Research	Classr.	Classr.	Classr.	Research
# questions	10	3	61	165	31	10
# answers (total)	17,207	903	566	4384	1926	6980

Another variable of potential influence is the *collection context* of the corpus. The ASAP data was collected from standardized testing in US schools and is accordingly graded by the test provider's trained annotators according to standardized rubrics. CREE, CREG and CSSAG assemble answers from small-scale ad-hoc testing, while data for ASAP-de and PG were collected by crowdsourcing. For all of these corpora, grading was done by the teachers, the researchers or research assistants.

Corpus sizes vary with collection context. In standardized testing, large amounts of data are available, and ASAP is the largest corpus in our data set. Ad-hoc testing yields the smallest corpora. We do not model corpus size as a predictor variable but account for it in the estimation of our model (cf. Sect. 3).

Finally, pseudonymized *student identities* sometimes appear as a corpus-level feature, when each student is given an individual ID code so their performance can be tracked across different questions in the corpus (this is the case in CSSAG, for example). For most corpora, however, student ID numbers appear to be re-initialized for each question, placing the property on the question level, which we discuss next.

3.3.2 Question level

As discussed directly above, *assessment task* is strictly speaking a question-level variable in our study, since the ASAP corpus contains both content assessment and language assessment questions. All other corpora contain data for a single task.

Another question-level property that we consider is *question difficulty level*. We assume that harder questions have more complex answers that are in turn harder to grade consistently. For content assessment, difficulty can be described with the Cognitive Process dimensions of Bloom's Taxonomy (Anderson & Krathwohl, 2014). There are six levels describing different cognitive processes necessary to answering questions. The first three levels are relevant here (none of the more advanced levels are present in the corpus data). They are *remember* (pure factual retrieval), *understand* (demonstrating comprehension of concepts, e.g. by explaining, comparing or classifying) and *apply* (using knowledge to solve a new problem). For language assessment, the taxonomy by Day and Park (2005) is more appropriate, since it is targeted directly at specifying different tasks relevant for reading comprehension questions in language teaching. The taxonomy also comprises six levels, of which only the three most basic levels are relevant. They are *literal comprehension* (repeating information from the text), *reorganization* (combining several explicit statements from the text), and *inference* (reasoning about information from the text). In case several tasks are required to answer a question, the category *several* can be assigned. CREE already contains Day & Park difficulty level annotation.

The *average answer length* per question differs widely between questions. We investigate this property to consider length effects on grading.

Finally, we look at *answer set homogeneity*, that is, the average similarity among answers for one question. Previous research has shown that this property is

correlated with question difficulty (Padó, 2017) and varies between CA and LA corpora (Padó, 2016). We expect that it should be easier to grade consistently if all the answers to a question are similar and the same grading decision can be re-applied multiple times. Annotator agreement will thus be improved if both annotators consistently apply the same grade to the homogeneous answer set. However, note that each of two annotators might also consistently apply a different grade. The result would then be maximal inconsistency between the annotators on the question level.

3.3.3 Answer level

On the answer level, we consider the *answer similarity* to the other answers for the same question. We include this property to complement the question-level factor of answer homogeneity. On the question level, answer homogeneity tells us how similar all answers to one question are on average. On the answer level, answer similarity indicates whether the answer question is very similar to other answers to the same question, or whether it is an outlier. Outliers might be easy to grade if they are clearly wrong or just a statement like “I don’t know”, or they might be hard to grade if they misinterpret the question or are highly original. Conversely, typical answers might be easier to grade because the same grading decision made elsewhere can just be re-applied.

A related concept, question–answer similarity, might also seem promising as a property at first glance, assuming that answers containing words that occur in the question address the correct topic. However, it is well known that students use repetition of question words as a filler strategy, so question–answer similarity is not informative. Therefore, we do not consider this property (cf. Mohler et al., 2011).

The final relevant property is the *correctness* of an answer. Correct answers are arguably easier to grade consistently by human graders since they can, for many difficulty levels, be compared against a reference answer or a positive statement in a scoring rubric. However, such matching strategies are not equally effective for all difficulty levels—the more creative the student has to be, e.g. in *reorganization*, the smaller the benefit for correct answers. We therefore expect an interaction between difficulty level and correctness (see Sect. 4.4).

4 Analysis framework

4.1 Multi-level structure

A particular challenge of the corpora described in the previous section is their multi-level structure, as shown in Fig. 1. We are interested in two fundamental types of corpora (language and content assessment corpora), with several individual corpora of either type. Within each corpus, we have a number of distinct questions, each of which comes with a number of answers given by various students. Since students typically provide answers to more than one question, the levels are not even

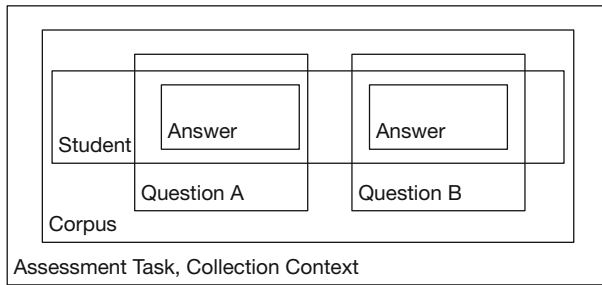


Fig. 1 Nested structure of data set

properly embedded in one another (note that student identification across questions is only possible in some corpora).

Recall that the focus of our interest is on the effects of various properties on grader agreement. These properties are located at various levels. Applying models of data analysis which treat all variables as if they were on the same level in such a situation typically leads to misleading outcomes: due to the collinearity between properties that describe the same phenomenon at different levels of the hierarchy, such analysis models will tend to assign all causal effects to the lower levels, since these enable the model to make finer-grained distinctions.

To give a concrete example, assume that language assessment questions are systematically more difficult to grade than content assessment questions (a corpus type-level difference). A single-level model can just as well capture the difference at the level of individual questions by simply assigning each language assessment question a higher difficulty. This is undesirable, since the model misses the higher-level generalization. Therefore, equipping the model with information about the level of each input variable allows us to draw more meaningful inferences.

