Original Article

Using Differential Item Functioning to Analyze the Domain Generality of a Common Scientific Reasoning Test

Ansgar Opitz, Moritz Heene, and Frank Fischer

Department of Psychology, LMU Munich, Germany

Abstract: A significant problem that assessments of scientific reasoning face at the level of higher education is the question of domain generality, that is, whether a test will produce biased results for students from different domains. This study applied three recently developed methods of analyzing differential item functioning (DIF) to evaluate the domain generality assumption of a common scientific reasoning test. Additionally, we evaluated the usefulness of these new, tree- and lasso-based, methods to analyze DIF and compared them with methods based on classical test theory. We gave the scientific reasoning test to 507 university students majoring in physics, biology, or medicine. All three DIF analysis methods indicated a domain bias present in about one-third of the items, mostly benefiting biology students. We did not find this bias by using methods based on classical test theory. Those methods indicated instead that all items were easier for physics students compared to biology students. Thus, the tree- and lasso-based methods provide a clear added value to test evaluation. Taken together, our analyses indicate that the scientific reasoning test is neither entirely domain-general, nor entirely domain-specific. We advise against using it in high-stakes situations involving domain comparisons.

Keywords: scientific reasoning, domain generality, Differential Item Functioning, assessment, higher education

OPEN DATA

The assessment of scientific reasoning has been highlighted as a particularly important challenge of this century (Osborne, 2013). Not all conceptualizations of scientific reasoning include the same skills, but there is considerable overlap and at its core the set comprises skills like formulating questions, formulating hypotheses, gathering evidence, evaluating evidence, explaining results, and communicating results (Fischer et al., 2014; National Research Council [NRC], 2012). These skills are not only relevant in conducting scientific studies but also in professional practice. While there is no shortage of scientific reasoning tests in the literature, a lot of them are of unknown quality and many assumptions remain untested (Opitz et al., 2017). This article will focus on one aspect of the assessment of scientific reasoning: the domain generality of scientific reasoning tests when applied to higher education students.

Whether scientific reasoning is domain-general or domain-specific has been an issue for decades. The position of domain generality entails that the core set of scientific reasoning skills is very similar or the same in domains like chemistry or physics. It also implies that the domain context of a test has no or almost no influence on the assessment of scientific reasoning skills. Proponents of this position claim that it is possible to negate the potential influence of test elements specific to a certain domain if content is used that students are familiar with and thus prevent that scientific reasoning tests are dominated by content effects (Harlen, 1999). Early on, test authors, inspired by the works of Jean Piaget (e.g., Inhelder & Piaget, 1958), tended toward such domain-general conceptualizations. The popularity of this position declined over time, though, as doubts arose about whether a universally applicable scientific method exists (Kind & Osborne, 2017). Thus, declaring scientific reasoning to be domain-specific seemed to be the logical conclusion. The position of domain specificity proposes a very close connection between scientific reasoning skills and the domain that they are tested in. According to this position, one cannot infer from the results of a scientific reasoning test in one context to the results in another context. The proponents of this position think that there should be no separation between reasoning and

https://doi.org/10.1027/1015-5759/a000662

knowledge and that real-life scenarios will always involve domain knowledge and thus no useful insights can be drawn from tasks that are knowledge-lean. (Kind, 2013; Osborne, 2013; Zimmerman, 2000). Additionally, some conceptualizations of scientific reasoning have moved away from a strict general versus specific dichotomy (Hetmanek et al., 2018; Karmiloff-Smith, 2012; Niaz, 1995). They postulate that scientific reasoning skills apply to more than just one specific context but they are also not as general as intelligence and some scientific reasoning skills can be more general than others.

We chose the area of higher education because knowing the degree of generality of scientific reasoning skills is especially relevant in order to find out how successful universities are in teaching scientific reasoning skills independent of a particular major. Considering the general relevance of scientific reasoning skills in both academic and nonacademic work environments, it would be ideal that students acquire these skills independent of their major, especially within the sciences. Assessing whether the acquisition of scientific reasoning skills independent of a major is indeed happening, requires an evaluation of whether a test can be given to students from different majors without producing biased results.

One test that makes such a claim about its domain generality is the Classroom Test of Scientific Reasoning (CTSR; Lawson, 2000). As it is one of the few common scientific reasoning tests that have been used in multiple studies, we selected it for this study for an evaluation of its domain generality assumption. The test has been used in higher education settings with study participants majoring in science as well as study participants not majoring in science (Bao et al., 2009; Coletta & Phillips, 2005; Lawson, Alkhoury, et al., 2000; Lawson, Clark, et al., 2000). Other tests that aim to measure scientific reasoning on a higher education level use items that are similar to CTSR items (see e.g., Gormally et al., 2012; Tobin & Capie, 1981), so it is reasonably representative of scientific reasoning assessments. Other authors who evaluated the CTSR described the relevance of knowledge that is domain-specific and required for the test as “minimal” (Osborne, 2013, p. 269).

