CITIZEN SCIENCE:
THEORY AND PRACTICE

Context Matters: Accounting for Item Features in the Assessment of Citizen Scientists' Scientific Reasoning Skills

METHOD

TILL BRUCKERMANN
TANJA M. STRAKA
MILENA STILLFRIED
MORITZ KRELL
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Citizen science (CS) projects engage citizens for research purposes and promote individual learning outcomes such as scientific reasoning (SR) skills. SR refers to participants' skills to solve problems scientifically. However, the evaluation of CS projects' effects on learning outcomes has suffered from a lack of assessment instruments and resources. Assessments of SR have most often been validated in the context of formal education. They do not contextualize items to be authentic or to represent a wide variety of disciplines and contexts in CS research. Here, we describe the development of an assessment instrument that can be flexibly adapted to different CS research contexts. Furthermore, we show that this assessment instrument, the SR questionnaire, provides valid conclusions about participants' SR skills. We found that the deep-structure and surface features of the items in the SR questionnaire represent the thinking processes associated with SR to a substantial extent. We suggest that practitioners and researchers consider these item features in future adaptations of the SR questionnaire. This will most likely enable them to draw valid conclusions about participants' SR skills and to gain a deeper understanding of participants' SR skills in CS project evaluation.

CORRESPONDING AUTHOR:
Till Bruckermann
Leibniz University Hannover, DE; IPN – Leibniz Institute for Science and Mathematics Education, DE
till.bruckermann@iew.uni-hannover.de

KEYWORDS:
scientific reasoning; assessment; explanatory Rasch model; evaluation; learning outcomes; science inquiry skills

TO CITE THIS ARTICLE:
Bruckermann, T, Straka, TM, Stillfried, M and Krell, M. 2021. Context Matters: Accounting for Item Features in the Assessment of Citizen Scientists' Scientific Reasoning Skills. *Citizen Science: Theory and Practice*, 6(1): 21, pp. 1–15. DOI: *https://doi.org/10.5334/cstp.309*

INTRODUCTION

Growing numbers of citizen science (CS) projects engage citizens in scientific research not only to collect and process large data sets (e.g., Zooniverse projects; Cox et al. 2015) but also to promote individual learning outcomes (ILOs) (Jordan, Ballard, and Phillips 2012). Enhanced participation of citizens in scientific research ultimately should promote not only knowledge of science but also science inquiry skills that can include skills related to scientific reasoning (SR) (Phillips et al. 2018). SR skills refer to the ability to solve a scientific problem in a particular situation by applying a set of scientific skills and knowledge, for example, to form hypotheses (Lawson et al. 2000; Bao et al. 2009). While science inquiry skills comprise all abilities that are required for tasks in the scientific endeavor, only some skills, such as designing investigations and analyzing data, are related to SR (Stylinski et al. 2020). Some science inquiry skills, such as identifying a species or taking measurements for data collection, are more common to CS projects. Fewer CS projects require skills that are related to SR, such as forming hypotheses, because only stronger commitment may facilitate those skills (NASEM 2018). Hence, SR skills comprise a subset of skills that are less common to CS projects than other science inquiry skills (Stylinski et al. 2020). In CS projects that involve participants in inquiry approaches for learning (e.g., Aristeidou et al. 2020), SR skills might foster the achievement of other ILOs (Edwards et al. 2017), for example, behavioral beliefs (Bruckermann et al. 2021).

Resources for evaluating CS projects' ILOs are scarce (Bonney et al. 2016), and there is a need for reliable and valid instruments to assess SR skills (Stylinski et al. 2020). Notably, in the evaluation of ILOs in CS, there is (1) a lack of clarity concerning the constructs, and (2) a lack of resources, time, and social science expertise for assessment (Phillips et al. 2018). While overarching evaluation frameworks regarding the assessment of ILOs exist (e.g., DEVISE; Phillips et al. 2014), only a few instruments to assess science inquiry skills are available, and SR skills are mentioned in only one percent of the literature reviewed (Stylinski et al. 2020). Therefore, the evaluation of SR skills in CS projects less often relies on standardized tests than on surveys of self-reported confidence in performing the SR skills (see overview in Stylinski et al. 2020)—despite validity concerns regarding self-reports (Critcher and Dunning 2009). To ensure that conclusions that are drawn from evaluations of ILOs in CS are valid, assessment instruments that do not rely solely on self-reports should be developed (Phillips et al. 2018).

With regard to assessing SR skills, several instruments have been proposed for formal education contexts (Hammann et al. 2008; Hartmann et al. 2015; Krell 2018; see overview in Opitz, Heene, and Fischer 2017). In most of those instruments, the items typically include some problems or background stories to contextualize the particular assessment, because SR depends on knowledge of the respective discipline (Fischer et al. 2014). However, there is little evidence for the validity of the instruments in the formal education context (Opitz, Heene, and Fischer 2017). Furthermore, the instruments typically do not contextualize items to represent various disciplines and contexts in CS research. The development of valid assessment instruments in CS projects faces the challenge that CS occurs in various contexts (e.g., astronomy, medicine, and biology; Follett and Strezov 2015). To develop assessment instruments appropriate to the variety of CS project contexts, the instruments often have to be purposefully designed for the specific project (Cronje et al. 2011).

The purpose of the study reported here is to describe the development of a multiple-choice scientific reasoning questionnaire (hereafter SRQ) that can be flexibly adapted to the different contexts of CS research. More specifically, assumptions about the cognitive processes underlying SR and, thus, about the participants' processing of the items, guided the development of the SRQ. Empirically, this study provides evidence that item features requiring specific cognitive processes of SR significantly contribute to the difficulty of multiple-choice items. Further, CS practitioners might benefit from a validated assessment instrument that could provide insights into the SR skills of CS project participants.

THEORETICAL BACKGROUND

The present study refers to three SR skills—forming hypotheses, testing hypotheses, and analyzing data—and takes a cognitive perspective on SR (Klahr and Dunbar 1988). The SR skill of forming hypotheses requires individuals to understand which hypotheses can be tested by a particular research design. Testing hypotheses requires individuals to develop a research design that is valid to test a particular hypothesis. Analyzing data refers to the skill of drawing a valid conclusion based on a particular research design and the data obtained from this research design. The cognitive perspective on SR assessments adopted in this study aims to explore individuals' thinking processes to make SR skills accessible for assessment purposes. The sociocultural perspective, in contrast, would provide a rationale on how SR historically developed to be a cultural product in different contexts (e.g., Kind and Osborne 2017). Our decision to adopt a cognitive perspective on SR assessments was further motivated by both the lack of construct clarity in previous research and the lack of resources available for developing valid assessments of participants' SR skills (Stylinski et al. 2020).

ITEM FEATURES AND COGNITIVE PROCESSES IN ASSESSMENTS OF SR

Assessments should represent the processes and strategies necessary for participants to perform on tasks that test a psychological construct (this is known as construct representation; e.g., Embretson 1983). SR skills depend on cognitive processes—such as identifying the variables under investigation (i.e., information encoding and retrieval)—as well as on the use of cognitive strategies—such as controlling several variables to avoid biases in the investigation (e.g., the control-of-variables strategy) (Morris et al. 2012). Items that assess SR skills usually include a scientific problem that can be solved by identifying the relevant variables and controlling other variables to facilitate unbiased conclusions. Previous research on SR questionnaires has explored how particular item features—such as the SR skill being investigated, the number of independent variables, and the research context—influence the thinking processes involved in the identification and control of variables (Krell 2018; Mannel, Walpuski, and Sumfleth 2015). In the development of SR items, researchers have to account for those item features so that the SR assessment instrument represents those cognitive processes and allows for valid interpretations of test scores (Hartig and Frey 2012). If the different item features are accounted for, it is possible to calculate the influence of each item feature in relation to the items' overall difficulty. Significant sources of item difficulty that are not related to the psychological construct pose a threat to validity because they suggest that other abilities are needed to solve the item in addition to the intended abilities. Hence, the identification of such sources of item difficulty has the potential to improve the validity of assessments. Moreover, the identification of sources of item difficulty that are related to the psychological construct can guide item development (Messick 1995).

