SEPARATING THE RELEVANT FROM THE IRRELEVANT: FACTORS INFLUENCING L1 STUDENT TEACHERS' ABILITY TO DISCERN (IR)RELEVANT ARGUMENTS IN TIME-PRESSURED GRAMMATICAL DISCUSSIONS

ARINA BANGA[1] AND JIMMY VAN RIJT[2]

1.Tilburg University of Applied Sciences, Department of Teacher Education
2.Tilburg University, Tilburg Center of the Learning Sciences

Abstract
Identifying relevant information and evaluating evidence are considered characteristics of critical thinking. These skills are important for language teachers, for example in evaluating pupils' grammatical reasoning in the context of grammar education. Therefore, the current study has examined whether Dutch language student teachers (N=298) in different educational tracks (Bachelor full-time, Bachelor part-time and Master) are able to distinguish relevant arguments from irrelevant (or incorrect) ones in two grammatical discussions. Results indicate that student teachers are better at evaluating relevant arguments in grammatical discussions than they are at evaluating irrelevant arguments. Multilevel analyses show that the factors partly explaining the Relevant Argument score are students' education and their Need for Cognition. The factors that partly explain the Irrelevant Argument score on the other hand are the perceived difficulty of the task, and strikingly, age. The paper discusses explanations for these findings, as well as practical implications for teacher education.

Keywords: grammar teaching; linguistic reasoning; grammatical argumentation; teacher education; critical thinking

Banga, A., & Van Rijt, J. (2023). Separating the relevant from the irrelevant: Factors influencing L1 student teachers' ability to discern (ir)relevant arguments in time-pressured grammatical discussions. L1-Educational Studies in Language and Literature, 23, 1-27. https://doi.org/10.21248/l1esll.2023.23.1.506

Corresponding author: Arina Banga, Fontys University of Applied Sciences, Tilburg, email: a.banga@fontys.nl

1. INTRODUCTION

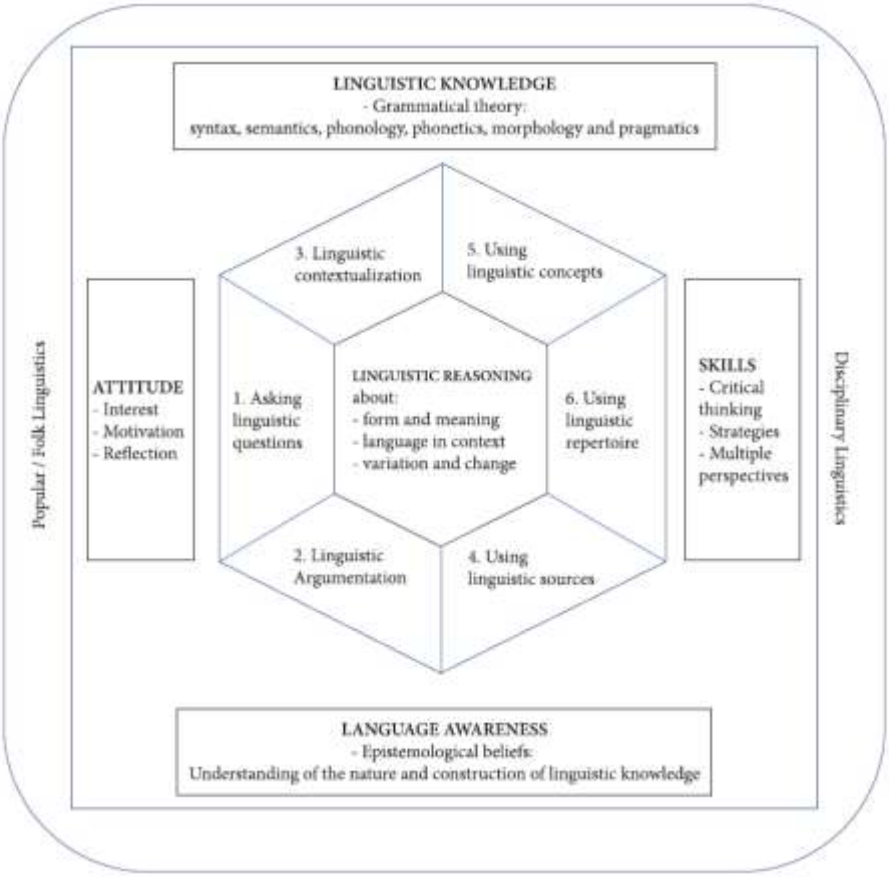
A recent trend in L1 grammar education research is that studies are increasingly focusing on the learning and teaching of grammar instead of merely providing theorized rationales for grammar teaching (Myhill, 2018; 2021). Some of the research that fits into this trend examines teachers' grammatical subject knowledge, sometimes coupled with their teacher beliefs about grammar teaching (e.g., Camps & Fontich, 2019; Gauvin et al., 2016; Macken-Horarik et al., 2018; Watson, 2012; Van Rijt et al., 2019a; 2022;). Language teachers' grammatical subject knowledge, as well as their pedagogical subject knowledge, is known to affect what students learn about grammar (Myhill et al., 2013). In other words: the more grammatical knowledge teachers have, and the more they know how to apply that knowledge in the classroom, the more likely it is that students develop grammatical understanding. The opposite also holds: teachers' lacking knowledge can cause students to develop misconceptions about the grammatical subject matter (Myhill, 2000; Myhill et al., 2013). Teachers will therefore need deep understanding about grammar, as well as a deep understanding of the pedagogical implementation thereof (Van Rijt et al., 2022).