Unfortunately, the domain generality assumptions of scientific reasoning tests, including the CTSR, are rarely tested (Opitz et al., 2017). The few studies that do test domain-related assumptions are problematic: They use methods from Classical Test Theory (CTT) that falsely equate observed test scores with underlying latent abilities (Cloonan & Hutchinson, 2011; Weld et al., 2011). In order to better understand why this is problematic, we can use the study by Weld et al. (2011) as an example. The authors compared test results from elementary education and biology majors and used the absence of a significant difference as support for their assumption of domain generality. However, this line of reasoning allows for a scenario where the results for the two groups of students are only on the same level because one group benefited from an unfair advantage. This advantage might have overshadowed the reality that the actual latent ability of one group is below the latent ability of the other group. This approach also allows for another problematic scenario that fails to separate between two potential realities: If the students majoring in biology would have achieved a significantly better result, there are two possibilities. First, it could mean that their latent scientific reasoning ability is higher. This would still be in accordance with an assumption of domain generality. Second, it could mean that students majoring in biology benefited from an unfair advantage due to biased elements in the test. This would contradict the assumption of domain generality. These two possibilities cannot be separated from one another by just looking at the difference of mean scores between two groups.

Another insufficiency of the previously used methods is that they only focus on the bias on the level of the total score but the absence of bias on this level is not necessarily an indicator for the absence of bias on the level of individual items (Borsboom, 2006). This means, for instance, that a test could contain biased items with a biology context and biased items with a physics context but a comparison of means between physics and biology students might not indicate any bias because the biased items are canceling each other out on the level of the total score.

This article will address these shortcomings of previously used CTT methods by employing analysis techniques that were developed within the framework of item-response theory (IRT). The IRT concept that is most relevant for our study is called differential item functioning (DIF). We will use DIF to judge whether the measurement properties of a test are invariant across different groups of interest. If the analyses point to an absence of measurement invariance, we have to assume that the results of an assessment were influenced by characteristics of group membership unrelated to the trait being measured. This would imply that it is not possible to compare group means without bias. We propose as a new idea for the present study that DIF analyses as an indicator of measurement invariance can be applied to the issue of domain generality. This idea is based on the following assumption: If DIF analyses indicate that the measurement properties of an assessment vary between students who are majoring in different domains, we should not expect that this test can assess scientific reasoning in a way that would be considered domain-general. While previous techniques to analyze DIF were limited to comparisons of two groups, we will employ more recent methods that are able to get rid of this limitation. We will employ these recent techniques to see whether a bias is found and if so which items are responsible for the bias.

Using more than one method allows us to see how stable the results of IRT-based bias analyses are when different methods are used on the same dataset.

Specifically, we selected so-called tree models as one of our methods, which are also known under their technical term as model-based recursive partitioning (Strobl et al., 2009). We employed two tree-model techniques: We used a Rasch tree, to analyze DIF on a global test level, that is, without looking at specific items (Strobl et al., 2015). The evaluation of DIF at the level of specific items was done with a second technique, namely item-focused trees (Tutz & Berger, 2016). Both tree-based techniques are following the same general procedure. The first step is to estimate all parameters jointly for the whole sample. The second step is to check how stable these parameter estimates are when the covariates are considered that might cause DIF. For instance, we will check in this article how stable parameter estimates are for different student groups. If DIF is indeed present, the introduction of the covariate leads to a systematic deviation in parameter estimates and not just a random fluctuation and thus the sample should be split into subsamples with different estimates. To test this, one transforms the deviation into a test statistic that can be submitted to a significance test. As this is done simultaneously for all potential splits at the same time the α-level of the significance test is adjusted for multiple testing. This adjustment is very important in order to control the false alarm rate, that is, without it we would find DIF in places where it does not exist. If at least one covariate leads to a significant deviation, the sample is split in a third step. If multiple splits would be significant the one that improves the model fit the most is chosen. These three steps are repeated within the subsample branches that are produced by the split until no more significant deviations remain (or the sub-sample size gets below a predetermined threshold). Thus, a tree emerges. As pointed out above, the technique by Strobl et al. (2015) applies these steps on a global level while the technique by Tutz and Berger (2016) applies it on the item level. The advantage compared to prior methods that checked DIF on an item level is that all items are considered simultaneously and not independently of another. The latter approach assumes for every item test that all other items are free of DIF, which is often unrealistic. Items that are selected for at least one split are considered to exhibit DIF, while items that are not selected at all do not exhibit DIF.