In the construction of valid assessments, previous research distinguished between two kinds of item features (Opfer, Nehm, and Ha 2012). First, deep-structure item features are held constant across all items because they aim to assess the cognitive processes and strategies related to SR, such as the identification and control of variables. Second, item surface features embed items in specific contexts of CS research designs (e.g., Follett and Strezov 2015). Individuals with high-level SR skills master the assessment despite the varying contexts (i.e., item surface features), but individuals with low-level SR skills are more likely to be distracted by such item surface features (Opfer, Nehm, and Ha 2012). If item features and the related cognitive processes are identified, it is possible to explain how an assessment instrument works, and this contributes to construct validity (Fischer 1995, 2005; Hartig and Frey 2012). Hence, we explore how different item features that indicate the cognitive processes required to solve the item contribute to item difficulty in SR assessments. From previous research, we identified two deep-structure item features that are essential for the assessment of SR: (1) the feature that one of the three different SR skills (i.e., forming hypotheses, testing hypotheses, and analyzing data) is required to solve the item, and (2) the feature that the number of independent variables (i.e., one or two independent variables) has to be accounted for to solve the item. Furthermore, three item surface features that might distract participants from successfully applying their SR skills were examined: research context, text complexity, and the use of specialist terms.

The three SR skills of forming hypotheses, testing hypotheses, and analyzing data are deep-structure item features—they relate to the cognitive processes and strategies of identifying and controlling variables—and have been shown to significantly influence item difficulty (Hammann et al. 2008; Mannel, Walpuski, and Sumfleth 2015; Krell 2018). Based on comparisons of item difficulties in the formal education context, research suggested that the SR skill of testing hypotheses requires different knowledge than forming hypotheses and analyzing data: the SR skills of forming hypotheses and analyzing data seem to require profound domain-specific content knowledge, while the SR skill of testing hypotheses is more closely, but not exclusively, related to knowledge of the processes (Hammann et al. 2008). Furthermore, previous research indicates that assessment items on forming hypotheses and testing hypotheses typically presuppose one part of the inquiry as given (i.e., the items provide either the research design or the hypothesis). For example, assessment items on testing hypotheses provide a hypothesis and ask the participant to propose a valid research design to test it. Items on data analysis, however, require participants to relate two parts of the inquiry process, that is, the research design and the observations (Krell 2018). In the formal learning context, studies revealed that assessment items on forming hypotheses and testing hypotheses are typically easier to solve for participants than assessment items on analyzing data (e.g., Krell 2018). In CS projects, all three SR skills are neither easy for participants nor common so that we aimed at comparing the item difficulties for the three SR skills in the informal education context of a CS project. Testing hypotheses will serve as the reference category in our analysis, that is, the item difficulty of the SR skills of forming hypotheses and analyzing data will be compared against it.

Item complexity is the second deep-structure item feature we identified. Item complexity in SR assessments is defined as the number of variables that individuals need to keep in mind to answer an item, that is, whether the

hypothesis, the research design, or the data refer to one or to two independent variables (Mannel, Walpuski, and Sumfleth 2015). Thinking of more than one independent variable at once increases the cognitive load, that is, the amount of information that individuals need to process (Kauertz et al. 2010). Therefore, the item complexity contributes to the item difficulty in an assessment of SR skills (e.g., Kauertz et al. 2010; Krell 2017).

Although the investigation of hypotheses as one form of scientific reasoning spans the sciences, SR skills must be applied in the various contexts of research. The research context is considered an item surface feature because individuals can apply their SR skills to different contexts, while the underlying thinking processes of variable identification and control remain the same. Previous research has indicated that SR also depends on domain-specific knowledge (Fischer et al. 2014). Individuals need to have domain-specific knowledge of the respective research context to identify the investigated variables and to represent them in a mental model (Morris et al. 2012). Domain-specific knowledge enables the adequate representation of variables that are relevant to problem-solving (Fischer et al. 2014). Especially unfamiliar contexts have been shown to make items more difficult to solve (Le Hebel et al. 2017). Hence, the context of the items on SR also contributes to the items' difficulty (Krell 2018).

Two further item surface features that can contribute to the items' difficulty are text complexity and the use of specialist terms. In the natural sciences, language is determined by its functional grammar, including specialist terms and complex sentences (Fang 2006). In SR assessments, items typically include a text-based description of a problem that can be solved by applying SR skills. The problem description often employs specialist terms and words and sentences with an above-average length. These constructs are due to the functional grammar that is used in the language of the natural sciences. The length of words and sentences (i.e., text complexity) and the use of specialist terms are both considered item surface features as they influence the readability of scientific texts. Individuals have to follow the grammar in the text and understand the specialist terms to be able to represent the problem mentally. Both text complexity and specialist terms seem to influence the item difficulty (Stiller et al. 2016; Krell, Khan, and van Driel 2021).

This study describes the development of a flexibly adaptable SRQ that systematically considers two deep-structure item features (the skills of forming hypotheses, testing hypotheses, and analyzing data; the two levels of item complexity) and three item surface features (the research context; the text complexity; the use of specialist terms). We compared the fit of statistical models on SR skills to provide empirical evidence for the item features' contribution to item difficulty. To do so, we first tested a descriptive Rasch model (one-parameter logistic model [1PLM]) that does not differentiate between the item features. Then, we compared the descriptive Rasch model against a basic linear logistic test model (LLTM) that accounts for three of the item features, that is, the SR skills, the levels of complexity, and the research contexts, and against an extended LLTM that additionally takes text complexity as well as specialist terms into account. Our research provides valuable insights for the increasing CS community and researchers who aim to assess participants' SR skills in CS projects by suggesting a blueprint and guidelines for the development of SRQs, using item features that can be adapted to different CS-relevant contexts.

METHODS

The study reported here is part of an interdisciplinary research project on CS, which comprised three CS projects ([1] urban wildlife ecology, [2] urban bat ecology, and [3] urban air pollution) in two large cities (Berlin and Leipzig) in the east of Germany. Each CS project involved several time-limited runs of data collection and analysis per year (hereafter field seasons). Although the three CS projects differed in their research context, all projects followed the same goal; that is, the examination of distribution patterns. To investigate participants' SR skills, we developed an SRQ accounting for the overarching factors of the underlying construct of SR skills (i.e., forming hypotheses, testing hypotheses, and analyzing data) while addressing the differing contexts of research (i.e., wildlife, bats, and air pollution). We report on data from two field seasons of one CS project (urban wildlife ecology) in the city of Berlin in which we assessed participants' SR skills. We chose to assess SR skills in only two field seasons of the urban wildlife ecology CS project because we did not want to overburden the participants by asking them to answer several questionnaires.

INSTRUMENT DEVELOPMENT

We applied an established blueprint for the systematic development of a multiple-choice SRQ (Krell 2018). The blueprint accounts for two deep-structure item features by addressing three skills of SR and two levels of item complexity. Furthermore, the blueprint allows for the contextualization in our particular CS projects (i.e., on wildlife ecology, bat ecology, and air pollution) by adapting the surface features to different research contexts. Finally, the blueprint guides the structure of the language (i.e., text complexity and specialist terms) and the figures (example in ***Figure 1***; see Supplemental file 1: Appendix 1 for the blueprint used in this study).