While grammatical knowledge is a key prerequisite for developing grammatical understanding, we argue that grammatical understanding encompasses more than just knowledge, as is generally accepted in the literature on the epistemology of understanding (Baumberger, 2019; De Regt, 2009). Given this, and tying in with work by Van Rijt (2020), grammatical understanding might be defined as one's ability to describe, comprehend and/or explain (aspects of) grammar, coherently using relevant grammatical (meta)concepts, terms and/or linguistic sources, grammatical repertoire, and argumentation.

Deep grammatical understanding, as outlined by several authors, might be achieved by developing one's linguistic reasoning abilities (Dielemans & Coppen, 2021; Honda & O' Neil, 2007; Leenders et al., 2021; Van Rijt et al., 2019b, 2020). Following Van Drie and Van Boxtel's definition of historical reasoning (2008, p. 4), linguistic reasoning might be defined as an activity in which a person organizes information about language or linguistics in order to describe, compare and/or explain linguistic phenomena.

Dielemans and Coppen (2021) developed a framework describing linguistic reasoning (see Figure 1), in which they argue that engaging in linguistic reasoning means reasoning about form and meaning, context or variation and change. Such reasoning can then be broken down into six components, and it is shaped by several key factors, such as attitude, epistemic beliefs (Elsner, 2021; Wijnands et al., 2021), general thinking skills and knowledge about linguistic theory (Van Rijt & Coppen, 2017).

Figure 1. Linguistic reasoning framework adopted from Dielemans & Coppen (2021)


LINGUISTIC KNOWLEDGE
- Grammatical theory:
syntax, semantics, phonology, phonetics, morphology and pragmatics
3. Linguistic contextualization
5. Using linguistic concepts
ATTITUDE
- Interest
- Motivation
- Reflection
1. Asking linguistic questions
LINGUISTIC REASONING about:
- form and meaning
- language in context
- variation and change
6. Using linguistic repertoire
SKILLS
- Critical thinking
- Strategies
- Multiple perspectives
Popular / Folk Linguistics
Disciplinary Linguistics
2. Linguistic Argumentation
4. Using linguistic sources
LANGUAGE AWARENESS
- Epistemological beliefs:
Understanding of the nature and construction of linguistic knowledge

While Dielemans and Coppen (2021) focus on linguistic reasoning in a broad sense, their framework can equally be applied to the domain of grammar specifically, which is a part of linguistics that describes the structure and meaning of sentences and phrases. In what follows, we will therefore focus specifically on the notion of *grammatical reasoning* and the related notion of *grammatical understanding*. An important facet of grammatical reasoning, and a subcomponent in Dielemans and Coppen's (2021) framework, is using grammatical argumentation. Not only do teachers need the ability to come up with relevant grammatical arguments to underpin a grammatical analysis they are conducting themselves, but they also need the ability to evaluate their students' grammatical argumentation. Given typical classroom activities, they will often need to do so under time pressure (e.g., in response to hearing students' argumentation). Several authors have argued that engaging in grammatical or linguistic reasoning is a very useful activity for secondary school students as well (Fontich, 2016; Honda & O’ Neil, 2007; Denham, 2020; Van Rijt et al., 2020), making it even more important for their teachers to possess grammatical reasoning abilities.

While there are considerable studies to show that (student) teachers’ grammatical subject knowledge is commonly underdeveloped (e.g., Brøseth & Nygård , 2023; Cajkler & Hislam, 2002; Myhill et al., 2013; Nygård & Brøseth, 2021; Macken-Horarik et al., 2018; Van Rijt et al., 2019a;), there is hardly any work on (student) teachers’ ability to construct or evaluate grammatical arguments (Van Rijt et al., 2021)—a potential consequence of the fact that critical thinking activities are hardly integrated into language teaching (Li, 2011; Weijun Liang & Fung, 2021). Van Rijt et al. (2021) examined how Dutch student teachers perform on grammatical odd one out tasks compared to 14 year old pre-university students. In these tasks, participants were asked to provide arguments to support one of three options (e.g., which one is the odd one out? Verb A, B, C). It was found that student teachers do not always outperform 14 year old secondary school students in such tasks, and that just over half of the arguments student teachers produced were actually linguistically valid. This shows that student teachers (and 14 year olds alike) also devised a lot of arguments that were not sufficiently relevant to tackle the problem at hand, as well as arguments that were simply false. In terms of student teachers’ grammatical reasoning ability (specifically, their grammatical argumentation), it seems that there is some reason for concern, and a need for more empirical research. The current project contributes to this.

In the current study we examine student teachers’ handling of grammatical arguments in a different way. In the Van Rijt et al. (2021) study, student teachers had to come up with grammatical arguments themselves; in the current study, we examine student teachers’ ability to process and evaluate given grammatical arguments under time pressure, as is the case during their work as teachers. In particular, the study examines student teachers’ ability to identify which arguments are relevant in a particular grammatical discussion and which ones are not. We also intend to examine which other variables might be predictive of students’ ability to separate relevant grammatical information from irrelevant information.