As our second method, we selected a technique that is known as the least absolute shrinkage and selection operator (lasso; Tibshirani 1996). Specifically, we chose a technique devised by Tutz and Schauberger (2015). This lasso-based technique introduces many parameters for potential DIF and then reduces them in a way that only genuine DIF is left in the model. In the first step, DIF parameters are introduced for every covariate in every item. The eventual goal is that every parameter unequal from zero will indicate DIF. The problem is that many of these parameters will vary from zero just by chance and we need to separate those from the parameters that indicate genuine DIF. This problem is solved by the lasso technique by introducing a penalization to the parameter estimation. The penalization is introduced in the form of the tuning parameter λ that shrinks the DIF parameter estimates. If λ is set to zero, we would get the standard estimate resulting in a high rate of false hits indicating DIF where there is none. When λ approaches infinity, all of these extra DIF parameters would shrink to zero and no item would be considered to exhibit DIF. However, these are extreme cases only used for explanation purposes, and for typical λ values most but not all parameters will be reduced to zero. If λ is carefully set then all the parameters which are not zero indicate genuine DIF. To find the optimal λ value the lasso uses the Bayesian information criterion (BIC; Schwarz, 1978), a criterion that balances model complexity with the model fit.

To summarize, we employed and compared the following methods to look for domain bias: On the level of the whole test, we compared the previously used CTT method of comparing means with a Rasch tree DIF analysis by Strobl et al. (2015). On the item level, CTT only allows for looking at item difficulties. This was compared with tree- and lasso-based IRT techniques (Tutz & Berger, 2016; Tutz & Schauberger, 2015) to analyze DIF.

We selected these techniques because it was shown that they can detect DIF because they rarely produce false alarms, that is, they rarely indicate DIF where it is not present, and because the models they produce are easy to interpret (Strobl et al., 2015; Tutz & Berger, 2016; Tutz & Schauberger, 2015). Additionally, the analyses on the item level provide an advantage if we discover DIF: In that case, we can gain insights about the connection of the domain specificity of items with their respective DIF results, that is, we can see if items that experts consider to be more domain-specific have a higher rate of DIF. We do want to make a note, though, about the lasso method: As all parameters get shrunk in its calculation, it underestimates the bias' exact size (Tutz & Schauberger, 2015).

Using the aforementioned methods, we wanted to address the following two research questions:

> *Research Question 1 (RQ 1)*: Can the CTSR be considered domain-general when used in higher education? A high amount of bias between domains and according to performance advantages would speak against this.

> *Research Question 2 (RQ 2)*: How useful are the employed tree- and lasso-based DIF analyses compared to previously used methods to analyze domain

generality assumptions? Especially the following two aspects would speak in favor of the new analyses being useful: First, do they effectively address the insufficiencies of previously used methods of classical test theory when analyzing domain bias? Second, do they provide conceptual and practical insights on an item level that go beyond just analyzing bias on the level of the complete test?

Method

Sample

We started data collection with the goal of including 500 participants. This number was informed by studies about the methods we wanted to use in which simulations had revealed an acceptable relation of true and false-positive indications of DIF for that sample size (Strobl et al., 2015; Tutz & Berger, 2016; Tutz & Schauberger, 2015). Our final sample was 507 students studying at university (249 male students, 256 female students). The students were M = 23.01 years (SD = 2.91) and had already been at university for M = 6.76 semester (SD = 2.53). The students were majoring in physics (192 participants), biology (167 participants), or medicine (148 participants). We chose biology and physics students for our sample because there are items in the CTSR with contexts from these two domains. Additionally, we included students majoring in medicine, because we were interested to see whether we could find a domain bias between participants enrolled in a science major – the combination of biology and physics majors in our case – and participants studying a related discipline.

In terms of exclusion criteria, we removed students who stopped taking the test before they completed the full assessment. This was the case for 27 participants. We also excluded students if their time for completing the test was below 40% of the full testing time (which was found to be the minimum amount of time to finish the assessment in a pilot study if one actually worked on the questions) and the rate of correct answers was also below the chance level at the same time. If these two criteria were met, we thought it safe to assume that the according student had randomly answered the test questions. We made one exclusion based on this procedure. The inclusion and exclusion criteria were established prior to data analysis.

Test Instruments

We translated the questions of the CTSR into German. The phrasing was checked in a pilot phase of the study and we observed no issues with it. The majority of the 24 CTSR questions are paired: First, the participants choose what they think is the correct answer. Second, the students choose what they think is the according to justification for this answer. The assessment is then scored based on a suggestion of the authors of the CTSR: The paired questions are merged into a single item for each pair. The only exception to this procedure is questions 23 and 24. They form items of their own. The total possible score resulting from this procedure is 13.