4.2 Mixed-effects logistic regression models

To address these challenges, we model the short answer corpora with a family of multi-level models called mixed-effects logistic regression models (see e.g. Gelman & Hill, 2006; Snijders & Bosker, 2012). These models are generalizations of the single-level logistic regression models which are well established in linguistics (Bresnan et al., 2007; Baayen et al., 2008); see Baayen (2008) for an introduction. Single-level logistic regression models assume a set of observations x , a set of feature functions $f_1 \dots f_n$ that compute the value of the i -th predictor for each observation, and a binary response variable y . They have the form

$$p(y = 1) = \frac{1}{1 + e^{-z(x)}} \text{ with } z(x) = \sum_i \beta_i f_i(x) \quad (1)$$

where β_i is the coefficient assigned to the predictor x_i . Model estimation sets the parameters β so that the likelihood of the observed data is maximized.

From a linguistic point of view, a very interesting property of these models is that they enable us to assess the importance of the different predictors for a given data set: For each predictor $f_i(x)$, the estimated value of its coefficient β_i can be compared to its standard error, which makes it possible to test the null hypothesis that $\beta_i = 0$, i.e., the predictor does not significantly contribute to the model. Furthermore, the absolute value of β_i can be interpreted as the *log odds*—that is, as the change in the probability of the response variable correlated with a change in x_i :

$$e^{\beta_i} = \frac{P(y = 1|f_i(x) = 1, \dots)/P(y = 0|f_i(x) = 1, \dots)}{P(y = 1|f_i(x) = 0, \dots)/P(y = 0|f_i(x) = 0, \dots)} \quad (2)$$

The fact that z is just a linear combination of predictor weights encodes the assumption that the log odds combine linearly among factors—this is, that factors are independent of one another. This means that collinearity among predictors (which violates the independence assumption) may lead to misleading estimations of log odds. Testing models for the presence of collinearity is therefore crucial.

The mixed-effects generalization of logistic regression models complements the predictors of single-level models, which are called *fixed effects* x , with a new set of predictors, called *random effects* x' . In this article, we restrict ourselves to so-called *random intercepts* models. These models assign each observation to a specific group (in our case, which student, which corpus, which question) and permit the model learn a coefficient *for each group* that linearly affects the prediction. Formally, assume a set of functions g_i which assign each observation to specific group g . Then the random intercept models redefines z from Eq. (1) as follows:

$$z(x) = \sum_i \beta_i f_i(x) + \sum_j \gamma_j g_j(x) \quad (3)$$

where the γ_j are the coefficients of the random effects.

In contrast to the coefficients of the fixed effects, β_i , the coefficients for each group g_j of the same type (e.g., all corpora, all students, etc.) are supposed to have a zero mean. That is, the random intercepts can account for the differences (i.e., otherwise unexplained variance) among the different groups at this level. This is particularly appropriate if we can only observe random samples for these levels, since it maximizes the generalizability of the model to unseen levels of the predictor (Clark, 1973). This is a typical situation in natural language processing in general, where we train a model on one existing data set and make predictions on another—in the educational case, for novel answers, questions, and students.

Figure 2 shows a simple example where data points from three groups are being fitted by a regression model. The left-hand side shows a single-level model that does not know about group membership. In this model, it seems that the predictor is positively correlated with the response variable. With knowledge of the groups, however, it becomes evident that the three largest values of the predictor all stem from data points from one group. If this pattern is accounted for by introducing a group-level intercept γ_{g_3} (a constant offset on all data points of the group), the resulting regression line shows a better fit with the data and is arguably less biased by the behavior of individual groups.

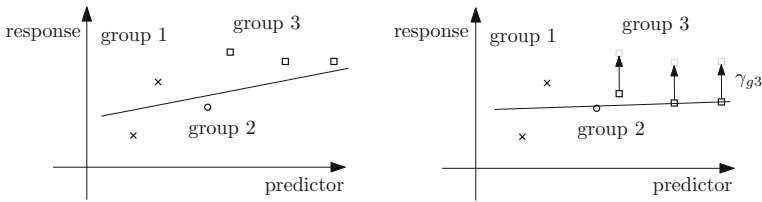


Fig. 2 Single-level regression (left-hand side) vs. random intercepts model (right-hand side): Modeling group membership (squares: group 3) as a constant offset γ_{g3} improves model quality

In our study, the methodological goal is a bit more specific: we use random effects to model the nested structure of predictors outlined above. Revisiting the example of corpus type vs. question from above, we can treat the more specific predictor (the identity of the question) as a random effect, and the less specific predictor (the corpus type) as a fixed effect—thus giving the model an incentive to preferably explain the response variable in terms of the corpus type, and only explain the remaining variance in terms of the specific questions.

4.3 Model evaluation and validation

Regression models are evaluated by measuring their ability to explain the data under consideration, often also taking into account model complexity. The most widely used measure, namely correlation between actual and predicted values of the response variable, is not applicable to our setup, which uses logistic regression to make categorical predictions. Instead, we report the Akaike Information Criterion (AIC):

$$\text{AIC} = 2k - 2 \ln(\hat{L}) \quad (4)$$

where k is the number of parameters in the model and \hat{L} the likelihood of the data given the model. AIC scores reflect that every extension of a model with new predictors increases the likelihood of the data. They are also widely used for model selection; the difference in AIC scores can be interpreted in analogy to a likelihood ratio (Burnham & Anderson, 2002).

While regression models technically support arbitrary sets of predictors, strong correlations among predictors, so-called multicollinearity, can distort the estimation of coefficients. Therefore, once the final model has been selected, it should be checked for the presence of multicollinearity. There is a wide range of tests available, see Ullah et al. (2019) for a recent overview. We use the so-called variance inflation factor (VIF). VIF measures how much the variance of a predictor's coefficient is inflated due to correlations with other predictors. Formally, the variance inflation factor for the predictor i , VIF_i , is defined as

$$\text{VIF}_i = 1 / (1 - R_i^2) \quad (5)$$

where R_i is the Pearson correlation coefficient obtained when predicting α_i from all

other predictors. Thus, the more collinearity is present, the higher VIF_i . VIF values of 4 or greater are generally interpreted as severe multicollinearity (Salmerón et al., 2018) and values greater than 2.5 warrant further investigation. A number of strategies are available to reduce collinearity, including dropping covariates, dimensionality reduction, and regularization; see Dormann et al. (2013) for details.