1.1 Separating relevant grammatical arguments from irrelevant ones

Identifying relevant information and evaluating evidence are considered hallmarks of critical thinking (Glassner, 2017; Kuhn, 1991; Paulsen & Kolstø, 2022). It is generally accepted that the ability to decide which information or argumentation is relevant and valid in a given context is what separates experts from novices (Glassner, 2017). This ability is particularly helpful in ill-structured knowledge domains (cf. King & Kitchener, 1994), such as linguistics or grammar (Wijnands et al., 2021). Grammatical problems can often be tackled from multiple perspectives (‘multiperspectivity’), which means that there is not always one way to analyze a grammatical structure or one perfect solution to a grammatical problem. Therefore, if conflicting evidence emerges when a grammatical problem arises, the careful use of grammatical arguments is required. A critical part of this process, especially for teachers, is being able to identify relevant grammatical arguments, or, looking at it from the other side of the same coin: being able to identify arguments that are either not true or irrelevant. We will provide an example of an ill-structured grammatical problem below, showing how the process of weighing arguments works in grammatical discussions.

A well-known discussion in Dutch grammar deals with so called aan het + infinitive constructions, which are durative constructions such as the one in (1), with the English translation below:

(1) Vader is aan het koken.
Father is to PREP the ART cooking INF
'Father is cooking.'

Among Dutch language teachers and even among authorative linguistic handbooks and reference grammars, the *aan het + infinitive* part is either analyzed as a part of the verbal predicate, or as a subject complement (Coppen, 2009); some handbooks make no mention of this type of construction at all. There are arguments to support both analyses (see also Appendix A). For example, the construction contains a linking verb (*is*, the finite form of the verb 'to be'), which is evidenced by the fact that this linking verb can be replaced by other linking verbs (e.g., *raken 'to become', lijken 'to appear' or blijven 'to stay'*), which is taken as evidence for the subject complement analysis. An argument in favor of the verbal predicate analysis is that a direct object can be added to the sentence, such as in (2):

(2) Vader is spaghetti aan het koken.
Father is spaghetti to PREP the ART cooking INF
'Father is cooking spaghetti.'

While there are many more arguments in favor of both analyses, these examples are sufficient to illustrate that this construction constitutes an ill-structured grammatical problem, which requires weighing grammatical arguments and being able to evaluate what the arguments are in favor of. Both arguments presented are relevant and valid. Of course, not all arguments used by teachers or language learners to underpin a certain analysis are equally valid, or equally relevant for the problem at hand. For example, in Coppen's (2009) description of teachers' discussion of this very construction, one teacher remarked that the linking verb argument does not hold, as the verb is ('to be') cannot be replaced by the finite form of the linking verb *worden* ('to become'). This argument is not valid, however, as there are several constructions in Dutch in which the verb *raken* ('to become') replaces the verb *worden* (Haeseryn et al., 1997, p. 1123) such as in the combination *bewusteloos raken* ('to become unconscious')—**bewusteloos worden* would be considered ungrammatical. Since a form of *raken* can be used instead of *worden* in the target construction, this supposed counter argument does not hold.

Apart from arguments not being valid, arguments can also be irrelevant for the discussion altogether. For example, students might remark that apart from the ability to include a direct object, it is also possible to add adverbials to the construction. This might be factually true, but this information does not support any of the possible analyses. Teachers are thus in need to do two things: they must be able to assess (a) whether an argument is true and relevant or not; and they must be able to assess (b) what an argument is in support of if they decide that it is relevant. While this ability is particularly helpful in grammatical discussions or when faced with ill-structured grammatical problems, the ability is equally helpful for evaluating students' reasons and arguments in support of a particular analysis in more straightforward cases (e.g., if students are required to explain why they think X is a preposition if there can be no doubt that X is a preposition). As stressed earlier, teachers will sometimes need to process grammatical arguments under time pressure during their work in the classroom. If they have to pay attention to several things, e.g., classroom management (Pillen et al., 2013), and are expected to reason grammatically at the same time, drawing heavily on their working memory capacity, this situation possibly puts teachers in danger of cognitive overload (Sweller et al., 1998). We therefore need more insights into student teachers' time-pressured coping with grammatical arguments. The current study subsequently set out to answer the following research questions:

1) To what extent are student teachers capable of separating relevant arguments from irrelevant arguments in grammatical discussions when under time pressure?
2) Which variables are predictive of the ability to identify relevant and irrelevant grammatical arguments?

2. METHOD

2.1 Relationship with previous work

The current study is part of a larger project, examining student teachers' ability to process different types of grammatical arguments. Its twin (Van Rijt et al., 2023) used the same experimental setup and dataset. However, the twin study was concerned with completely different research questions and was subsequently aimed at different aspects of the task. In the twin study, we examined whether arguments based on linguistic manipulations (LM) posed a higher cognitive load on student teachers than arguments based on rules of thumb (RoT), examining reasoning scores, perceived mental effort (Paas, 1992) and response times. The latter two measures are frequently used for measuring cognitive load. For the perceived mental effort, we used the Mental Effort Rating Scale (Paas, 1992). It is a one item scale on which participants indicate how difficult they find a certain task. We also examined whether any processing differences between these two types of arguments could be attributed to students' Need For Cognition (cf. Cacioppo et al., 1984) and their willingness to teach grammar (their Grammar Willingness). A participant's Need For Cognition reflects the tendency to enjoy effortful cognitive processing. This may influence the (perceived) cognitive load; even if a high cognitive load is needed to complete a certain task, individuals may not report a high cognitive load if their Need For Cognition is relatively high. In addition, Nussbaum (2005) concluded that participants with high Need For Cognition scores generated more elaborated arguments, so this may influence linguistic reasoning. Another factor that may influence this, is Grammar Willingness, i.e., a participant's affinity with grammar and grammar teaching. Finally, we examined whether teacher training impacts students' ability to process these two different types of arguments. The current study is not concerned with the difference between LM and RoT. It solely investigates the relevance of arguments. For more background details on the measures used, we refer the reader to the twin study (Van Rijt et al., 2023).