In order to check the degree that the solution to biased items depends on domain-specific aspects, we gave the items to two researchers from the department of physics and two researchers from the department of biology. They provided us with a context rating for the items in terms of their domain dependency. We had created a rating scheme for this task that had four possible classifications: Items with no biology or physics context were rated with a zero. A rating of 1 was given if an item had a biology or physics context but the raters did not expect that context to induce an advantage for a specific domain. An item received a rating of 2 if the raters did expect an advantage for students from a specific domain when solving the item, but the raters also thought it possible that a scientific reasoning skill that might be useful across domains, such as the interpretation of data, could also be applied in solving the item. Last, the raters applied a rating of 3 if they thought it necessary for test-takers to achieve mastery of a domain-specific concept to find the correct solution for an item. Table 1 contains a list of item numbers, information about which questions were combined to form the item, abbreviated names for the items, a simplified explanation of the contents of the questions (which is different from their exact wording), as well as the rating given to the item context by the experts. The context rating was established via consensus ratings of the experts.

Additionally, in order to control whether possible domain differences are based on differences in general reasoning, we chose three subscales from the Intelligence Structure Test, revised version (IST 2000 R; Amthauer et al., 2001), selected 10 items from each subscale, and added them to the test booklet. These subscales assessed numerical, verbal, and figural reasoning.

Analysis

The DIF analyses that were used do not return parameter estimates in situations in which every student (or all students from one domain if domains were compared) had correctly solved an item and for students who correctly answered all questions. Based on this, we had to exclude one item from the DIF calculations because every student majoring in physics had answered it correctly (Item 1 in Table 1). Additionally, we had to remove the students who had achieved the maximum score of 13 in the CTSR from

Table 1. Overview of scientific reasoning test items and their context rating

Item no	Question	Item	Item description	Context rating[a]
1	1 + 2	Clay balls	What happens to the weight of a clay ball when the ball is flattened?	3 (Phy)
2	3 + 4	Marbles	When a steel marble is put into a cylinder filled with water, what happens to the water level compared to a glass marble?	3 (Phy)
3	5 + 6	Water tubes 1	How high will water rise in a narrow cylinder when water from a wide cylinder is poured into it?	0
4	7 + 8	Water tubes 2	How high will water rise in a wide cylinder when water from a narrow cylinder is poured into it (the amount of water is different from the item before)?	0
5	9 + 10	Strings	Which strings (out of 3 possible strings) have to be used to find out if the length of a string has an effect on the time of one swing of the string?	1 (Phy)
6	11 + 12	Flies 1	What does a figure (showing the results of an experiment with flies) tell you about two possible influences on the behavior of flies?	1 (Bio)
7	13 + 14	Flies 2	This item is the same as item 6 with one influence factor and the results being changed.	1 (Bio)
8	15 + 16	Urn 1	What are the chances of drawing a certain kind of wooden piece out of a given set of pieces?	0
9	17 + 18	Urn 2	This item is the same as item 8 with the target piece and the set of pieces being changed.	0
10	19 + 20	Mice	Based on a figure about mice that carry a combination of two possible traits: Is there a link between the two traits?	2 (Bio)
11	21 + 22	Candle	How can a suggested explanation for a given observation be tested and which result would show that the explanation is wrong?	1 (Phy)
12	23	Blood cells 1	Which result of an experiment would show that the explanation for an observation is wrong?	1 (Bio)
13	24	Blood cells 2	Which result of an experiment would show that the explanation for an observation is wrong (the explanation is different from the item before)?	1 (Bio)

Note.· [a]Context rating: 0 = no item context from physics or biology; 1 = item has physics or biology context but no domain-specific aspects that help solve the item; 2 = domain-specific aspects help solve the item, but the item can also be solved by a cross-domain valid skill; 3 = mastery of domain-specific aspects are necessary for solving the item. Phy = physics; Bio = biology.

our DIF calculations. After this removal, we ended up with n = 461 students who could be included in the DIF calculations. We created two dummy variables to represent the three domains of our participants in the DIF calculations. Domain differences were additionally analyzed with CTT-based methods that were used to establish the domain generality status of tests in the past. These methods included an analysis of variance (ANOVA), an analysis of covariance (ANCOVA), and the calculation of item easiness. Numerical, verbal, and figural reasoning served as control variables in the ANCOVA. While we found ceiling effects (see the Results section for more details) that violate the normality assumption of the ANOVA and ANCOVA, the assumption of homogenous slopes was met for the ANCOVA.

We calculated descriptive analyses as well as the CTT-based scale characteristics and mean comparisons (i.e., ANOVA and ANCOVA) with IBM® SPSS® 23. We conducted the analysis of DIF with R, version 3.2.5, using the DIFlasso (Schauberger, 2016), DIFtree (Berger, 2016), and psychotree (Strobl et al., 2015) packages.