4.4 Experimental setup

Experimental rationale Given the small size of individual SAG corpora, our *Experiment 1* builds models for groups of corpora. We group the data by elicitation task: one model for language assessment (LA), and one for content assessment (CA). This grouping is natural, since the established classifications used for question difficulty levels are incomparable across these two tasks (cf. Sect. 3.3). This experiment tells us what properties these corpora have in common. The models on the level of corpus groups will identify important factors in predicting grader agreement separately for each elicitation task. Comparing and contrasting the patterns found for each elicitation task will yield insight about which factors are generally relevant in SAG and which are specific to the task. This will also shed light on how different the data collected in each of the tasks really is. Researchers so far have often used corpora from different elicitation tasks side by side, without strong evidence for their general comparability.

Subsequently, *Experiment 2* analyzes individual corpora and checks to what extent the shared patterns we found in the general models in Experiment 1 carry over to individual corpora. This comparison will show how these corpora differ in their basic characteristics. This will help researchers who are working with the individual corpora know what to expect from each corpus.

Response variable Our response variable is a binarized version of agreement between the human graders on the level of individual answers, as described in Sect. 3.2.

Fixed effects We use the following predictors as fixed effects, as described in Sect. 3.3 above.

- *Difficulty levels* Manual annotation of difficulty levels was only available for a subset of corpora. On the CA side, difficulty annotation only existed for CSSAG (Padó, 2017). For the CA part of ASAP and PG, two annotators independently annotated the difficulty levels according to the definition in Anderson and Krathwohl (2014), with the option of adjudication in case of non-agreement. The questions in ASAP(-de) and PG were not tied to a specific instructional context, so there is reason to assume that the caveat from Padó (2017) concerning the fact that explicit previous exposure to the specific subject matter of question changes the difficulty level of that question does not apply here. The PG and ASAP content assessment questions were annotated at 100% agreement, and the annotations were transferred to the ASAP-de questions (which correspond directly to the English ASAP questions). On the LA side, difficulty annotation already exists for CREG. Again, two annotators annotated CREE and the LA part of ASAP on the basis of Day and Park (2005) following a similar

procedure. The five ASAP LA questions were again easy to annotate, with 100% annotator agreement. For CREE, there was disagreement for 20% of questions (12 out of 61) before adjudication. The annotators agreed best on *literal*, the most frequent difficulty level, at 96% agreement, and least on *reorganization*³. The levels *literal* (LA) and *remember* (CA) are chosen as reference levels, since they are the lowest levels in their respective hierarchies.

- *Answer Length* We measure answer length by using z-scores of answer length in white space separated tokens, normalized for each corpus.
- *Answer Similarity* We compute answer similarity as the average similarity of an answer to all other answers for the same question. We measure similarity with Greedy String Tiling (GST), a non-symmetric string similarity algorithm (Wise, 1996) used for paraphrase detection. GST tests whether substrings of an answer are also present in other answers, but does not penalize substrings from the other answers missing in the current answer. Therefore, “high similarity” is often assigned to short answers that contain only words also found in other answers—but possibly leave out part of the intended solution. “Low similarity” answers contain words or phrases not often found in the other answers—alternative formulations, wrong concepts or erroneous spellings or usage. In order to make answer similarities comparable across questions and data sets, we discretize answer similarity into three bins (low, mid, high) by z-scoring the average similarities and using -0.5 as 0.5 and cutoff points.
- *Answer Set Homogeneity* For each question in a corpus, we compute answer set homogeneity as the average answer similarity.
- *Correctness* We use the ground truth grade provided in each data set. In order to correctly interpret the grade, we have to consider the relevant grading scale. Since our goal is a comparable, easily interpretable value, we consider an answer to be correct when the student is awarded 50% or more of the possible points, and incorrect otherwise.

We leave out question length: Due to the generally small number of questions in each data set (cf. Table 3), estimates often depend strongly on (properties of) single questions and therefore do not yield generalizable insight.

Random effects We adopt the general strategy proposed by Matuschek et al. (2017) positing only theoretically motivated random effects. We learn random intercepts for the following predictors, permitting the model to explain the variance introduced by them.

- Identity of corpus
- Identity of question
- Identity of student (where available)

We do not include collection context because of the sparsity of this effect: for some collection contexts, we just have one or two corpora, as a result of which the parameters are not uniquely identifiable.

³ The annotations are available; see the end of Sect. 4 for details.

Model shape Our final model look as follows:

We treat all fixed effects as main effects that are independent of one another. The one exception is an interaction between difficulty level and correctness, which permits the model to account for the arguably plausible hypothesis that the effect of correctness differs in strength for different types of questions.

Corresponding to the choice of a dichotomous response variable, we train a *logistic* mixed effects model, using a logistic link function. Importantly, we reweigh the contribution of each data point inversely by corpus size, so that each corpus has the same total influence on the model parameters. Our rationale is that we are mainly interested in observations that hold across corpora irrespective of corpus size.

Implementation We implemented all models in R, using the package *blme* (Chung et al., 2013) which implements Bayesian estimation of mixed effects models to improve convergence, using the default priors throughout. The complete analysis is available, together with the data files and created annotations, as a Jupyter/R notebook that can be downloaded from <https://www.ims.uni-stuttgart.de/data/sag-grader-agreement>.

5 Experiment 1: Modeling language and content assessment corpora in aggregate

We fit two logistic mixed effects models to the aggregates of language and content assessment corpora, respectively. We call these the “Overall” models. The Overall language assessment model has an AIC of 17,319 (log likelihood -8647) and has variance inflation factors of < 2 for all fixed effects. The Overall content assessment model has an AIC of 10,033 (log likelihood -5006) and all variance inflation factors < 1.5 . The fixed effects of the two models are shown in Tables 4 and 5, respectively.