2.2 Participants

In the current study, 298 student teachers, training to become Dutch language and literature teachers, participated. 229 of them identified as female, 68 identified as male and 1 student preferred not to disclose this. The students were recruited at 8 (out of 9) institutions for Dutch teacher training. They were either full-time bachelor students (*N* = 156, *M* age = 20.56 years, *SD* = 2.28), part-time bachelor students (*N* = 99, *M* age = 36.72 years, *SD* = 10.64) or part-time master students (*N* = 43, *M* age = 34.95, *SD* = 11.44). Bachelor students will receive a second degree teaching license and a Bachelor of Education (BEd) degree upon completing their studies, which will allow them to teach in the lower levels of secondary schools (where most grammar is taught in the Netherlands). Master students will obtain a first degree teaching license as well as a Master of Education degree (MEd), enabling them to also teach in the higher levels of secondary schools. Note that there is a considerable age difference between the full-time students and the part-time students. This is because full-time students typically enroll in teacher education immediately after completing secondary education (usually senior general secondary education or *havo*), whereas BEd part-time students typically start later as a result of a career switch. MEd students can only start their track after having graduated as a BEd teacher. The participants were not equally divided over the various institutions, which was a result of practical limitations for institutions to take part in the investigation (as the experiment was conducted during regular class hours). This means that some institutions were only able to provide us with certain groups of students (e.g., only master students). Participants voluntarily participated in the study, signing active consent forms, which enabled us to process their data anonymously. The institutions approved of the investigation. Students were only informed of the specific aim of the study afterwards.

2.3 Experimental setup and materials

One of the purposes of the present experiment was to measure student teachers' linguistic reasoning under time pressure, because they experience this, to a certain extent, in real classroom situations in which they must consider grammatical arguments from their pupils at the same time as other teaching tasks (e.g., classroom management).

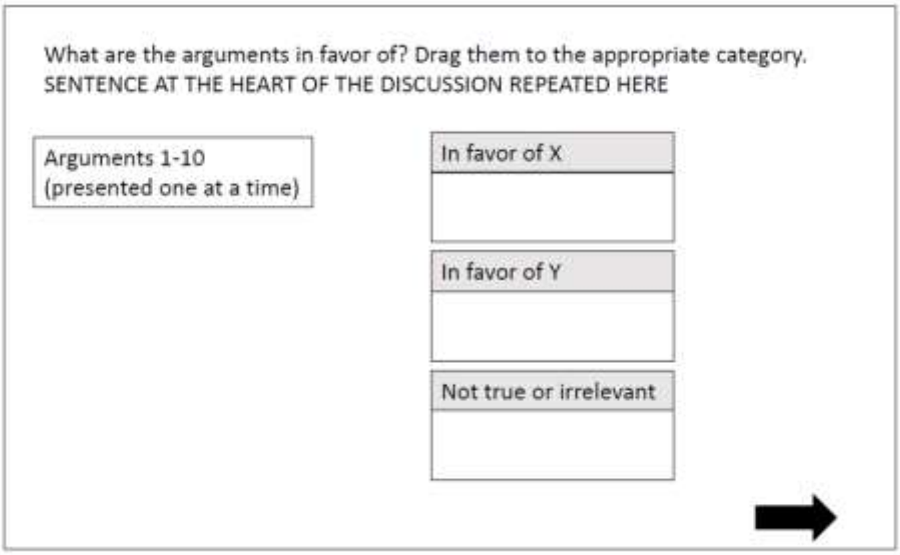
Sometimes it can be useful to make explicit the uncertainty in grammatical reasoning. Of course, it is also possible for a teacher to say that he will come back to a certain question or argument in the next lesson if a pupil comes up with something unexpected, as an anonymous reviewer pointed out. However, not all linguistic reasoning is unexpected; language teachers must master this up to a point.

For this study, we developed an experimental task in Qualtrics that featured two grammatical discussions. See Table 1 for the two discussions and a few example arguments. Each grammatical discussion consisted of a sentence that was being analyzed by two fictional teachers, where teacher A would argue that a certain part of the sentence should be analyzed as X, whereas teacher B would argue that the same part should be analyzed as Y. Per grammatical discussion, participants were then presented with a randomly ordered set of 10 grammatical LM-based arguments, and subsequently with a randomly ordered set of 10 RoT-based arguments, or vice versa. The arguments for RoT were matched with the arguments for LM for number of words, number of grammatical terms and syntactic structure. These characteristics namely may influence the required mental effort during the categorization task. For case A the number of words was 181, the number of grammatical terms was 5, there were 2 singular sentences, there were 4 compound sentences containing conjunction *if* only and there were 4 compound sentences containing other conjunctions than *if*. For case B the number of words was 153, the number of grammatical terms was 8, there were 5 singular sentences, there were 2 compound sentences containing conjunction *if* only and there were 3 compound sentences containing other conjunctions than *if*.

Which set of 10 arguments they were given first was randomly decided as a result of our Qualtrics-setup. Participants were then asked to categorize each argument (presented one at a time) in the following way: the argument could either be in favor of what teacher A argued (we call this a relevant argument given the context), it could be in favor of what teacher B argued (we call this a relevant argument given the context), or the argument could be not true or irrelevant for the discussion. An example of the latter category is: *You can at least add another adverbial to this sentence without it becoming ungrammatical* in Case A, in which two teachers argue about *de 400 meter*, which can be analyzed as an adverbial or as a direct object. This argument is irrelevant given the context. This category also includes incorrect arguments, for example in Case A: The relevant phrase should actually be *weer de 400 meter* instead of *de 400 meter*. Appendix A for all relevant and irrelevant arguments and the mean percentage of the correct categorizations per argument.