Results

Before we present the results of the DIF analyses, several important descriptive statistics are given. The mean total score in the CTSR for the whole sample was M = 9.91 (SD = 2.22), with a maximum total score of 13. The reliability (Cronbach's α) of the CTSR was .63 and for figural, verbal, and numerical reasoning it was .67, .47, and .77, respectively. Table 2 displays the easiness of the CTSR items according to CTT. The table contains the values for the complete sample and the three subsamples grouped by major. It is noteworthy that all values for biology majors are below the value for physics majors (i.e., the items were more difficult for biology students in CTT terms).

Differential Item Functioning Analyses

The first DIF analysis was conducted on the level of the complete test using the tree-based technique described in the introduction. The analysis suggested two splits: The first split was made between physics students and all remaining students, p < .001. The second split was made within the group of non-physics students for biology versus medicine, p = .014. Based on these splits we would expect that comparing any of the three subgroups of students will produce a biased result.

As the previous analysis only covered the level of the whole test, we need to turn to the analyses using the item-focused tree as well as the lasso to single out the items that violate the assumption of measurement invariance. The item difficulties (in the form of logit values) produced by these analyses are shown in Table 3. When looking at

Table 2. Item easiness of the scientific reasoning items according to CTT

		Item easiness[a]			
Item no	Item	Complete sample	Physics	Biology	Medicine
1	Clay balls	.98	1.00	.95	.99
2	Marbles	.88	.94	.85	.83
3	Water tubes 1	.76	.87	.69	.70
4	Water tubes 2	.53	.64	.43	.49
5	Strings	.89	.93	.81	.92
6	Flies 1	.44	.47	.46	.39
7	Flies 2	.65	.71	.62	.59
8	Urn 1	.90	.95	.83	.91
9	Urn 2	.82	.89	.76	.79
10	Mice	.70	.72	.62	.77
11	Candle	.73	.86	.63	.66
12	Blood cells 1	.78	.84	.67	.82
13	Blood cells 2	.86	.86	.79	.93
	Mean	.76	.82	.70	.75

Note. [a]Item easiness: 1 = item difficulty; calculations based on CTT

Table 3. Item difficulty parameter from the tree- and lasso-based analyses

		Tree-based analysis			Lasso-based analysis			
Item no	Item	Diff_{S}	Diff_{Phy}	Diff_{Bio}	Diff_{S}	Diff_{Phy}	Diff_{Bio}	Context rating
1	Clay balls	na	na	na	na	na	na	3 (Phy)
2	Marbles	−1.15			−0.59			3 (Phy)
3	Water tubes 1	−0.13			0.42			0
4	Water tubes 2	1.29			1.83			0
5	Strings	−1.27			−0.70			1 (Phy)
6	Flies 1	2.18		1.03	2.31	2.41	2.07	1 (Bio)
7	Flies 2	0.87		0.18	1.17	1.19	1.10	1 (Bio)
8	Urn 1	−1.39			−0.82			0
9	Urn 2	−0.56			0.00			0
10	Mice	−0.04	0.88		0.84	1.02	0.87	2 (Bio)
11	Candle	0.12			0.67	0.62	0.67	1 (Phy)
12	Blood cells 1	−0.23			0.32			1 (Bio)
13	Blood cells 2	−1.99	−0.23	−0.89	−0.36	−0.22	−0.30	1 (Bio)

Note. Diff_{S} = baseline difficulty of the sample; Diff_{Phy} = difficulty for physics students; Diff_{Bio} = difficulty for biology students. If a cell is empty the difficulty for that group is the same as for the whole sample. In case the difficulty diverges for both physics and biology students, Diff_{S} becomes the difficulty for medical students. If both the Diff_{Phy} and Diff_{Bio} cells are empty the according item is unbiased as Diff_{S} denotes the difficulty for the whole sample in this case. Item 1 was left out of the analyses because 100% of physics students solved the item. The displayed values are logit values. Smaller numbers indicate a lower difficulty. The logit values between the two analyses cannot be directly compared because of different parameter identification procedures and the shrinkage that is happening in the lasso analysis.

the numbers in the Table 3, it is important to remember the note from the introduction: The lasso analysis underestimates the absolute value of the difference between student groups. The emphasis of the lasso analysis is more about accurate and important information regarding which items are biased in which direction and less about estimating the bias size.

The two analyses consistently pointed to domain bias in four items (6, 7, 10, and 13). One additional biased item (Item 11) was present in the lasso results. A comparison between participants from biology and physics regarding these biased items shows consistent matches between the bias and the context domain (which can be seen in the last columns of Tables 1 and 3). In the biased items with a biology context (6, 7, 10, and 13) the bias favored participants from biology. The bias in Item 11, which is embedded into a physics context, favors participants from physics. At the same time, two other items (2 and 5) that were embedded into a physics context seem to be unbiased based on the results. However, these items were not especially difficult.

When compared with all other items in the DIF calculations, there was just one other item (Item 8) that had a lower difficulty.