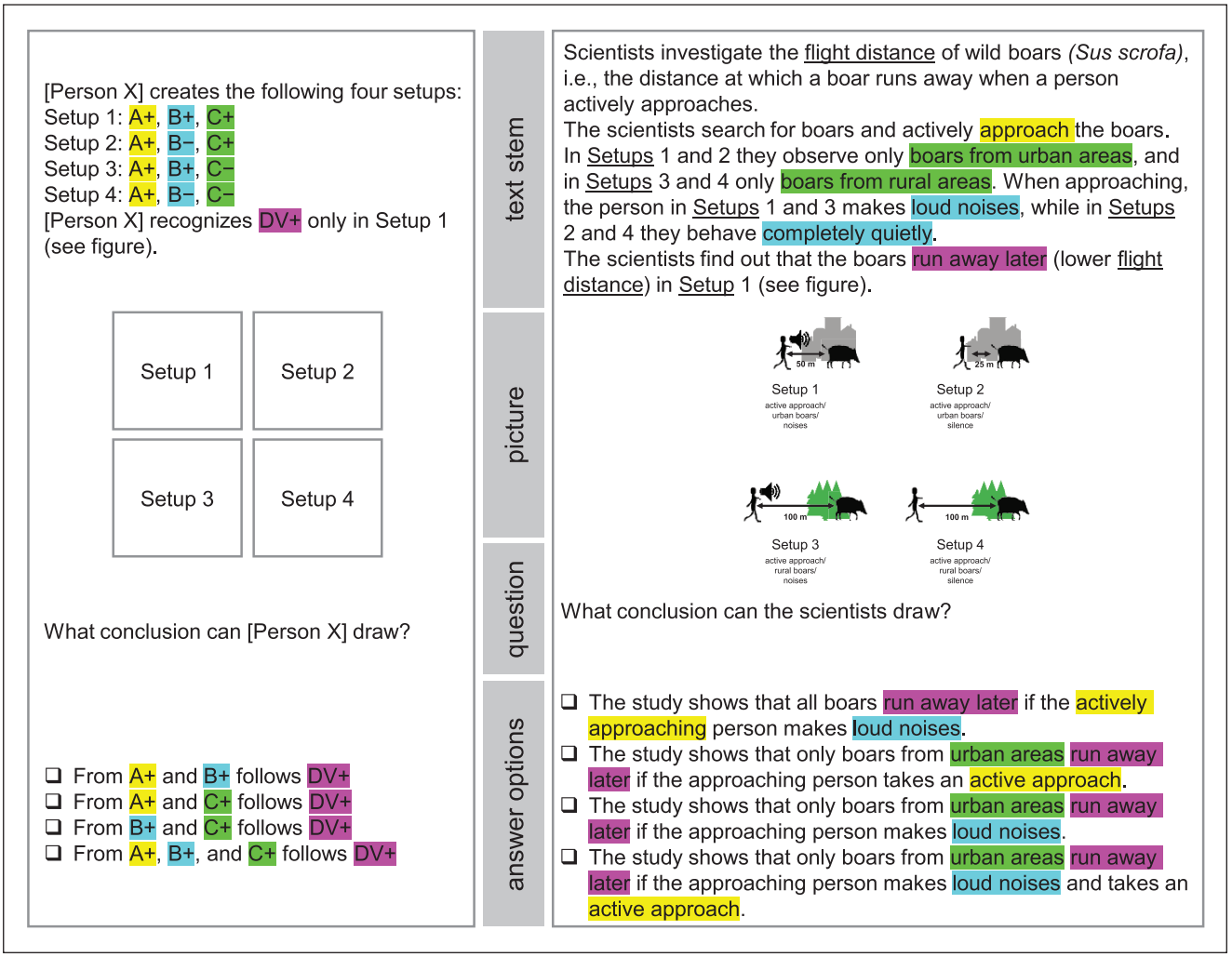

Figure 1 Example of the structure of an item in the blueprint (deep-structure item feature: analyzing data, high item complexity) and its adaptation for Item 6 in the SRQ (item surface feature: context of urban wildlife ecology). Color coding represents the corresponding variables in the blueprint (*left*) and the example (*right*). Underlined words are specialist terms in this example.

To adapt the blueprint and contextualize it in authentic CS research, experts first identified research designs within actual research on the respective topics (e.g., flight distance of urban wild boars: Stillfried et al. 2017; effect of artificial light at night and tree cover on bats: Straka et al. 2019). Second, the experts reviewed the research regarding its central variables, hypotheses, design, and the data obtained from this research. Third, we adopted the respective variables of the chosen research contexts to each of the three SR skills (i.e., forming hypotheses, testing hypotheses, and analyzing data). To differ between item complexities, we varied the number of independent variables under consideration by using two levels, that is, the consideration of one or two independent variables (i.e., low and high item complexity). The SRQ comprised three contexts (wildlife ecology, bat ecology, and air pollution) for three SR skills (forming hypotheses, testing hypotheses, analyzing data) and two item complexity levels (one independent variable and two independent variables). The complete crossing of the three contexts, three SR skills, and two complexity levels resulted in 3 × 3 × 2 = 18 items in total (***Table 1***).

With regard to the remaining two item surface features, we did not purposefully vary the length of words and sentences or the specialist terms between the items; the blueprint aimed to keep the complexity of language comparable for all items. All 18 items had a comparable structure (see example in ***Figure 1***). First, the text stem introduced a research design with all relevant dependent and independent variables. Second, the picture represented the setup of this research and named all independent variables. Third, the question prompted participants to provide a valid hypothesis, suggest an additional setup, or draw a valid conclusion. Fourth, each item provided four answer options.

Although we aimed to keep the complexity of language comparable, the different research contexts required using words and terms of different length and familiarity in the

	ITEM COMPLEXITY (NUMBER OF INDEPENDENT VARIABLES)	
	LOW = ONE VARIABLE	HIGH = TWO VARIABLES
Scientific reasoning skill		
Forming hypotheses	Wildlife ecology (item 1) Bat ecology (item 7) Air pollution (item 13)	Wildlife ecology (item 4) Bat ecology (item 10) Air pollution (item 16)
Testing hypotheses	Wildlife ecology (item 5) Bat ecology (item 11) Air pollution (item 17)	Wildlife ecology (item 2) Bat ecology (item 8) Air pollution (item 14)
Analyzing data	Wildlife ecology (item 3) Bat ecology (item 9) Air pollution (item 15)	Wildlife ecology (item 6) Bat ecology (item 12) Air pollution (item 18)

Table 1 Specification of item features for item numbers 1–18 in the development of the SRQ.

assessment items. We analyzed the text complexity and the use of specialist terms to control for their effects on item difficulty. To monitor the influence of text complexity on item difficulty, we calculated the Flesch Reading Ease Index (FRE; Flesch 1948) in its German adaptation that accounts for the mean sentence length and the mean number of syllables per word. For the FRE, values below 60 indicate a high text complexity, that is, sentences are longer and a word has more syllables. Furthermore, we counted the percentage of specialist terms (ST) in every item because the ability to identify the variables being investigated (i.e., the cognitive process of information encoding) also depends on knowledge of specialist terms. We formed a list of specialist terms that are not commonly used in everyday language (e.g., transect, particulate measure, flight distance) and consistently applied it to all items. The number of specialist terms varied across the 18 items depending on the respective research contexts. The less tangible research context of air pollution used more specialist terms than the research contexts of wildlife ecology and bat ecology. More than seven specialist terms in 100 words (ST > 7%) are considered cognitively demanding (Kulgemeyer and Starauschek 2014).

PARTICIPANTS

Participants were recruited via media sources such as radio, newsletter, or posters in public places. They could apply to participate in one of the two field seasons of the CS project. Given the diversity of sociodemographic factors within the city's districts and to ensure an equal distribution of participants across the city, citizens were selected for participation based on the location where they lived. The participants were evenly distributed across the districts of the city of Berlin by the design of this study. $N = 374$ citizens participated, of which 198 were identified as being female and one as having a non-binary gender. Their mean age was $M = 53.22$ ($SD = 11.92$; range: 25–81). As in many CS projects, this sample was well educated, with most participants holding an upper secondary school certificate (82.6%) and fewer participants holding a certificate from the upper secondary vocational track (15.5%). Furthermore, more than half of the participants held a university degree (59.9%) and some also even had a doctoral degree (11.5%).