5.1 Analysis of fixed effects

The first part of our analysis compares the estimates for the fixed effects obtained by the Overall LA and CA models. The tables give the model estimates for each predictor (or each level for categorical predictors) as well as standard errors, z scores, and bounds for the p values. We organize the discussion of the results along six main observations.

Observation 1: Longer answers are significantly harder to grade for both LA and CA. As expected, answer length is a major factor in predicting grader agreement. Presumably, longer answers are often more complex and require detailed assessment – similar to how, e.g., longer sentences are harder to rank in the manual assessment of machine translation output (Bojar et al., 2011). Longer answers can also result from more complex question types, e.g. questions that

Table 4 Fixed effects in language assessment corpora

	Predictor	Level	Estim.	SE	z	p
Main effects	Answer length		-0.38	0.04	-8.62	< 0.001
	Answer set homog.		1.40	1.00	1.40	> 0.1
	Correctness		3.14	0.09	34.72	< 0.001
	Answer similarity	Mid	-0.19	0.05	-3.65	< 0.001
		High	-0.22	0.06	-3.50	< 0.001
	Difficulty level	Reorganization	1.57	0.50	3.16	< 0.001
Inference		2.24	0.54	4.15	< 0.001	
Interactions	Correctness × diff.	Corr. – reorg.	-2.49	0.12	-20.64	< 0.001
		Corr. – inf.	-3.25	0.13	-25.66	< 0.001

Table 5 Fixed effects in content assessment corpora

	Predictor	Value	Estim.	SE	z	p
Main effects	Answer length		-0.44	0.08	-5.56	< 0.001
	Answer set homog.		2.07	3.33	0.62	> 0.1
	Correctness		1.13	0.11	10.52	< 0.001
	Answer similarity	Mid	0.29	0.09	3.26	< 0.001
		High	0.30	0.11	2.80	< 0.001
	Difficulty level	Understand	-1.17	0.72	-1.62	> 0.1
Interactions	Correctness × diff.	Corr. – Und.	-0.85	0.17	-4.94	< 0.001

require multi-part answers (“Name three reasons ...”). Since we do not have a predictor that characterizes this aspect of questions, answer length is likely to pick up such effects as a proxy.

Observation 2: Answer Set Homogeneity is not significant, for neither LA nor CA. We hypothesized that homogeneous answer sets could be easier to grade, because the opportunity to re-apply a grading decision frequently to a similar answer should lead to larger consistency within the grades of each annotator for the set of similar answers and therefore to convergent grades (provided that the annotators agree on the grade for the answer set). This is not the case: we see a positive coefficient for homogeneity in both Overall models, but it does not reach statistical significance. In fact, the standard errors are large, so answer set homogeneity manifests only as a very mild tendency. The extremely low multicollinearity scores we find for homogeneity ($VIF < 1.05$) make it unlikely that this null finding is an artifact of a strong correlation with other predictors (for example the difficulty level). Therefore, it seems that any effect of higher consistency within the grades of one annotator is not strong enough to raise annotation agreement between several annotators.

Observation 3: Correct answers are significantly easier to grade for both LA and CA. In both Overall models, this is the strongest main effect. Our explanation is

similar to Observation 1: Correct answers require less assessment by the grader. Note that this is not necessarily true: it would also be conceivable that there are clearly wrong answers that can be recognized easily and graded consistently. That the data are so emphatically biased in one direction is due, we believe, to the *asymmetry* of testing, and thus of testing materials: for each question, test developers typically have one correct answer in mind, and graders have access to a reference answer or positive rubrics, but not to a comparable set of wrong answers: there is one way to be right, but many ways to be wrong.

Observation 4: For LA, low-similarity answers are easiest to grade. For CA, in contrast, low-similarity answers are hardest. This observation calls for a more detailed discussion. Recall from Sect. 4 that low similarity generally means that answers are unusual (creative, or wrong), while high similarity means answers contain frequently used words, but are potentially incomplete. Despite the contradictory patterns for LA and CA, there is common ground: mid and high similarity answers cluster together, away from the low similarity answers. This implies that it is the difference between potentially incomplete answers and incorrect or unexpected answers that drives this pattern.

For LA, the data shows that high-similarity (incomplete) answers leave room for interpretation by the graders. This is compounded by the use of a two-point grading scale in CREE and CREG, two of the three LA corpora. For example, a question from CREG asks students to name three requirements listed in a German job ad. One high-similarity answer lists only two: “Selbständige Arbeitsweise gute Englischkenntnisse” (independent worker, good knowledge of English). One grader accepted the answer because the listed properties are correct, the other rejected it based on the requirement to name three properties. No partial credit was possible that the graders could have agreed on. In contrast, low-similarity answers are often clearly incorrect or just add phrases that other answers left out (answering “Watching television and sleep are two major activities.” when all other answers were simply “watching television and sleeping”, in CREE). In consequence, the majority of grader disagreement for LA corpora is among high-similarity questions.

For CA, all corpora (with the exception of PG) use grading scales that allow partial credit, so incomplete answers can be consistently graded as such—and the binarization method we use implies that assignment of partial credit by both graders are almost guaranteed to be interpreted as agreement. Thus, the potential incompleteness of high-similarity answers causes no increase in grader disagreement. Instead, we find that unusual, low-similarity answers give rise to much more need for interpretation, and lead to more grader disagreement, than for LA. For example, is it acceptable to answer the PG question “What is one reason the original colonists came to America?” with the word “gold”? It is definitely a non-canonical answer, but the set of reference answers includes “economic opportunity”, which could be interpreted to include the colonists’ intention to acquire gold in America.

Observation 5: For LA, higher difficulty levels are easier to grade. For CA, there is an opposite non-significant trend. The CA results match our intuition that harder questions require more complex, longer answers that are in consequence harder to grade consistently.