Domain Differences in ANOVA and ANCOVA Analyses

A one-way ANOVA comparing the scientific reasoning scores of the student subgroups revealed a significant difference, $F(2, 503) = 24.82$, $p < .001$, $\eta_p^2 = .09$. We used a Bonferroni corrected post hoc analysis to compare the three subgroups and found that biology majors achieved significantly lower scores than medical students, who achieved significantly lower scores than physics students. In a subsequent ANCOVA, the domain effect remained significant after we controlled for numerical, verbal, and figural reasoning, $F(2, 473) = 11.94$, $p < .001$. The effect size was reduced to a $\eta_p^2 = .05$.

Discussion

With the help of DIF analyses, we discovered a domain bias in multiple CTSR items between two domains of science, physics and biology. Thus, on the one hand, the answer to our first research question, whether the CTSR can be considered domain-general, seems to be that the evidence produced by this study is weak. The context rating conducted by domain experts for this study shows two things: First, the biased items are embedded in a context that matches the bias, for example, the items that were biased toward biology majors are embedded into a biological context. Second, that it is possible, in theory, to solve these items without domain-specific knowledge. It seems that bias can occur in an item just because some of its context features are domain-specific, even if no domain-specific knowledge is absolutely necessary for providing the correct solution. On the other hand, the CTSR does not seem to be a completely domain-specific test either. When we look at the items that favored biology majors, we do not observe a high fail rate for physics majors. In fact, one could argue that the opposite is more accurate: In terms of CTT, the biased items were actually more difficult for biology majors when compared to physics majors. It seems as if the physics students managed to use their scientific reasoning skills to solve these items. The same picture emerges on the level of the complete test: The physics students were significantly better than the biology students, even when controlling for general reasoning.

Based on these observations, we advise being cautious if authors of scientific reasoning assessments claim that their assessment is measuring scientific reasoning in a fully domain-specific or domain-general fashion. This has implications for the debate surrounding the domain generality or specificity of scientific reasoning. Based on the presented results it seems reasonable for researchers to explore alternative conceptualizations that go beyond the classical dichotomy (Hetmanek et al., 2018; Karmiloff-Smith, 2012; Niaz, 1995; Zimmerman, 2000) and instead imagine scientific reasoning as a set of skills that are relevant in some but not all contexts. The way forward then is to test the limits of such conceptualizations: For instance, looking at the item content of the biased items in this study, we might hypothesize that items involving the interpretation of experimental designs are more specific.

The results also have implications for the evaluation and construction of scientific reasoning tests. Based on this study we can only speculate about the cause of the domain bias, for example, whether motivational factors are at work. To explore this in more detail we might have to analyze the way the items are solved with the help of think-aloud interviews. A study by Adams and Wieman (2015) using problem-solving tasks did exactly that and might serve as an inspiration for such an approach.

As it is very common for scientific reasoning items to be embedded in a domain context (see e.g., Gormally et al., 2012; Schwichow et al., 2016) test creators should pay particular attention to the potential bias this induces, especially in regard to the interpretation of test results. It might be tempting to simply scratch all biased items but this could leave out important aspects of scientific reasoning so this should be considered very carefully. As an alternative for long assessments with many questions, it might be feasible to produce domain-specific difficulty estimates for items exhibiting DIF by using the items without bias as a fixed comparison (Boone et al., 2014).

Besides these theoretical and psychometrical implications, we also want to consider the practical implications of our results. First, it should be noted that an absence of measurement invariance does not necessarily imply a simultaneous absence of predictive invariance (Millsap, 1995). Whether the presence of bias means that the bias will be a concern in the application of the test, depends on the purpose the test is used for (Borsboom, 2006). Considering that predictive invariance might be present if the focus of an application of the CTSR is on the relationship with other variables, the test might still produce unbiased results for that specific purpose. However, the bias we found should not be neglected. A good way to look at this that is more convenient to interpret is the transformation of logit values into the probabilities that a hypothetical person would solve the item with and without the bias: The largest difference in item difficulty was 1.76 logits between medicine and physics in Item 13. A bias of that size means that the probability to solve such an item for a person who has a 50% chance of solving the item without the bias, would jump to 85% if the person benefits from the bias. Therefore, we would advise against using the CTSR, or similar scientific reasoning tests, in situations that involve comparisons of students who were not previously enrolled in the same major, such as the selection of students for PhD positions, as the results would be of doubtful validity. Another area that might be affected by biased test results is the evaluation of educational programs. The outcomes of the PISA assessment (Organisation for Economic Co-operation and Development [OECD], 2007) had a substantial influence on educational policies in secondary education, and similar assessments are in demand for higher education (Zlatkin-Troitschanskaia et al., 2015). If these assessments want to measure scientific reasoning for a wide range of students with supposedly domain-general items, it is important to consider the domain bias that is introduced by item context alone in order to not make decisions based on biased results.