PROCEDURE

To participate in one of the two field seasons, participants signed up on an online platform. For two months, the participants formed an online community to share and analyze the data they had collected, as well as to discuss their findings with other participants. Participants filled in the questionnaire before and after they took part in the field season. In this study, we report on data that was collected from two field seasons, one in fall 2018 and one in spring 2019. In detail, we analyzed the answers of participants to the SRQ before the project. The reason for analyzing the data (i.e., participants' answers to the questionnaire) collected before participation in the CS project was that this assured that the SR skills assessed had not been explicitly trained by participation in the CS project. Participants gave their informed consent for this study, and an external ethics board approved the SRQ.

DATA ANALYSIS

To estimate how the different item features contributed to the items' difficulty, we applied the LLTM. The LLTM assumes that item difficulty is a linear combination of the different item features (Fischer 1995, 2005). The LLTM belongs to the Rasch models, a family of established psychometric models applied in psychological and educational research (Embretson and Reise 2000). The family of Rasch models includes descriptive psychometric models, such as the 1PLM, which allows for the holistic estimation of individual person ability (θ_s) and item difficulty (β_i) parameters. In the 1PLM, it is assumed that the probability of a correct item

response depends only on θ_s and β_i (Embretson and Reise 2000):

$$P(X_{is}) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)}.$$

In contrast to descriptive models such as the 1PLM, explanatory models consider different item features to estimate each feature's influence analytically (Wilson, Boeck, and Carstensen 2008). From this perspective, the LLTM can be seen as an item explanatory model because it replaces the β_i parameter with a linear combination of the basic parameters α_k : $\beta_i' = \sum_{k=1}^{N}(\alpha_k \chi_{ik})$ (Fischer 1995). The LLTM splits up the item difficulty of whole items (i.e., β_i parameter in the 1PLM) into the individual contribution of different item features to the item difficulty (i.e., α_k parameters). Hence, if an LLTM can be shown to fit the given data, the estimated parameters α_k provide a measure of each item feature's contribution—such as the different SR skills, the levels of item complexity, and the research context—to the item difficulty.

To evaluate the model fit of an LLTM, a two-step procedure is proposed: First, the 1PLM has to fit "at least approximately" (Fischer 2005, p. 509) to the data because further decomposition should concern a unidimensional measure of SR skills. Second, the decomposition of β_i needs to be checked for empirical validity. For this purpose, item difficulty parameters estimated in the 1PLM and the LLTM can be compared (e.g., graphically or by calculating the Pearson correlation), assuming that they positively correlate (Baghaei and Kubinger 2015). Furthermore, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the log-likelihood difference test can be applied to compare the fit of both models, as well as the plausibility of different LLTMs (Fischer 2005; Wu, Adams, and Wilson 2007). The AIC and BIC are relative fit indices that allow model comparison, but they do not allow an absolute evaluation of model fit. The higher the AIC and BIC values, the more the data deviates from the specified model (Wu, Adams, and Wilson 2007). In the present study, we used the software ACER Conquest (Wu, Adams, and Wilson 2007) and the R package eRm (Mair and Hatzinger 2007) for parameter estimation.

RESULTS

SPECIFICATION OF A DESCRIPTIVE RASCH MODEL (1PLM)

We specified a one-dimensional 1PLM that reflected the view of SR as a general ability without disentangling the influence of different item features, for example, the SR skills or the text complexity. Hence, the 1PLM provides an estimation of item difficulty without an account of specific item features. The 1PLM showed appropriate mean square (MNSQ) item-fit statistics ($0.7 \leq \text{MNSQ} \leq 1.8$; not distorting measurement; Wright and Linacre 1994). The item separation reliability was very high ($\text{rel.}_{\text{SEP}} = .98$) and the person reliability was good ($\text{rel.}_{\text{EAP/PV}} = .74$) compared with previous SR assessments (Hartmann et al. 2015: 0.54; Mannel, Walpuski, and Sumfleth 2015: 0.76; Krell 2018: 0.64).

In the Wright map in ***Figure 2***, we inspected the distribution of item difficulties in the SRQ and the distribution of person ability in our sample on the same linear scale (equal-interval logits) as computed based on marginal maximum likelihood (MML) estimation. Higher logit values indicate a higher person ability (see ***Figure 2***; green dots represent

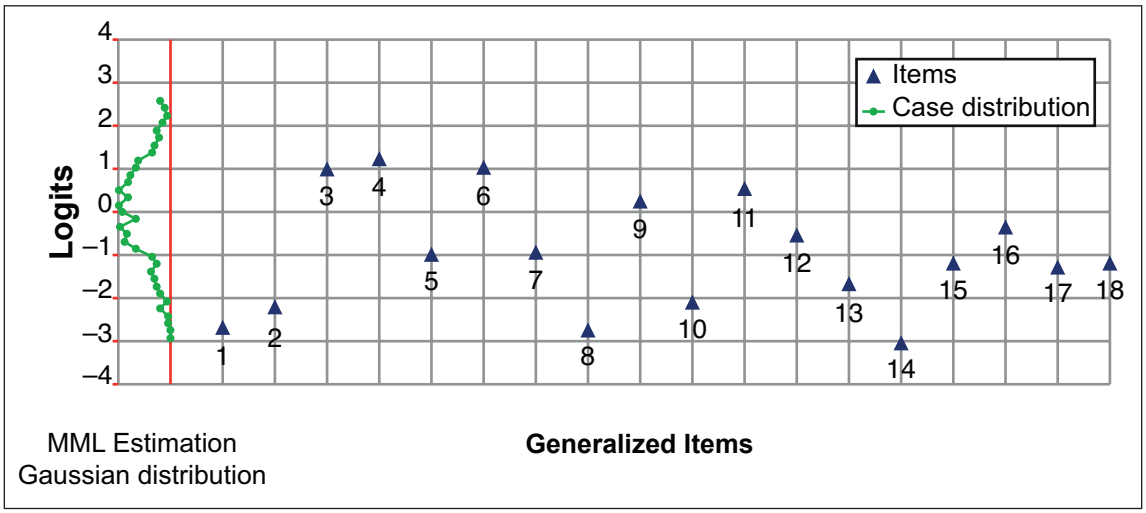

Figure 2 Wright map of $N = 374$ participants' scientific reasoning (SR) abilities (marginal maximum likelihood [MML] estimation) and item difficulties (logits) for items 1–18 in the one-parameter logistic model (1PLM). Table 1 presents the item features for items 1–18 that were varied on purpose.

distribution of participants). The estimated person ability followed a Gaussian distribution. Furthermore, higher logit values indicate a greater item difficulty. For example, Item 4 (i.e., forming hypotheses with two independent variables in the context of urban wildlife ecology; see Supplemental file 2: Appendix 2) was the most difficult item with a difficulty of 1.25 logits (see ***Figure 2***, blue triangles). Based on their difficulty, the items were evenly scattered across the person ability in our sample.

SPECIFICATION OF AN EXPLANATORY RASCH MODEL (LLTM)

To further explain the difficulty of items, we specified two LLTMs, that is, a basic model and an extended model (see ***Table 2*** for item features that are included in the models). The smaller values obtained from the AIC and the BIC suggested a better fit between model and data for the extended model compared with the basic model (***Table 2***), as did the log-likelihood difference test ($p < .001$); however, the extended model still showed an inferior fit compared with the 1PLM based on both the smaller information criteria AIC and BIC and the significant log-likelihood difference test ($p < .001$). These findings indicate the least deviation between model and data for the 1PLM, followed by the extended model and the basic model.