Table 6 Correctness \times difference level interactions in language assessment corpora

	Literal	Reorganization	Inference
Incorrect	0.00	1.57	2.24
Correct	3.14	2.22	2.13

Table 7 Correctness \times difference level interactions in content assessment corpora

	Remember	Understand
Incorrect	0.00	-1.17
Correct	1.13	-0.89

When we investigate the inverse pattern in the LA corpora, we find that the difficulty levels are distributed almost in complementary fashion between the corpora: 83% of the answers in CREE and 81% of the answers in CREG are at the *literal* difficulty level. In contrast, ASAP contains no *literal* answers at all, but covers the *reorganization* and *inference* levels that the other corpora populate relatively sparsely. Given this correlation of difficulty levels with corpus identity, the data do not allow us to determine whether higher-level answers really are easier to grade for LA corpora, or whether this observation results from systematically more consistent grading for ASAP compared to CREE and CREG. Recall from Table 2 that ASAP is overall one of the most consistent corpora, more so than CREE and CREG.

Note that not all difficulty levels are present in every corpus for the CA data, either, but CSSAG and ASAP both cover the two difficulty levels we consider, so we do not have a quasi-complementary distribution as for LA.

Going forward, the results regarding difficulty level and its interaction with correctness are hard to interpret directly from individual estimates. To simplify the interpretation, Tables 6 (for language assessment) and 7 (for content assessment) show the contribution to rater reliability made by every combination of the difficulty level and correctness level. Here, our main observation is:

Observation 6: For both LA and CA, the effect of correctness is significantly less positive for higher difficulty levels. For LA, the largest effect of correctness—i.e., the mean difference in agreement between correct and incorrect answers for a set difference level, which manifests here as the difference between the corresponding model estimates—occurs for the *literal* level, and for CA, for the *remember* level. Moving to the highest difficulty level, it decreases from 3.14 to -0.11 (LA) and from 1.13 to 0.28 (CA). This indicates that as questions become harder and answers require more argumentation on the part of the student and interpretation on the part of the grader, it is no longer easier to identify correct answers consistently, as there is no simple check against an answer key available anymore. For CA, the picture

Table 8 Standard deviation accounted for by random effects

	LA	CA
Student ID	0.74	3.57
Question ID	2.20	1.96
Corpus ID	0.56	2.24

Table 9 Random effect coefficients for LA and CA corpora

	LA	CA
ASAP (LA)	-0.26	ASAP (CA) 0.23
CREE	0.20	ASAP-de -1.16
CREG	-0.49	CSSAG -3.87
		PG -3.36

matches our O3 (correct answers are easier to grade consistently) and the trend in O5 (harder questions are harder to grade).

For LA, the observed pattern is again due to the complementary distribution of difficulty levels between corpora: CREE and CREG show a positive effect of correctness (covering the *literal* level), while we find no effect of correctness for ASAP, which provides the majority of data for the *reorganization* and *inference* levels (see also our discussion of corpus-specific models in Sect. 6). Again, as for O5, we cannot distinguish whether the effect is properly absent or whether its absence results from the imbalance of the corpora regarding difficulty levels.

5.2 Analysis of random effects

We now proceed to analyze the random effects in our models. We again contrast results for language and content assessment. Table 8 shows the standard deviation accounted for by the random effects in the two corpora (that is, the standard deviation of the grader agreement across levels of the random effect on the log-odds scale). Our observations are:

The property 'student ID' is the strongest random effect in the CA but not in LA. The reason is that in CA, we have multiple answers for 832/8400 students (about 10%), while this is only the case for 395/9400 students in LA (about 4%). Thus, this property has a lower explanatory potential in LA; this is however an artifact of corpus collection rather than an actual difference.

The property 'question ID' is the strongest random effect in LA and still accounts for a fair amount of variation in CA. In general, it is not surprising to find that the question level accounts for variation in the data. On the question level, grader agreement depends, among others, on the existence and quality of grading rubrics or guidelines, on the clarity of the question and the consistency with which the students interpret it and, for reading comprehension questions, on the difficulty of the reading passage. These effects have been accounted for nowhere else in the model and

therefore are being modelled here. We have 229 questions in LA, but only 44 in CA, which leads to the larger explanatory power for LA.

The property 'corpus ID' explains a substantial amount of variance for CA, but little for LA. To explain this observation, we list the coefficients for the individual corpora in Table 9. These coefficients describe the difference between the observed grader agreement for the individual corpora and the systematically expected grader agreement based on the Overall LA and GA models. Most of the coefficients are negative: the Overall models expect corpora to be somewhat more consistent than they actually are.

The Overall LA model appears to make good predictions for agreement in the individual corpora, which leads to a small effect of corpus ID. Indeed, an inspection of AIC scores shows that the effect of corpus ID does not contribute significantly to the explanation of grader agreement for LA, while all other random effects do, for both LA and CA (see Jupyter notebook for details). In contrast, the coefficients are large for CA corpora. Two corpora show particularly large negative coefficients, namely CSSAG and PG. We have reason to believe that these two corpora are not good representatives of the global CA data set. CSSAG has the lowest agreement of all corpora (cf. Table 2), pointing towards disagreements among graders that go beyond the corpus properties we model. PG, while showing very good agreement, is unusual in that it consists of questions on a single difficulty level.

5.3 Discussion

We have presented a set of observations about two overall models that aggregate multiple SAG corpora by elicitation task, since difficulty levels are incomparable across Language Assessment and Content Assessment. This enables us to assess influences on grading agreement separately for each elicitation task as well as probing the consistency of SAG corpora across elicitation tasks. Overall, we found a relatively high level of consistency: We found a negative effect of answer length on grader agreement for both LA and CA models (Observation 1), but an absence of effect for answer homogeneity (Observation 2). Our strongest effect is that correct answers are graded more consistently (Observation 3). However, this effect diminishes for harder questions (Observation 6).

The main difference between LA and CA corpora concerns the role of answer similarity, that is the quantification of how similar an individual answer is to all other answers to the same question. For LA, answers that are very dissimilar to all other answers are easiest to grade, for CA, there is a small inverse trend (Observation 4). There is a striking difference in LA and CA with regard to question difficulty level: While for CA, the higher levels are harder to grade consistently, for LA corpora, the higher difficulty levels are, somewhat counter-intuitively, easier to grade (Observation 5).

The outcome regarding random effects is mostly expected: both student and question identity account for substantial amounts of variance for both LA and CA. The main difference lies with the outcome of the random effect of corpus, which indicates that the LA corpora are overall more similar to one another, while the CA aggregate contains two 'outliers', CSSAG and PG.