Students individually had to drag each argument to the appropriate box using a computer mouse or a touch pad, and they were instructed to do this as well as they could, but also as quickly as they could, to add some time pressure to the task. The time students took per set of arguments was tracked invisibly in the background. Prior to tackling the grammatical discussion tasks, students completed the Need For Cognition test (Cacciopo et al., 1984), which is an instrument measuring a person's enjoyment of cognitively demanding tasks. The Need For Cognition test was highly reliable (Cronbach's alpha = .85). Student teachers also responded to four five-point Likert scale questions that aimed to measure students' willingness to engage in grammar (teaching), the Grammar Willingness test, which was also sufficiently reliable (Cronbach's alpha = .71). Prior to tackling the grammatical discussion tasks, participants were presented with a practice task, in which they categorized arguments in favour of either online teaching or teaching on campus. This was done to make sure students were familiar with the procedure. Figure 2 is a representation of participants' task screen. In Table 1, we will present the two grammatical discussions we used.

Figure 2. Representation of participants' task screen. Note 'In favour of X/Y' corresponds with 'In favour of what teacher A/B says'. Clicking the arrow would take students to the next screen and register the time taken. Figure adopted from Van Rijt et al. (2023).



What are the arguments in favor of? Drag them to the appropriate category.
SENTENCE AT THE HEART OF THE DISCUSSION REPEATED HERE

After each categorization task (i.e., after having categorized a set of 10 arguments), we asked students to indicate how difficult they felt the task was. To this end, we used Paas' (1992) one-item Mental Effort Rating Scale (MERS), which participants used to indicate how difficult they perceived the task to be on a scale ranging from 1-9, where 1 means *very, very little effort* and 9 means *very, very much effort*.

The experiment took between 15 and 20 minutes in total, in which student teachers completed the practice task, took the Need For Cognition test and the Grammar Willingness test and categorized 40 grammatical arguments, 20 per grammatical discussion, followed up by the Mental Effort Rating Scale after each set of ten arguments. See Appendix A for an overview of all arguments. The arguments were constructed by the authors, based on reference grammars, and they were independently verified by another linguist, who did not participate in the experiment.

Table 1. The two grammatical discussions and a few examples of arguments presented in the task

Grammatical discussion	**Argument in favor of what teacher A says**	**Argument in favor of what teacher B says**	**Not true/irrelevant argument**
1. *Afgelopen winter hebben we voor het eerst in tijden weer de 400 meter geschaatst.* ('Last winter, we ice-skated the 400 meters again for the first time in ages'). Teacher A argues that *de 400 meter* is an adverbial; teacher B argues that it is a direct object.	If you want to replace the phrase *de 400 meter*, you can replace it by the phrase *400 meter ver*.	If you passivize the sentence, the result is *'De 400 meter wordt door ons geschaatst'.*	If you put the phrases in a different order, *de 400 meter* can be placed sentence-first.
2. *Arie is na het puzzelen met plezier aan het koken.* ('Arie is happy to cook after puzzling'). Teacher A argues that *aan het koken* is a verbal predicate; teacher B argues that it is a subject complement.	If you remove the finite verb ('is'), the verb *kookt* will appear in that place.	You can replace *is* with other linking verbs, for example *blijft*, *lijkt* and *raakt*.	The verb *koken* is plural in this sentence.

In total, student teachers could earn 40 points, 1 point per correctly categorized argument in one of the three categories. 25 points could be earned for relevant arguments (that pointed either in the direction of what teacher A says, or in the direction of what teacher B says); 15 points could be earned by correctly categorizing arguments as not true or irrelevant. We therefore calculated Relevant Argument scores and Irrelevant Argument scores.

2.4 Analysis

To examine whether student teachers can separate relevant arguments from irrelevant arguments in grammatical discussions when under time pressure, we compared their Relevant Argument scores with their Irrelevant Argument scores using a simple paired samples T-test. Because the total number of points that could be earned for each category differs (25 for the relevance category vs. 15 for the irrelevance category), we calculated relative scores (i.e., the percentage of correctly categorized Relevant and Irrelevant arguments) that served as the input for the comparison (and for all further analyses).

For our second research question, in which we wanted to examine which factors might account for students' ability to separate relevant information from irrelevant information, we first examined correlations between potentially relevant variables and the relative Relevant Argument and Irrelevant Argument scores. Significantly correlating variables were then used as input for further analysis. To this end, we constructed two multilevel regression models (Restricted Maximum Likelihood), taking into account the fact that students are nested within institutions: one model for the Relevant Argument score, and one model for the Irrelevant Argument score, based on the significantly correlating variables. This enabled us to see whether different variables might account for participants' Relevant Argument score than for participants' Irrelevant Argument score. For the multilevel modelling, we used the GAMLj package available in jamovi (Gallucci, 2019).