The item difficulty parameters estimated in the 1PLM positively correlated with those estimated in the basic model ($r = .58$, $p = .011$, 95% CI [0.51, 0.64]) and the extended model ($r = .62$, $p = .006$, 95% CI [0.55, 0.68]). This means that about 34% (basic model: 95% CI [26, 41]) or 39% (extended model: 95% CI [30, 46]) of the individually estimated item difficulties in the 1PLM can be explained with the respective parameters specified in the LLTMs. The graphical model tests of the basic model in ***Figure 3a*** and the extended model in ***Figure 3b*** reveal that the items were scattered around the 45° line moderately well.

The estimated α_k parameters (***Table 3***) showed that all item features contributed significantly to the items' difficulty because their 95% CI did not include zero. For example, the SR skills of forming hypotheses and analyzing data seemed to be rather difficult (i.e., relatively high positive α_k parameter) compared with testing hypotheses, which served as the reference category in our comparison. As already found in the item parameters of the 1PLM, a higher item complexity reduced item difficulty (i.e., negative α_k parameter). The consideration of text complexity and specialist terms (item surface features) in the present study reduced the estimated effect of the context on item difficulty: The α_k parameters for the contexts of wildlife and of bats were smaller in the extended model compared with the α_k parameters for the contexts in the basic model. This indicates that the difficulty of different research contexts is to some degree related to the use of specialist terms and the complexity of the text that is used to describe the research context.

DISCUSSION

This research investigated the influence of item features on item difficulty in a scientific reasoning questionnaire (SRQ). The identification of item features that influence item difficulty is crucial in the assessment of the SR skills of CS participants. From the item features, it is possible to infer the thinking processes related to SR skills and to determine whether the SRQ assesses what it is supposed to assess (i.e., to obtain validity evidence for the assessment of SR). In line with our assumptions, we were able to provide statistical evidence on item features (in the SRQ) that represent the thinking processes in relation to SR such as forming hypotheses. The variance that is explained by the item features indicates that the item features stimulate thinking processes, for example, with regard to hypotheses formulation. The item features we proposed to be relevant for SR skills influence the item difficulty and hence could be used to represent participants' SR skills. Furthermore, we showed how deep-structure item features, namely three SR skills (i.e., forming hypotheses, testing hypotheses, and analyzing data) and two levels of item complexity (i.e., one and two independent variables), contribute to the items' difficulty. In addition, we found that the three research contexts, the text complexity, and the use of specialist

	SR SKILLS; ITEM COMPLEXITY; RESEARCH CONTEXT	TEXT COMPLEXITY; SPECIALIST TERMS	ESTIMATED PARAMETERS	DEVIANCE	AIC	BIC
1PLM			19	6,737	6,775	6,849
LLTM (basic model)	X		5	7,856	7,866	7,886
LLTM (extended model)	X	X	7	7,803	7,817	7,845

Table 2 Model-fit indices of the Rasch models specified in the present study.

Note: The "X" marks which item features are included in the respective model for analysis. SR: scientific reasoning; AIC: Akaike information criterion; BIC: Bayesian information criterion; PLM: parameter logistic model; LLTM: linear logistic test model.

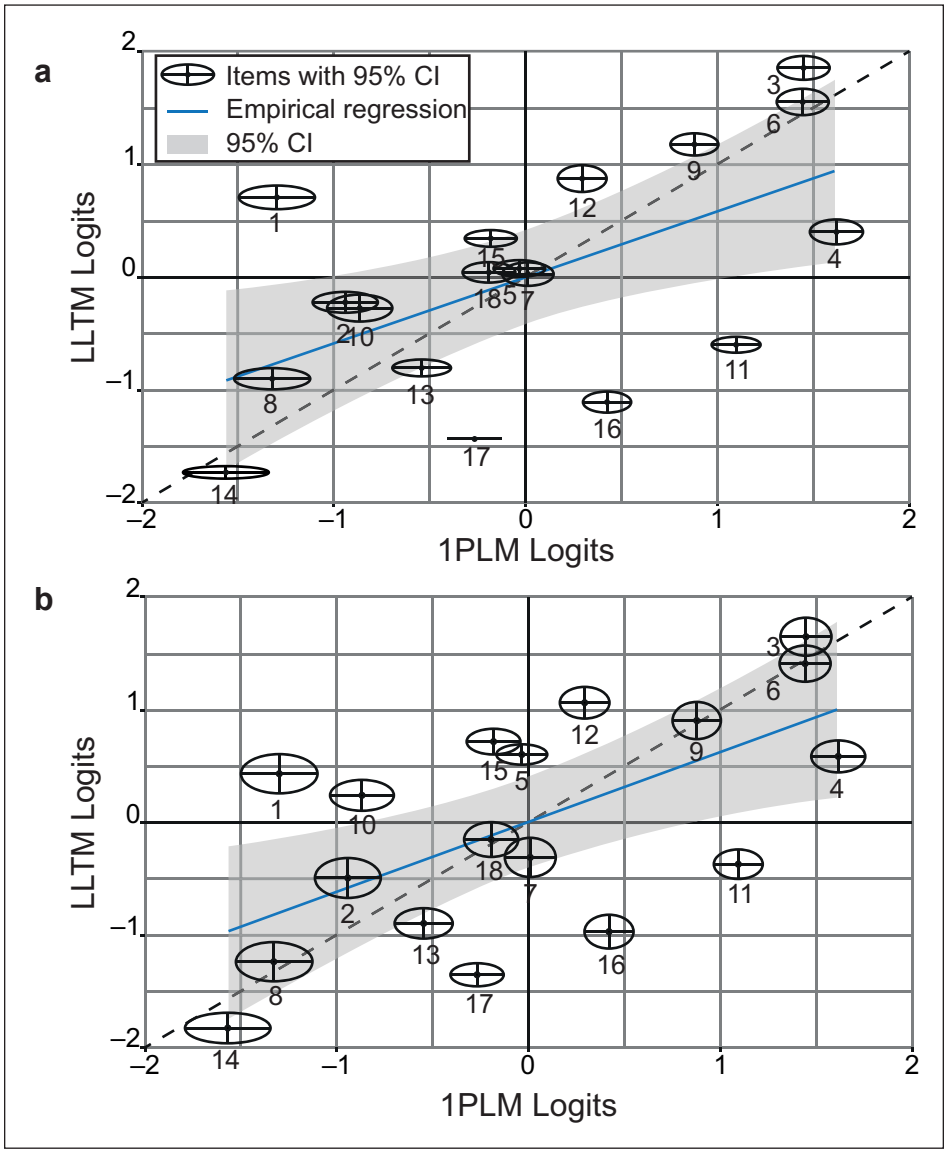

Figure 3 Graphical test of (*a*) the basic model and (*b*) the extended model that compares the item difficulties (logits) for items 1–18 between the descriptive one-parameter logistic model (1PLM) Rasch model (x-axis) and the explanatory linear logistic test model (LLTM) Rasch model (y-axis).

terms (i.e., item surface features) influence the items' difficulty. We were able to establish the item features that are crucial in SR assessment instruments to provide valid conclusions. The SR skills in the SRQ, however, do not equal the variety of science inquiry skills that might be required in other CS projects, such as identification of species. The systematic identification of item features that contribute to item difficulty in the SRQ provides guidelines for further flexible adaptation to the variety of CS contexts. Furthermore, it showcases a method to investigate item features in formal tests of science inquiry skills.