Table 10 Comparison of assessment task-level and corpus-level models

	O1	O2	O3	O4	O5	O6
Overall LA	–	0	+	–	+	–
ASAP (LA part)	–	0	0	0	0	–
CREE	0	0	+	0	0	0
CREG	–	0	+	–	0	–
Overall CA	–	0	+	+	0	–
ASAP (CA part)	0	NA	+	0	–	NA
ASAP-de	–	–	+	0	0	–
CSSAG	–	0	0	0	0	0
PG	0	0	+	+	NA	NA

+ significant positive effect on rater agreement, – significant negative effect, 0 no effect, NA predictor does not exist, Above: LA, below: CA

6 Experiment 2: Individual corpora

We now test to what extent our observations also hold for the individual corpora. Are the observations of the Overall models supported by all (or most) corpora individually, or do they just describe an artificial aggregate that behaves differently from any individual corpus?

For this analysis, we estimate models for individual corpora (for ASAP, we estimate two separate models for the language and content assessment portions, respectively). We start with the same model structure used for the Overall LA and CA models in Experiment 1 (of course, leaving out the random effect of corpus). For corpora with only one difficulty level, we also omit the interaction term. We test each individual model for multicollinearity and find VIF scores < 1.6 throughout, indicating no problems with collinearity (see Jupyter Notebook for details). The only exception is the CA part of ASAP, where difficulty level and answer homogeneity are highly collinear. For this corpus, we remove both answer homogeneity and the interaction.

Table 10 shows the presence of observations O1 to O6 from the Overall models in the individual corpora. Overall model patterns are shown for reference. + denotes a significant positive effect, – denotes a significant negative effect, and 0 denotes the absence of a significant effect. NA indicates that the observations cannot be replicated by a model, due to the absence of a relevant predictor or the relevant level(s), as described above.

The first observation is that there are many zeros in the matrix, indicating that the effects we found in the Overall models are not present in each individual corpus. This is to be expected, given the smaller size of the individual corpora. Conversely, we rarely see the case that null effects in the Overall models (O2, O5 for CA) are significant in individual corpora. This is reassuring, since a large numbers of such cases would indicate that individual corpora differ strongly from the Overall models.

For the language assessment corpora (upper half of the table), we find generally good agreement with the Overall models, in line with the analysis of the corpus random effect in Sect. 5.2 which indicated consistency between the LA corpora. The LA part of ASAP is fairly similar to the Overall model, with matches for three

observations, and null effects for three others. CREE is more dissimilar to the Overall model: the only significant effect we find is Correctness (O3). This is presumably the case because CREE is the smallest corpus with only 566 answers (cf. Table 3). In contrast, CREG is again very similar to the Overall model. The only major difference is that CREG does not significantly show O5 (higher difficulty levels easier to grade), although there is a non-significant tendency in this direction.

For the content assessment corpora (lower half of the table), we see a larger amount of variance, again in line with the random effect of corpus discussed in Sect. 5.2. The CA part of ASAP is relatively dissimilar to the Overall model, with one match, three mismatches, and two NA observations. The clearest difference concerns O5 (higher difficulty levels are easier to grade) which is a significant only for ASAP (cf. the discussion in Sect. 5). Next, ASAP-de is very similar to the Overall model, with four matches and two mismatches. The major divergence is O2 (answer homogeneity): ASAP-de is the only corpus where homogeneity shows a significant negative effect on grader agreement. Next, CSSAG is again fairly similar, with three matches and three mismatches. Answer length is the only significant effect. We believe that this could be the case because CSSAG uses the most detailed (question-specific) grading scheme among all corpora (cf. Table 3) and therefore loses most information in the binary treatment of correctness that we assume. Finally, PG is very similar to the Overall model, with four matches out of four (O5 and O6 are not observable since PG covers only one difficulty level). The main difference is that the answer length is not significant. This is presumably the case because PG has by far the shortest answers of all corpora (median: 18 tokens, quartiles: 11, 31). Compare this to, e.g., the median of 128 tokens in CSSAG (quartiles: 74, 197). This indicates that answers vary much less in PG than in other corpora with regard to their level of detail.

A different perspective on Table 10 is provided by an analysis by observations: how many corpora conform to the Overall pattern in each column? We find that the observations fall into two broad groups in this respect: O1 through O3 show good support, with more than half of the corpora supporting the Overall observations (O1: 4/7, O2: 5/6, O3: 5/7), while support for O4 and O5 is much more sparse (O4: 2/7, O5: 2/6). O6 is a borderline case, in particular since we miss observations for two corpora (3/5).

The two main groups correlate with the type of the predictors: O1–O3 concern binary and continuous predictors, for which only a single parameter is learned, while O4–O6 concern three-level categorical predictors. These involve multiple parameters, which are learned independently by the models, and which therefore run a larger risk of showing different patterns across corpora, in particular when learned from smaller amounts of data. These technical considerations coincide with our findings in Experiment 1, where our qualitative analysis led us to mistrust O4–O6 on the basis of their connection to other factors, namely the grading scale in the case of O4 and corpus imbalance for O5 and O6. In sum, we believe that O1–O3, describe patterns that are generally valid for SAG corpora (except where obviated by specific properties of individual corpora, cf. the case of PG discussed above), while we cannot be sure that this is the case for O4–O6.

7 Discussion and conclusion

Overall findings The starting point for our study was the fragmentation of the short answer grading (SAG) data landscape. The availability of a large number of relatively small corpora that differ along multiple dimensions represents a major obstacle for research since generalizations beyond individual corpora are difficult. We have analyzed the corpora from the point of view of their grader agreement, a crucial prerequisite for reliable analysis and modeling. We have modeled various properties that we expect to influence grader agreement. Since these properties are hierarchically structured on the answer, question, and corpus levels, we use a logistic mixed effects model which is well suited to teasing apart the contributions of properties at different levels.