The second goal of this study was to find out how useful IRT-based, and in particular DIF-based, methods are to evaluate the domain generality assumption of a scientific reasoning test. This is in line with others who have pointed toward the importance of IRT-based methods in the assessment of scientific reasoning (Edelsbrunner & Dablander, 2019) and the analyses in this study clearly show the added value of the applied DIF techniques. They helped us to reveal biased items, mostly in favor of biology students, and we would have not flagged these as biased toward biology majors if we had only considered CTT-based difficulty values. The added value of the DIF techniques becomes even more apparent if we contrast them with simple comparisons of mean values. These mean comparisons are one way that has been previously used by test authors to make claims about their assumptions regarding domain specificity or generality (Cloonan & Hutchinson, 2011; Weld et al., 2011). If we would have used only this previous method we might have come to a different bias evaluation compared to the bias analyses based on DIF, namely that the bias is directed against biology students. In contrast, the results from the DIF calculations imply that, on a latent level, our biology majors had lower scientific reasoning skills, and any differences in comparisons of means would actually be more substantial without the bias that favors biology majors.

Among the strengths of the applied analyses was the detailed information provided on an item level, which was one benefit we aimed to achieve by using more sophisticated analyses. It was this item-level information that allowed us to understand the role of domain specificity of item contexts. This detailed item-level information can help identify problematic items during test development and evaluation. The high convergence of two techniques using different calculations indicates the reliability of the findings gained from them. Based on these strengths we recommend continuing to apply the methods developed by Tutz and Schauberger (2015), Tutz and Berger (2016), and Strobl et al. (2015).

In terms of limitations of the present study, we need to mention that the CTSR scores of our participants were on the higher end of possible values. We are not the first to encounter this problem with the CTSR. In a study by Bao et al. (2009) participants from a higher education setting achieved a similar mean score at around 75% of the maximum score. Studies in higher education settings likely ask for scientific reasoning items with a more advanced difficulty. Additionally, we want to mention that our reliability value for the CTSR is below the value of .81 that at least one other study achieved (Lawson, Alkhoury, et al., 2000). We want to point out, though, that consistency of .63 is also not without precedent. Lawson, Clark, et al. (2000) recorded a comparable reliability value of .65 in a group of 667 college students. As we cannot rule out that the low difficulty and reliability affected the results, it is worthwhile to consider how this effect might look like. Based on what is known in general about the influence of ceiling effects and low reliability, it seems most reasonable to assume that they lead to a reduction of systematic variance, therefore making it harder to detect differences between groups (Charter, 2003; Šimkovic & Träuble, 2019). Thus, it seems reasonable to assume that the bias we found exists, especially as it was found with two different techniques, but it might be an underestimation of the overall bias. In particular, we consider it possible that items with a strong physics context, for example, Item 2, might have exhibited bias if only its overall difficulty would have been higher. It should be stressed that this assumption holds only if the measurement error is purely random. Over- or underestimations of group differences can, on the other hand, occur under systematic or differential measurement error (van Smeden et al., 2020).

Last, it could be said that our conclusions are tied to one particular test. In response, we would ask to consider how commonly used the CTSR is and its similarities to other scientific reasoning tests, which commonly cover skills such as generating and evaluating evidence and drawing conclusions, too (Opitz et al., 2017). Thus, we are confident that this study has consequences beyond just one test and that our conclusions are valid for scientific reasoning assessments in general.

Conclusion

In summary, based on our findings we advise against using the CTSR in high-stake situations that involve domain comparisons. Furthermore, we demonstrated that at a higher education level DIF offers insights about domain-induced bias that go beyond the insights offered by CTT. DIF methods offer more information not only on tests as a whole but also on specific items. We think this line of research deserves to be continued.

References

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics, 83*(5), 459–467. https://doi.org/10.1119/1.c4913923

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R* [Intelligence-Structure-Test 2000 R]. Hogrefe.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., Liu, Q., Ding, L., Cui, L., & Luo, Y. (2009). Learning and scientific reasoning. *Science, 323*(5914), 586–587. http://www.physics.ohio-state.edu/∼lbao/Papers/Bao_Learning-Scientific-Reasoning.pdf

Berger, M. (2016). *DIFtree: Item focused trees for the identification of items in differential item functioning* (Version 2.0.4) [R package]. https://CRAN.R-project.org/package=DIFtree

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Borsboom, D. (2006). When does measurement invariance matter? *Medical care, 44*(11), 176–181. http://journals.lww.com/lww-medicalcare/Abstract/2006/11001/When_Does_Measurement_Invariance_Matter_.23.aspx

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290–304. https://doi.org/10.1080/00221300309601160

Cloonan, C. A., & Hutchinson, J. S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice, 12*(2), 205–209. https://doi.org/10.1039/c1rp90025k