3. RESULTS

3.1 Descriptives

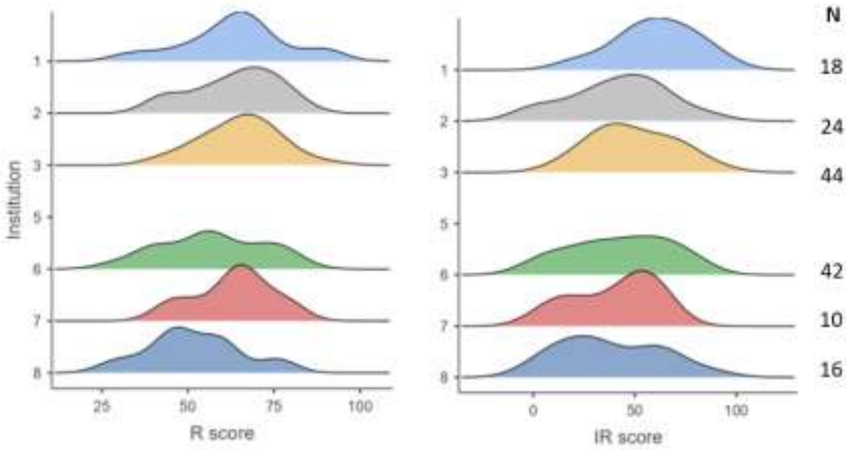
Table 2 lists the relevant descriptive statistics for this study.

Table 2. Mean scores and standard deviations per educational track per variable: Grammar Willingness (GW), Need For Cognition (NFC), Response time, Mental Effort Rating Scale score (MERS), Relevant Argument score (R score), Irrelevant Argument score (IR score)

Track	**N**	**GW (SD)**	**NFC (SD)**	**Response time (SD)**	**MERS (SD)**	**R score / 25 (SD)**	**IR score / 15 (SD)**
BEd full-time	156	3.62 (0.67)	3.32 (0.52)	105.76 (39.35)	6.13 (1.45)	15.3 (3.45)	6.88 (3.45)
BEd part-time	99	3.86 (0.58)	3.58 (0.49)	120.33 (36.83)	5.80 (1.28)	14.7 (3.55)	8.83 (3.07)
MEd	43	3.92 (0.61)	3.82 (0.54)	132.60 (60.60)	5.30 (1.19)	15.3 (4.26)	8.12 (3.13)
Total	298	3.74 (0.64)	3.48 (0.54)	114.00 (43.3)	5.90 (1.39)	15.1 (3.61)	7.71 (3.39)

Figure 3 shows how Relevant Argument and Irrelevant Argument scores (relative measures, i.e., percentages) differ across the different institutions for our largest subgroup of students (bachelor full-time), making it plausible that potential differences may be caused by the institutions that students are nested in. This warrants the use of multilevel modelling, to take the influence of institution out of the equation. This assumption was further strengthened by calculating the ICC for the total reasoning score (combining Relevant Argument and Irrelevant Argument scores), which was 0.09, showing that between-institution differences can account for 9% of the variation in reasoning scores. As even small ICC's are a reason to adopt multilevel models (Nezlek, 2008) we have done so for answering our second research question.

Figure 3. Density plots for Full-time Bachelors students' (N = 154) Relevant Argument (R score) and Irrelevant Argument (IR score) scores per institution. The figure shows some variation over institutions in both scores. Institution 4 did not provide any bachelor full-time students; institution 5 only provided 2, so no density plot could be generated for that institution.



3.2 Differences between R and IR scores

As Figure 4 shows, there are substantial differences between student teachers' Relevant Argument score (*M* = 60.4, *SD* = 14.4) and their Irrelevant Argument score (*M* = 51.4, *SD* = 22.6). A paired samples T-test shows that this difference is statistically significant *t*(297) = 5.82, *p* <.001, Cohen's *d* = 0.34.

To gain a sense of how the different arguments were categorized over the three categories (In favor of A / in favor of B / not true or irrelevant), Appendix A shows how the different arguments were categorized per task. Since we were interested in the overall difference between Relevant Argument and Irrelevant Argument scores, regardless of different tasks, the following analyses will only report on total R and IR scores.

3.3 Factors influencing relevant argument and irrelevant argument scores

To examine which factors might influence student teachers' Relevant Argument and Irrelevant Argument scores, we constructed a Pearson's correlation matrix for each dependent variable, using the other variables as input. See Tables 3 and 4.

Figure 4. Differences between Relevant Argument (R) and Irrelevant Argument (IR) scores. The Figure shows the mean and median percentage of correct categorizations per score type

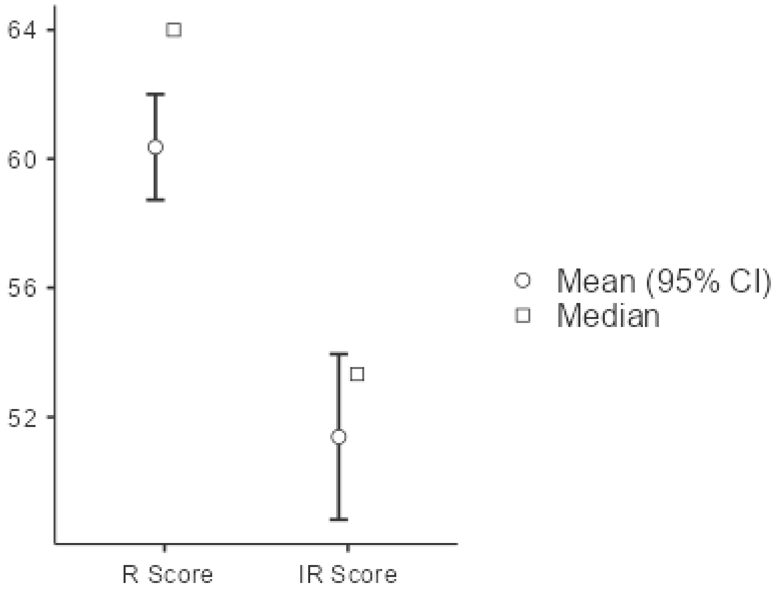


Table 3. Pearson's correlation matrix showing correlations between the Relevant Argument score (R score) and age, Grammar Willingness (GW), Need For Cognition (NFC), Mental Effort Rating Scale (MERS) and response time