Our results corroborate previous research on the validity of SRQs by considering both the deep-structure item features and the item surface features of SR (i.e., SR skills, item complexity, and research context: Hammann et al. 2008; Stiller et al. 2016; Krell 2018). Furthermore, our results expand previous validity evidence that has been found for SR assessments in formal education (Hartmann et al. 2015; Stiller et al. 2016) to a sample of CS participants. In line with previous research (Stiller et al. 2016; Krell 2018), our explanatory modeling of item features explained a significant amount of the variance in item difficulty in the SR assessment. Our results from the basic model indicate that the item features accounted for 34% of the variance in participants' answers in the SR assessment (large effect: $R^2 > .25$; Hartig and Frey 2012). Previous validation studies of SR assessments explained comparable amounts of variance (43% for secondary school students: Krell 2018;

	BASIC MODEL				EXTENDED MODEL			
	α_k	$SE\alpha_k$	$CI_{95\%}$		α_k	$SE\alpha_k$	$CI_{95\%}$	
Research context								
Air pollution[1]	—	—	—	—	—	—	—	—
Wildlife (1 = yes)	1.04	0.07	1.18	0.91	0.69	0.09	0.86	0.52
Bats (1 = yes)	0.58	0.07	0.72	0.44	0.43	0.08	0.58	0.28
Scientific reasoning skill								
Testing hypotheses[1]	—	—	—	—	—	—	—	—
Forming hypotheses (1 = yes)	0.43	0.07	0.57	0.29	0.34	0.07	0.48	0.19
Analyzing data (1 = yes)	1.23	0.07	1.37	1.09	1.23	0.07	1.37	1.08
Item complexity (number of independent variables)								
Low[1]	—	—	—	—	—	—	—	—
High (1 = two variables)	−0.21	0.06	−0.10	−0.32	−0.35	0.07	−0.22	−0.47
Text complexity								
Low[1]								
High (1 = FRE < 60)	—	—	—	—	−0.46	0.08	−0.31	−0.61
Specialist terms								
Low[1]	—	—	—	—	—	—	—	—
High (1 = ST > 7%)	—	—	—	—	−0.29	0.07	−0.16	−0.43

Table 3 Estimated parameters in the LLTMs with standard error (*SE*) and confidence interval (CI).

Note: FRE: Flesch Reading Ease Index; ST: specialist terms.

[1] We compared estimated α_k parameters with the values of this reference category.

32% for university students: Stiller et al. 2016) despite validating the assessments in more homogenous samples with regard to participants' age range and formal training in SR. The substantial effect size found in our study indicates that the item features represent SR and that the interpretation of test scores to draw conclusions about the SR skills of participants in this sample of citizen scientists is, therefore, valid to a considerable extent (Hartig and Frey 2012). Adapting the SRQ for further research will provide other researchers with valid assessment scores if they account for the item features presented here in their questionnaires. Although validity is not a stable characteristic of an assessment instrument as it depends on the specific sample and assessment situation, accounting for item features makes it more likely that other researchers will reproduce our findings in their sample. The adaptation to other CS samples is feasible because our findings provide evidence that the item features explained an amount of variance in the assessment of SR comparable to the results from the samples of secondary school students and university students. The adaptation to participants in a CS project and the contextualization in research designs from different disciplines did not pose a threat to the validity of the conclusions about participants' SR skills that were drawn from the assessment instrument.

Regarding the deep-structure item features in SR assessments, our findings correspond to previous research on the influence of SR skills and item complexity (Hammann et al. 2008; Mannel, Walpuski, and Sumfleth 2015; Krell 2018). However, our findings extend previous research concerning the effects of item surface features, such as research contexts that better represent CS research as well as text complexity and specialist terms. Considering text complexity and specialist terms, the extended model explained another 5% of variance (39%). Furthermore, adding text complexity and specialist terms as item features to the basic model influenced the previously tested contribution of the research context in the extended model. We discuss the effects of the different item features in the following.

The research context matters in the assessment of SR in CS projects because applying SR skills in different contexts affects the item difficulty. When assessing SR skills in the context of a particular project, the participants' scores might not be comparable to participants' scores in CS projects that

probed their SR skills in another context. For participants, SR might be more difficult in some CS projects than in others, depending on the context. Our results build on previous findings on SR item contextualization in research designs from school curricula (e.g., Le Hebel et al. 2017; Krell 2018) and extend them to research designs that are more authentic for CS (e.g., wildlife ecology). Although we tested the same SR skills in all items, the varying contexts in which the SR skills were applied affected the item difficulty. We suggest that the development of SR assessments in CS projects should account for the different research contexts in the items because knowledge of the respective research domain is likely to influence the item difficulty.

The item features SR skills and item complexity contributed to the items' difficulty, in line with previous research (Hammann et al. 2008; Mannel, Walpuski, and Sumfleth 2015; Krell 2018). For participants in CS projects, questions are difficult to answer depending on the particular SR skill and the item complexity. Questions on the SR skills of analyzing data from a given research design and forming a hypothesis are more difficult than questions on the SR skill of testing hypotheses. In line with previous research, we assume that testing hypotheses probably requires stronger procedural knowledge, whereas forming hypotheses and analyzing data rely more heavily on domain-specific content knowledge (Hammann et al. 2008). Our results show that even without formal training, the items on some SR skills, such as hypotheses testing, are more easily mastered by participants in CS projects than others. Therefore, when evaluating SR skills, researchers might find more pronounced individual learning outcomes for SR skills that challenge participants less in the assessment. These findings also correspond to actual participation in CS projects because participants less frequently engage with forming hypotheses or analyzing data (Phillips et al. 2019), be it for motivational or cognitive reasons. Even though the number of variables (i.e., item complexity) contributed to item difficulty, this effect's direction cannot be determined. We further discuss this effect in the Limitations section.

Considering the text complexity and use of specialist terms in the SR assessment, we found that both item features affected item difficulty and reduced the research context's effect on item difficulty. Although, in previous research, an explanatory modeling of the items in an SR assessment indicated that text complexity and specialist terms influence the item difficulty (e.g., Stiller et al. 2016; Krell, Khan, and van Driel 2021), the systematic development of SR assessments has not yet considered the language aspect (Krell 2018). Our findings indicate that text complexity and specialist terms impact the research context's influence on item difficulty. Obviously, the cognitive processing of both the specialist terms and the research context relies on domain-specific knowledge (Le Hebel et al. 2017). We recommend that the language used in different research contexts should be accounted for in SR assessment because this reduces the research context's effect on item difficulty.

LIMITATIONS

Despite the significant contribution of the item features examined in our study, another 61% of variance remained unexplored. Future research should explore further item features that rely on the cognitive processes of solving items that require SR. For example, in this study, we did not consider how pictorial representations influence the item difficulty in SR assessments. At least for students, pictures may reduce the difficulty of items as they reduce the cognitive effort required to construct a mental model of the problem (Lindner et al. 2018).

Furthermore, our sample is a convenience sample from two field seasons of a CS project, and we did not compile it based on theoretical considerations. The participants who were interested in the project were also quite well educated. Although this sample is comparable to other CS projects (e.g., Trumbull et al. 2000), some participants' expert status might have led to the counterintuitive finding that items with two independent variables were less difficult. Highly skilled participants might perceive the variation of only one variable as easy and, therefore, be prompted to invest less thinking effort in the task. Further validation of the SRQ in more heterogeneous samples of CS participants should be put forward in further studies.

IMPLICATIONS

Our research provides practical implications for evaluating ILOs in CS (Jordan et al. 2012) as it shows how item features in questionnaires influence the item difficulty. We suggest that practitioners and researchers in CS account for the different SR skills and the number of variables when developing questionnaires to evaluate participants' SR skills for the investigation of hypotheses. Regarding SR skills that have been less common in evaluations of CS projects, such as forming hypotheses (Stylinski et al. 2020), the blueprint might help to standardize assessment instruments in the different research contexts. We further suggest accounting for the research context in which SR skills have to be applied. The research context influences the item difficulty in our study—in addition to the deep-structure item features that directly relate to SR.