The outcome of our experiments is a nuanced view of general and specific properties of SAG data. We identify a number of effects that hold consistently across corpora—a common denominator is that consistent grading is easy if the task is cognitively simple (short answers, matching against reference answers), while there is, naturally, also some room for corpus specific behavior. The different SAG corpora are overall surprisingly consistent, despite their visible differences in elicitation task or grading scheme: The Overall models for language assessment and content assessment correspond well to one another, and to individual corpora. That being said, content assessment corpora tend to be somewhat more dissimilar to one another than language assessment corpora—which is to be expected, given their stronger dependence on subject specific expectations.

An important contributor to the success of this study was careful normalization of the predictors in the model—for example, answer lengths were represented as z-scores; similarly, answer similarity was z-scored and binned. This corresponds to the assumption that it is the relative magnitude of these predictors within each corpus, rather than their absolute magnitude, that matters for difficulty. The results validate this assumption: the effects generalize well not only between language and content assessment corpora, but also between English and German corpora.

Lesson learned: correlation and interpretation The present study showcases the usefulness of mixed effects models for understanding and comparing the structures of a wide range of corpora. At the same time, it also clearly shows the limitations of this approach. The most important one is that regression models only lend themselves to detailed interpretation if the predictors are only weakly correlated (Snijders & Bosker, 2012). This limits the number of properties that can be included in an analysis. We consider a relatively rich set of properties, but have to exclude, for example, the collection context of the corpora—classroom vs. standardized testing vs. scientific study, cf. Sect. 3.3—because this property is too strongly correlated with corpus identity, with one or two instances for each collection context. Similarly, we exclude some properties on a case-by-case basis in Experiment 2 due to high multicollinearity (cf. Sect. 6).

A more subtle case of a similar kind is presented by the difficulty levels in the LA corpora which show an initially puzzling pattern (O5). This turns out to be due to a complementary distribution of levels across the corpora, essentially creating a

correlation between difficulty level and corpus. Consequently, the model is unable to disentangle the effect of corpus from difficulty level (cf. Sect. 5). Finally, we find that some properties 'pick up' effects, beyond their original operationalization, that are not represented as predictors in the model—for example, answer length also mirrors question complexity. Again, this can be seen as a case of correlation.

For all of these reasons, it is crucial that one must proceed cautiously in the interpretation of the results of correlation models, and qualitative validation should be a central component of such studies.

Lessons learned: future directions for SAG corpora For practical as well as technical reasons (see previous paragraph), our study mostly focused on properties that were available for all corpora under consideration; we only added manual annotation for difficulty levels. However, we have come across a number of properties that would be interesting to include in future similar studies, and that are therefore worthwhile targets for the extension of annotation in SAG corpora. Some of them concern richer characterization of the question level, like the complexity of the expected answer (for example in terms of the expected number of facts), or the relative difficulty of the relevant passage for reading comprehension questions. Others concern a better understanding of the students, such as their language ability according to a standardized scale or even just a pseudonymized identification of all answers from the same person. Finally, the whole area of grader training is currently a kind of black box: we typically do not know how graders are trained, what the concrete grading process looks like, and how good the annotation rubrics or guidelines are.

Another aspect of interest for future SAG corpora concerns the corpus imbalance with regard to certain properties that we already discussed above. While of course complete representativeness is impossible to achieve, future corpus collection efforts should strive to balance the important properties of the data as well as possible. For example, a new language assessment corpus with a more even distribution of answers over the three commonly observed difficulty levels would allow us to tease apart the corpus effects from the effects of difficulty level discussed in Sect. 5.

Lesson learned: the grading scale matters Since the goal of our study was to identify shared patterns of grader agreement across different SAG corpora, we explicitly attempted to abstract away from the individual grading scales through binarization. Overall, we were successful: We were able to show consistent effects across corpora and generate general insights about the available corpora.

However, different grading scales are not a problem that can simply be binarized away. In our study, we find an implicit effect of grading scale for O4 (Sect. 5). As explained above, binarization arguably brings out the influence of the grading scale on O4 rather clearly by making the high-similarity CA answers appear more consistently graded than they were, while disagreement on the binary scales of the LA corpora remained unaltered.

This observation underscores the general significance of the grading scale for grading consistency: A grading scale that is too coarse may foster inconsistencies in corner cases like incomplete answers. Conversely, a grading scale that is too fine will generate disagreement on the absolute grades, since these become harder to

distinguish the more fine-grained they are. Beyond the grading task itself, fine-grained grading scales also make the task of machine learning models harder (cf. the discussion in Horbach & Zesch, 2019).

Lesson learned: correctness remains crucial One of our most robust findings in the Overall models (also well borne out by the corpus-specific models) is that correct answers are easier to grade than incorrect ones. To our knowledge, this effect has not been generally acknowledged in the literature, but it is highly relevant both for future grading endeavors and for the interpretation and use of existing grade annotations.

We believe that the effect of correctness (diminishing for harder questions in the CA corpora) comes about because the grading rubrics and reference answers are formulated positively and do not give examples of negative answers. This makes it easier to identify correct answers by direct comparison to the guideline materials. Therefore, it would be worth investigating in future research on grading consistency whether agreement can be improved by explicitly instructing the graders how to deal with incorrect or partially correct questions.

For the existing data, the imbalance in grading consistency that we found between correct and incorrect answers has important repercussions for model induction and evaluation. While machine learning algorithms are generally able to cope with (some) noise in the labels, the noise is often assumed to be randomly distributed (Reidsma & Carletta, 2008). Patterns in labeling inconsistencies may therefore easily be mirrored in the machine grader models, and consequently we would have to expect lower model reliability for incorrect answers. This is clearly undesirable in the application of machine grading models in real-world assessment.

More recently, Loukina et al. (2020) have demonstrated that variations in grader agreement substantially influence the comparability of model performance across corpora and have proposed a new evaluation measure that factors out the effect of grader agreement. This is a much-needed development; however, the measure also assumes that grading error is randomly distributed, which we have found not to be the case for the corpora we have analyzed. The impact of this mismatch remains a topic of future work.

Outlook: Generalizability Our study analyzes a set of SAG corpora based on their hierarchically structured properties (such as answer length or similarity between student and reference answers). Therefore, our specific observations apply primarily to these corpora, although we believe that they are sufficiently robust to generalize to other corpora of this family.