		R score	Age	GW	NFC	MERS	Response time
R score	Pearson's r	-					
	p-value	-					
Age	Pearson's r	-0.030	-				
	p-value	0.603	-				
GW	Pearson's r	0.074	0.065				
	p-value	0.203	0.267				
NFC	Pearson's r	0.159**	0.261***	0.246***	-		
	p-value	0.006	< .001	< .001	-		
MERS	Pearson's r	-0.063	-0.134*	-.346***	-0.304	-	
	p-value	0.276	0.021	<.001	<.001	-	
Response time	Pearson's r	0.049	0.201***	0.087	0.108	0.023	-
	p-value	0.400	<.001	0.134	0.064	0.692	-

*Note. * p < .05, ** p < .01, *** p < .001*

Table 3 shows a modest significant positive correlation between student teachers' Need for Cognition score and the Relevant Argument score, indicating that the more they enjoy complex problems, the better they seem to perform on this variable. No other variables correlate significantly with the Relevant Argument score. There are however positive correlations between Need For Cognition and Grammar Willingness (i.e., the higher student teachers' Need For Cognition, the more likely they are to enjoy grammar), and there is a negative correlation between students perceived mental effort (MERS) and their Need For Cognition, indicating that the more difficult the tasks were perceived, the lower the Need For Cognition. There was also a negative correlation between the Mental Effort Rating Scale score and Grammar Willingness, indicating that the more difficult the tasks were perceived, the less a participant liked to engage in grammar. Mental Effort Rating Scale scores also negatively correlates with age (i.e., the older a student is, the less difficult they perceived the tasks to be). There are positive correlations between Need For Cognition and age (i.e., the older a student is, the higher the Need For Cognition). Lastly, age and response time correlate significantly, which is to be expected (participants react slower as they age).

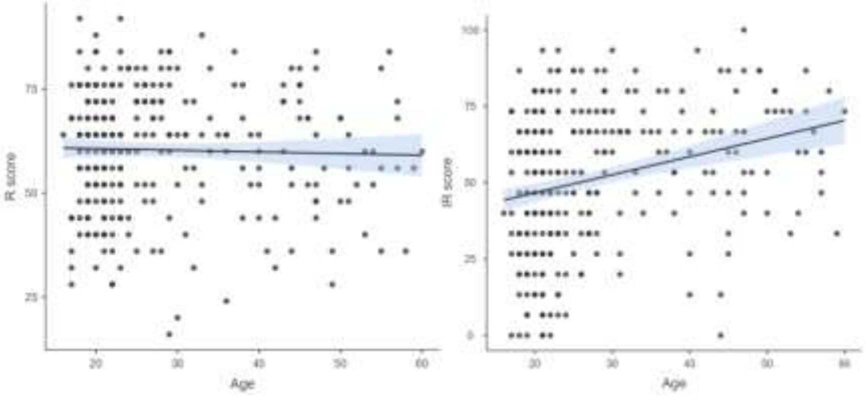
Table 4. Pearson's correlation matrix showing correlations between the Irrelevant Argument score (IR score) and age, Grammar Willingness (GW), Need For Cognition (NFC), Mental Effort Rating Scale score (MERS) and response time

		IR score	**Age**	**GW**	**NFC**	**MERS**	**Response time**
IR score	Pearson's r	-					
	p-value	-					
Age	Pearson's r	0.287***	-				
	p-value	<.001	-				
GW	Pearson's r	0.074	0.065				
	p-value	0.200	0.267				
NFC	Pearson's r	0.127*	0.261***	0.246***	-		
	p-value	0.028	< .001	< .001	-		
MERS	Pearson's r	-0.251***	-0.134*	-.346***	-0.304	-	
	p-value	<.001	0.021	<.001	<.001	-	
Response time	Pearson's r	0.010	0.201***	0.087	0.108	0.023	-
	p-value	0.858	<.001	0.134	0.064	0.692	-

*Note. * p < .05, ** p < .01, *** p < .001*

Table 4 shows significant positive correlations between Irrelevant Argument score and age, indicating that the older student teachers are, the higher their Irrelevant Argument score tends to be. The table also shows a modest positive correlation between Need For Cognition and Irrelevant Argument score. In addition, perceived mental effort (MERS) correlated negatively with Irrelevant Argument score, meaning that Irrelevant Argument score is more likely to be higher if Mental Effort Rating Scale scores are lower. Interestingly, age seems to matter more when it comes to deciding what is irrelevant than it is when it comes to deciding what is relevant, as is shown visually in Figure 5.

Figure 5. Correlations between Relevant Argument score (first panel) and Irrelevant Argument score (second panel) and age. Each dot represents an individual observation. The correlation is shown with the linear regression line, with standard error marked in blue.