In our study, the explanatory modeling based on item features provided evidence for the validity of the assessment. Following our theoretical assumptions, our results indicate that the research context, the SR skills,

and the item complexity accounted for 34% (or 39% with the item surface features text difficulty and specialist terms added) of the item difficulty that individual citizens encountered while solving the assessment items on SR. The substantial amount of variance explained can be traced back to the systematic development of the SRQ in regard to item features. We suggest that practitioners use this blueprint when adapting the SRQ to the research context and participant sample of their CS project.

Furthermore, the analysis of item features revealed that the item difficulty differs depending on the SR skills, the item complexity, and the research context in an assessment of SR. This confirms that the development of items and the interpretation of test scores in SR assessments should consider the particular item features. When comparing participants' proficiency in SR across CS projects that assess different SR skills or SR in different disciplines, researchers should consider that the test items are not equally difficult. Similar to our overview of item features in this SR assessment on the investigation of hypotheses, future research should describe which science inquiry skills were tested in which contexts and using how many variables in the assessment. Given the number of research contexts and the different science inquiry skills addressed by CS projects, further research on the evaluation of SR in samples of citizen scientists should systematically explore item feature effects.

DATA ACCESSIBILTY STATEMENT

The research data is available upon request to the corresponding author because the data analyses are still ongoing.

SUPPLEMENTARY FILES

The supplementary files for this article can be found as follows:

- **Supplemental file 1: Appendix 1.** Blueprint for the development of a scientific reasoning questionnaire (based on Krell 2018). DOI: *https://doi.org/10.5334/cstp.309.s1*
- **Supplemental file 2: Appendix 2.** Exemplary items. DOI: *https://doi.org/10.5334/cstp.309.s2*

ETHICS AND CONSENT

Participants gave their informed consent for this study, and an external ethics board approved the SRQ.

ACKNOWLEDGEMENTS

We want to thank Gráinne Newcombe for language review.

FUNDING INFORMATION

This study was supported by the German Federal Ministry of Education and Research (grant numbers 01|O1727, 01|O1725). The publication of this article was funded by the Open Access Fund of the Leibniz Universität Hannover. The funding sources were involved neither in conducting the research nor in preparing the article.

COMPETING INTEREST

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualization, M.K. and T.B.; methodology, M.K. and T.B.; formal analysis, M.K.; investigation, M.S., T.B., and T.M.S.; resources, M.S., T.B., and T.M.S.; data curation, M.K. and T.B.; writing – original draft preparation, M.K., T.B., and T.M.S.; writing – review and editing, M.K., M.S., T.B., and T.M.S.; visualization, T.B.

AUTHOR AFFILIATIONS

Till Bruckermann *orcid.org/0000-0002-8789-8276*
Leibniz University Hannover, DE; IPN – Leibniz Institute for Science and Mathematics Education, DE
Tanja M. Straka *orcid.org/0000-0003-4118-4056*
Technische Universität Berlin, DE
Milena Stillfried
Leibniz Institute for Zoo and Wildlife Research, DE
Moritz Krell *orcid.org/0000-0003-2226-0383*
IPN – Leibniz Institute for Science and Mathematics Education, DE

REFERENCES

Aristeidou, M, Scanlon, E and **Sharples, M.** 2020. Learning outcomes in online citizen science communities designed for inquiry. *International Journal of Science Education*, Part B 10(4):277–294. DOI: *https://doi.org/10.1080/21548455.2020.1836689*

Baghaei, P and **Kubinger, KD.** 2015. Linear Logistic Test Modeling with R. *Practical Assessment, Research & Evaluation*, 20(1).

Bao, L, Cai, T, Koenig, K, Fang, K, Han, J and **Wang, J,** et al. 2009. Learning and scientific reasoning. *Science*, 323(5914): 586–587. DOI: *https://doi.org/10.1126/science.1167740*

Bonney, RE, Phillips, TB, Ballard, HL and **Enck, JW.** 2016. Can citizen science enhance public understanding of science? *Public Understanding of Science*, 25(1): 2–16. DOI: *https://doi.org/10.1177/0963662515607406*

Bruckermann, T, Greving, H, Schumann, A, Stillfried, M, Börner, K, Kimmig, SE, Hagen, R, Brandt, M and **Harms, U.** 2021. To know about science is to love it? Unraveling cause-effect relationships between knowledge and attitudes toward science in citizen science on urban wildlife ecology. *Journal of Research in Science Teaching*, 58(8): 1179–1202. DOI: *https://doi.org/10.1002/tea.21697*

Cox, J, Oh, EY, Simmons, B, Lintott, C, Masters, K and **Greenhill, A,** et al. 2015. Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects. *Computing in Science & Engineering*, 17(4): 28–41. DOI: *https://doi.org/10.1109/MCSE.2015.65*

Critcher, CR and **Dunning, D.** 2009. How chronic self-views influence (and mislead) self-assessments of task performance: Self-views shape bottom-up experiences with the task. *Journal of Personality and Social Psychology*, 97(6): 931–945. DOI: *https://doi.org/10.1037/a0017452*

Cronje, R, Rohlinger, S, Crall, A and **Newman, G.** 2011. Does participation in citizen science improve scientific literacy?: A study to compare assessment methods. *Applied Environmental Education & Communication*, 10(3): 135–145. DOI: *https://doi.org/10.1080/1533015X.2011.603611*

Edwards, R, McDonnell, D, Simpson, I and **Wilson, A.** 2017. Educational backgrounds, project design, and inquiry learning in citizen science. In: Herodotou, C, Sharples, M and Scanlon, E (eds.) *Citizen inquiry: Synthesising science and inquiry learning*. Routledge. pp. 195–209. DOI: *https://doi.org/10.4324/9781315458618-11*

Embretson, SE. 1983. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1): 179–197. DOI: *https://doi.org/10.1037/0033-2909.93.1.179*

Embretson, SE and **Reise, SP.** 2000. *Item response theory for psychologists*. Mahwah, NJ: Erlbaum. DOI: *https://doi.org/10.1037/10519-153*

Fang, Z. 2006. The Language Demands of Science Reading in Middle School. *International Journal of Science Education*, 28(5): 491–520. DOI: *https://doi.org/10.1080/09500690500339092*

Fischer, F, Kollar, I, Ufer, S, Sodian, B, Hussmann, H and **Pekrun, R,** et al. 2014. Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. *Frontline Learning Research*, 2(3): 28–45.

Fischer, GH. 1995. The Linear Logistic Test Model. In: Fischer, GH and Molenaar, IW (eds.) *Rasch Models: Foundations, recent develeopments, and applications*. New York: Springer. pp. 131–155. DOI: *https://doi.org/10.1007/978-1-4612-4230-7_8*

Fischer, GH. 2005. Linear Logistic Test Models. In: Kempf-Leonard, K (ed.) *Encyclopedia of social measurement*. Amsterdam: Elsevier. pp. 505–514. DOI: *https://doi.org/10.1016/B0-12-369398-5/00453-9*

Flesch, R. 1948. A new readability yardstick. *The Journal of applied psychology*, 32(3): 221–233. DOI: *https://doi.org/10.1037/h0057532*

Follett, R and **Strezov, V.** 2015. An analysis of citizen science based research: Usage and publication patterns. *PloS one*, 10(11): e0143687. DOI: *https://doi.org/10.1371/journal.pone.0143687*

Hammann, M, Phan, TTH, Ehmer, M and **Grimm, T.** 2008. Assessing pupils' skills in experimentation. *Journal of Biological Education*, 42(2): 66–72. DOI: *https://doi.org/10.1080/00219266.2008.9656113*

Hartig, J and **Frey, A.** 2012. Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Construct validation and scale description in competence diagnostics by predicting task difficulties]. *Psychologische Rundschau*, 63(1): 43–49. DOI: *https://doi.org/10.1026/0033-3042/a000109*

Hartmann, S, Upmeier zu Belzen, A, Krüger, D and **Pant, HA.** 2015. Scientific reasoning in higher education. *Zeitschrift für Psychologie*, 223(1): 47–53. DOI: *https://doi.org/10.1027/2151-2604/a000199*

Jordan, RC, Ballard, HL and **Phillips, TB.** 2012. Key issues and new approaches for evaluating citizen-science learning outcomes. *Frontiers in Ecology and the Environment*, 10(6): 307–309. DOI: *https://doi.org/10.1890/110280*

Kauertz, A, Fischer, HE, Mayer, J, Sumfleth, E and **Walpuski, M.** 2010. Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I [Modeling competence according to standards for science education in secondary schools]. *Zeitschrift für Didaktik der Naturwissenschaften*, 16: 135–153.