In contrast, our method is not limited to this specific kind of corpus or specifically to grade annotations. Recall that we made a point of operationalizing annotator agreement in a way that abstracts away from the concrete grading scheme in each of the corpora. Also, mixed effects regression models can accommodate essentially arbitrary feature hierarchies modulo concerns about model selection (Barr et al., 2013; Matuschek et al., 2017). Therefore, our methodology could be directly applied to corpora annotated with ordinal ratings, also beyond the SAG domain, for example reviews (Gao et al., 2019). Corpora with nominal annotation require some extra work in the form of a user-defined mapping from label pairs to agreement scores. An example of such nominal labels is found for example in CREG, where in

addition to the binary correct/incorrect annotation, there are extra labels that group incorrect answers into classes such as 'missing concept', 'incorrect concept', etc. (Bailey & Meurers, 2008). The simplest mapping from pairs of these labels to agreement scores would be exact match, but if the labels can be reasonably clustered by similarity, the mapping can also reflect partial matches.

In sum, we believe that our methodology is sufficiently general to be applicable to many different data sets—at any rate, the importance of understanding the inherent properties of the data set and its annotations remain crucial for the success of the analysis.

Acknowledgements The authors thank two anonymous reviewers for their comments and criticisms.

Author Contributions UP: idea, data preprocessing, interpretation, writing. SP: statistical analysis, interpretation, writing

Funding Open Access funding enabled and organized by Projekt DEAL. This study was not carried out under external funding.

Data and code availability The datasets and Jupyter Notebook used in this study, together with the created annotations, are all publicly available. See Sect. 4 for details.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anderson, L. W., & Krathwohl, D. A. (Eds.). (2014). *A taxonomy for learning, teaching and assessing: A revision of Bloom's*. Pearson Education.
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Workshop of Building Educational Applications*, (pp. 107–115). Columbus, Ohio.
- Baldwin, D., Fowles, M., & Livingston, S., (Eds.) (2008). *Guidelines for constructed-response and other performance assessments*. Educational Testing Service.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- 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.

- Bojar, O., Ercegovčević, M., Popel, M., & Zaidan, O. (2011). A grain of salt for the WMT manual evaluation. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 1–11). Edinburgh, Scotland.
- Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Science.
- Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: using clusters to grade short answers at scale. In *Proceedings of the Conference on Learning at Scale* (pp. 89–98). Atlanta, GA.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25, 60–117.
- Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4), 685–709.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- Dagan, I., Bill, D., Bernardo, M., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering*, 15(4), 105.
- Day, R. R., & Park, J.-S. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1), 60–73.
- Dormann, C., Elith, J., Bacher, S., Buchmann, C., Carl, G., Gabriel Carré, T., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36, 27–46.
- Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H. T. (2013). SemEval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Conference on Semantic Evaluation* pp. 263–274. Atlanta, GA.
- Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., & Miyao, Y. (2019). Does my rebuttal matter? Insights from a major NLP conference. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1274–1290). Minneapolis, Minnesota.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research*. Cambridge University Press.
- Horbach, A., Palmer, A., & Wolska, M. (2014). Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 588–595). Reykjavík, Iceland.
- Horbach, A., Stenmanns, S., & Zesch, T. (2018). Cross-lingual content scoring. In *Proceedings of the Workshop of Building Educational Applications* (pp. 410–419). New Orleans, LA.
- Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, 4, 28.
- Loukina, A., Madnani, N., Cahill, A., Yao, L., Johnson, M. S., Riordan, B., & McCaffrey, D. F. (2020). Using PRMSE to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Workshop of Building Educational Applications* (pp. 18–29). Seattle, WA.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.
- Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011a). Evaluating answers to reading comprehension questions in context: results for German and the role of information structure. In *Proceedings of the TextInfer Workshop on Textual Entailment* (pp. 1–9). Edinburgh, Scotland, UK.
- Meurers, D., Ziai, R., Ott, N., & Bailey, S. (2011b). Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(4), 355–369.
- Mieskes, M. (2009). *Exploring methods for the automatic summarization of meetings*. PhD diss: Friedrich-Alexander-Universität Erlangen-Nürnberg, Technische Fakultät.
- Mieskes, M., & Padó, U. (2018). Work smart—reducing effort in short-answer grading. In *Proceedings of the Workshop on NLP for Computer-Aided Language Learning* (pp. 57–68). Stockholm, Sweden.

- Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 752–762). Portland, OR.
- Padó, U. (2016). Get semantic with me! The usefulness of different feature types for short-answer grading. In *Proceedings of the International Conference on Computational Linguistics* (pp. 2186–2195). Osaka, Japan.
- Padó, U. (2017). Question difficulty—how to estimate without norming, how to use for automated grading. In *Proceedings of the Workshop of Building Educational Applications* (pp. 1–10). Copenhagen, Denmark.
- Padó, U., & Kiefer, C. (2015). Short answer grading: when sorting helps and when it doesn't. In *Proceedings of the Workshop on NLP for Computer-Aided Language Learning* (pp. 42–50). Vilnius, Lithuania.
- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5), 238–243.
- Reidsma, D., & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34, 319–326.
- Salmerón, R., García, C. B., & Pérez, J. G. (2018). Variance inflation factor and condition number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12), 2365–2384.
- Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage Publishers.
- Ullah, I., Muhammad, M. A., Altaf, S., & Ahmed, M. (2019). Some new diagnostics of multicollinearity in linear regression model. *Sains Malaysiana*, 48(2), 2051–2060.
- Way, W. D., Vickers, D., & Nichols, P. (2008). Effects of different training and scoring approaches on human constructed response scoring. In *Proceedings of the Annual Meeting of the National Council for Measurement in Education*. New York City, NY.
- Weiss, Z., Riemenschneider, A., Schröter, P., & Meurers, D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the Workshop of Building Educational Applications* (pp. 30–45). Florence, Italy.
- Wise, M. J. (1996). YAP3: Improved detection of similarities in computer program and other texts. *Bulletin of the ACM Special Interest Group on Computer Science Education*, 28(1), 130–134.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.