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics, 73*(12), 1172–1182. https://doi.org/10.1119/1.2117109

Edelsbrunner, P. A., & Dablander, F. (2019). The Psychometric Modeling of Scientific Reasoning: A Review and Recommendations for Future Avenues. *Educational Psychology Review, 31*(1), 1–34. https://doi.org/10.1007/s10648-018-9455-5

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45. https://doi.org/10.14786/flr.v2i2.96

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *Cell Biology Education, 11*(4), 364–377. https://doi.org/10.1187/cbe.12-03-0026

Harlen, W. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice, 6*(1), 129–144. https://doi.org/10.1080/09695949993044

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: Scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge* (pp. 203–226). Taylor & Francis.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structure*. Routledge & Kegan Paul.

Karmiloff-Smith, A. (2012). Is development domain specific or domain general? A third alternative. In J. Shrager & S. Carver (Eds.), *The journey from child to scientist: Integrating cognitive development and the education sciences* (pp. 127–140). American Psychological Association. https://doi.org/10.1037/13617-006

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning: Assessment scales with scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530–560. https://doi.org/10.1002/tea.21086

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

Lawson, A. E. (2000). *Classroom test of scientific reasoning* (rev ed.). http://modeling.asu.edu/modeling/weblinks.html

Lawson, A. E., Alkhoury, S., Benford, R., Clark, B. R., & Falconer, K. A. (2000). What kinds of scientific concepts exist. Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching, 37*(9), 996–1018. https://doi.org/10.1002/1098-2736(200011)37:9%3C996::AID-TEA8%3E3.0.CO;2-J

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., & 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. http://onlinelibrary.wiley.com/doi/10.1002/tea.20172/abstract

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577–605. https://doi.org/10.1207/s15327906mbr3004_6

National Research Council [NRC]. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://doi.org/10.17226/13165

Niaz, M. (1995). Enhancing thinking skills: Domain specific/domain general strategies – a dilemma for science education. *Instructional Science, 22*(6), 413–422. https://doi.org/10.1007/BF00897976

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

Opitz, A., Heene, M., & Fischer, F. (2021). *Using Differential Item Functioning to analyze the domain generality of a common scientific reasoning test* [Dataset]. https://doi.org/10.17605/OSF.IO/B9TP4

Organisation for Economic Co-operation and Development. (2007). *Science competencies for tomorrow's world. Volume I: Analysis*. OECD. https://doi.org/10.1787/9789264040014-en

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Schauberger, G. (2016). *DIFlasso: A penalty approach to differential item functioning in Rasch models* (Version 1.0-2) [R package]. https://CRAN.R-project.org/package=DIFlasso

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education, 38*(2), 216–237. https://doi.org/10.1080/09500693.2015.1137651

Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLoS One, 14*(8), Article e0220889. https://doi.org/10.1371/journal.pone.0220889

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348. https://doi.org/10.1037/a0016973

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society; Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tobin, K. G., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement, 41*(2), 413–423. https://doi.org/10.1177/001316448104100220

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika, 81*(3), 727–750. https://doi.org/10.1007/s11336-015-9488-3

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Van Smeden, M., Lash, T. L., & Groenwold, R. H. H. (2020). Reflection on modern methods: Five myths about measurement error in epidemiological research. *International Journal of Epidemiology, 49*(1), 338–347. https://doi.org/10.1093/ije/dyz251

Weld, J., Stier, M., & McNew-Birren, J. (2011). The development of a novel measure of scientific reasoning growth among college freshmen: The Constructive Inquiry Science Reasoning Skills Test. *Journal of College Science Teaching, 40*(4), 101–107. https://www.jstor.org/stable/42992885

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411. https://doi.org/10.1080/03075079.2015.1004241

History
Received August 13, 2020
Revision received April 2, 2021
Accepted April 23, 2021
Published online July 22, 2021
EJPA Section / Category Educational Psychology

Acknowledgment
We thank Richard Shavelson for his input on the study design, Stella Bollmann for answering questions about the DIF analysis methods, and Benjamin Schweizer for his help in constructing the online version of the test instruments.

Open Science
We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p*-values, effect sizes, and 95% confidence or credible intervals. All data files and analyses scripts for this study can be found under this link: https://osf.io/b9tp4/?view_only=489519a8762a4781b51e722f798eb0f0.

Open Data: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant (Opitz et al., 2021).

Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. The CTSR can be requested from the test authors.

Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding
This work was supported by the Elite Network of Bavaria [K-GS-2012-209].

ORCID
Ansgar Opitz
https://orcid.org/0000-0002-4753-2157

Ansgar Opitz
Department of Psychology
LMU Munich
Leopoldstr. 13
80802 Munich
Germany
ansgar.opitz@psy.lmu.de