Kind, P and **Osborne, J.** 2017. Styles of scientific reasoning: A cultural rationale for science education? *Science Education*, 101(1): 8–31. DOI: *https://doi.org/10.1002/sce.21251*

Klahr, D and **Dunbar, K.** 1988. Dual space search during scientific reasoning. *Cognitive Science*, 12(1): 1–48. DOI: *https://doi.org/10.1207/s15516709cog1201_1*

Krell, M. 2017. Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, 4(1): 79. DOI: *https://doi.org/10.1080/2331186X.2017.1280256*

Krell, M. 2018. Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie [Difficulty-generating task characteristics in multiple-choice tasks for experimental competence in biology teaching: A replication study]. *Zeitschrift für Didaktik der Naturwissenschaften*, 24(1):1-15. DOI: *https://doi.org/10.1007/s40573-017-0069-0*

Krell, M, Khan, S and **van Driel, J.** 2021. Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear

Logistic Test Model (LLTM). *Education Sciences*, 11(9): 472. DOI: *https://doi.org/10.3390/educsci11090472*

Kulgemeyer, C and **Starauschek, E.** 2014. Analyse der Verständlichkeit naturwissenschaftlicher Fachtexte. In: Krüger, D, et al. (eds.) *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin, Heidelberg: Springer. pp. 241–253. DOI: *https://doi.org/10.1007/978-3-642-37827-0_20*

Lawson, AE, Clark, B, Cramer-Meldrum, E, Falconer, KA, Sequist, JM and **Kwon, Y-J.** 2000. Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, 37(1): 81–101. DOI: *https://doi.org/10.1002/(SICI)1098-2736(200001)37:1<81::AID-TEA6>3.0.CO;2-I*

Le Hebel, F, Montpied, P, Tiberghien, A and **Fontanieu, V.** 2017. Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, 39(4): 468–487. DOI: *https://doi.org/10.1080/09500693.2017.1294784*

Lindner, MA, Ihme, JM, Saß, S and **Köller, O.** 2018. How Representational Pictures Enhance Students' Performance and Test-Taking Pleasure in Low-Stakes Assessment. *European Journal of Psychological Assessment*, 34(6): 376–385. DOI: *https://doi.org/10.1027/1015-5759/a000351*

Mair, P and **Hatzinger, R.** 2007. Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, 20(9): 1–20. DOI: *https://doi.org/10.18637/jss.v020.i09*

Mannel, S, Walpuski, M and **Sumfleth, E.** 2015. Erkenntnisgewinnung: Schülerkompetenzen zu Beginn der Jahrgangsstufe 5 im naturwissenschaftlichen Anfangsunterricht [Acquirement of Knowledge: Students' Competencies at the Beginning of Secondary Schooling in the Natural Sciences (Grade 5)]. *Zeitschrift für Didaktik der Naturwissenschaften*, 21(1): 99–110. DOI: *https://doi.org/10.1007/s40573-015-0028-6*

Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American psychologist*, 50(9): 741–749. DOI: *https://doi.org/10.1037/0003-066X.50.9.741*

Morris, BJ, Croker, S, Masnick, AM and **Zimmerman, C.** 2012. The Emergence of Scientific Reasoning. In: Kloos, H, Morris, BJ and Amaral, JL (eds.) *Current topics in children's learning and cognition*. Rijeka, Croatia: InTech. pp. 61–82.

National Academies of Sciences, Engineering, and Medicine (NASEM). 2018. *Learning through citizen science: Enhancing opportunities by design*. Washington, DC: National Academies Press (US).

Opfer, JE, Nehm, RH and **Ha, M.** 2012. Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching*, 49(6): 744–777. DOI: *https://doi.org/10.1002/tea.21028*

Opitz, A, Heene, M and **Fischer, F.** 2017. Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation*, 23(3-4): 78–101. DOI: *https://doi.org/10.1080/13803611.2017.1338586*

Phillips, T, Ferguson, M, Minarchek, M, Porticella, N and **Bonney, RE.** 2014. User's Guide for Evaluating Learning Outcomes in Citizen Science, Cornell Lab of Ornithology. Ithaca, NY.

Phillips, T, Porticella, N, Constas, M and **Bonney, RE.** 2018. A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science: Theory and Practice*, 3(2): Article 3. DOI: *https://doi.org/10.5334/cstp.126*

Phillips, TB, Ballard, HL, Lewenstein, BV and **Bonney, R.** 2019. Engagement in science through citizen science: Moving beyond data collection. *Science Education*, 103(3): 665–690. DOI: *https://doi.org/10.1002/sce.21501*

Stiller, J, Hartmann, S, Mathesius, S, Straube, P, Tiemann, R and **Nordmeier, V,** et al. 2016. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, 41(5): 721–732. DOI: *https://doi.org/10.1080/02602938.2016.1164830*

Stillfried, M, Gras, P, Börner, K, Göritz, F, Painer, J and **Röllig, K,** et al. 2017. Secrets of Success in a Landscape of Fear: Urban Wild Boar Adjust Risk Perception and Tolerate Disturbance. *Frontiers in Ecology and Evolution*, 5: 683. DOI: *https://doi.org/10.3389/fevo.2017.00157*

Straka, TM, Wolf, M, Gras, P, Buchholz, S and **Voigt, CC.** 2019. Tree Cover Mediates the Effect of Artificial Light on Urban Bats. *Frontiers in Ecology and Evolution*, 7: 27. DOI: *https://doi.org/10.3389/fevo.2019.00091*

Stylinski, CD, Peterman, K, Phillips, T, Linhart, J and **Becker-Klein, R.** 2020. Assessing science inquiry skills of citizen science volunteers: A snapshot of the field. *International Journal of Science Education, Part B*, 10(1): 77–92. DOI: *https://doi.org/10.1080/21548455.2020.1719288*

Trumbull, DJ, Bonney, RE, Bascom, D and **Cabral, A.** 2000. Thinking scientifically during participation in a citizen-science project. *Science Education*, 84(2): 265. DOI: *https://doi.org/10.1002/(SICI)1098-237X(200003)84:2<265::AID-SCE7>3.0.CO;2-5*

Wilson, MR, Boeck, P de and **Carstensen, CH.** 2008. Explanatory Item Response Models: A Brief Introduction. In: Hartig, J, Klieme, E and Leutner, D (eds.) *Assessment of Competencies in Educational Contexts*, 1st edn. Göttingen: Hogrefe Publishing. pp. 83–110.

Wright, B and **Linacre, J.** 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8: 370.

Wu, ML, Adams, RJ and **Wilson, MR.** 2007. *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell, Vic.: ACER Press.

TO CITE THIS ARTICLE:
Bruckermann, T, Straka, TM, Stillfried, M and Krell, M. 2021. Context Matters: Accounting for Item Features in the Assessment of Citizen Scientists' Scientific Reasoning Skills. *Citizen Science: Theory and Practice*, 6(1): 21, pp. 1–15. DOI: *https://doi.org/10.5334/cstp.309*

Submitted: 28 January 2020 Accepted: 20 October 2021 Published: 25 November 2021

Citizen Science: Theory and Practice is a peer-reviewed open access journal published by Ubiquity Press.