To gain a more robust idea of the relationship between these significant variables and Relevant Argument and Irrelevant Argument scores, we entered the variables in a multilevel regression model. Because the twin study revealed that master student teachers outperformed bachelor students teachers, we took educational level into account as a predictor, and only entered the significantly correlating variables after that. Table 5 shows the multilevel regression model for the Relevant Argument score; Table 6 shows the multilevel regression model for the Irrelevant Argument score. The marginal R2 value of the model presented in Table 5 was 0.05, indicating that the model can explain 5% of the variance in R scores. The marginal R2 value of the model presented in Table 6 was 0.14, meaning that the model can explain 14% of the variance in Irrelevant Argument scores. The syntax for the model presented in Table 5 is as follows: R score ~ 1 + Education + NFC+(1 | Institution) +(StudentID), where R score means Relevant Argument score and NFC means Need For Cognition score. The syntax for model 6 is as follows: IR score ~ 1 + Education + MERS + Age+(1 | Institution)+(StudentID), where IR score means Irrelevant Argument score and MERS means Mental Effort Rating Scale. Both models showed better fit than an intercept only model (Model 5's intercept only model: AIC = 2440.8761; Model 6's intercept only model: AIC = 2699.194). Lower AIC's reflect better models.

Table 5 shows that students' education can partly account for differences in the R score: the difference between full-time and part-time bachelor students can significantly better predict students' R scores than an intercept only model (p = 0.026). The model shows that part-time bachelor students on average score -5.42% compared to the grand mean of intercepts (59.00). Knowing whether a student is a master of bachelor full-time student does not improve the model but knowing a student's Need For Cognition does (p<.001): for each point more on Need For Cognition, students' R score increases with 5.28%.

Table 5. Multilevel regression model for Relevant Argument score (AIC = 2432.8004); NFC = Need For Cognition

				95% confidence interval				
	Effect	***b***	**ß**	**Lower**	**Upper**	**df**	**t**	***p***
(Intercept)	(Intercept)	59.00	1.49	56.08	61.916	7.94	39.61	< .001
Education1	BEd part-time - BEd full-time	-5.42	2.37	-10.06	-0.774	50.84	-2.29	0.026
Education2	MEd - BEd full-time	-3.13	2.81	-8.64	2.368	109.17	-1.12	0.267
NFC	NFC	5.28	1.59	2.17	8.391	296.03	3.33	< .001

Table 6. Multilevel regression model for Irrelevant Argument score (AIC = 2674.698); NFC = Need For Cognition, MERS = Mental Effort Rating Scale.

				95% confidence interval				
Names	**Effect**	**Esti-mate**	**SE**	**Lower**	**Upper**	**df**	**t**	***p***
(Intercept)	(Intercept)	51.414	1.992	47.510	55.319	7.03	25.8082	< .001
Education1	BEd part-time - BEd full-time	6.027	4.184	-2.173	14.226	68.49	1.4405	0.154
Education2	MEd - BEd full-time	-1.525	4.630	-10.600	7.550	134.58	-0.3293	0.742
MERS	MERS	-3.415	0.926	-5.230	-1.600	297.33	-3.6871	< .001
Age	Age	0.422	0.159	0.109	0.734	297.93	2.6466	0.009
NFC	NFC	0.150	2.456	-4.664	4.963	296.11	0.0610	0.951

Table 6 shows that students' education seems to have no significant predictive value for the Irrelevant Argument score. Students' MERS, their perceived mental effort, does have predictive value, as it outperforms a model that includes intercept and education type (p <.001). The model predicts that the higher the Mental Effort Rating Scale score is (i.e., the more perceived mental effort), the poorer a student's IR score. For each step up on the Mental Effort Rating Scale, the model estimates that the Irrelevant Argument score drops by 3.42%. Student teachers' age is also a predictor for Irrelevant Argument score: with each increasing year of age, Irrelevant argument score increases with 0.42%. Strikingly, Need For Cognition correlated significantly with Irrelevant Argument score (Table 4), although it does not hold up in the regression model. Removing the Need For Cognition variable from the model presented in Table 6 would improve the model's AIC to 2672.701.

4. DISCUSSION

4.1 Interpretation of main findings

The present study set out to investigate a) to what extent student teachers are capable of separating relevant arguments from irrelevant arguments in grammatical discussions when under time pressure and b) which variables are predictive of the ability to identify relevant and irrelevant grammatical arguments. In examining these questions, the study attempted to contribute to knowledge on reasoning skills in subject specific contexts, in this case, grammar.

With regard to a), the results show that when time pressured, student teachers can more adequately handle relevant grammatical arguments than irrelevant grammatical arguments. On average, student teachers manage to correctly assess 60.4% of all relevant arguments, and only 51.4% of all irrelevant arguments, which is just slightly above chance level. These percentages show that student teachers' overall score is not as high as one would hope for student teachers who are engaged in reasoning tasks about their own area of expertise. Van Rijt et al. (2021, p. 10) note for similar reasoning tasks that a threshold of 80% correct is a minimum requirement to be able to effectively teach grammar. Student teachers' current performances are far removed from that threshold. If it is considered that assessing grammatical arguments in real classroom situations is likely to be even more time sensitive than in the current task, these percentages may even overestimate student teachers' time pressured abilities.

The question might arise why student teachers are better at categorizing relevant arguments than at categorizing irrelevant arguments in grammatical discussions. This seems difficult to explain at first sight (i.e., if a student knows what is relevant in a certain grammatical case, this student should also know what is irrelevant in the same case), this could partly be explained by the kind of grammar education they received. The student teachers in our study predominantly attended traditional grammar classes (mostly parsing exercises), as is common praxis in secondary school and later at the teacher training institutes in the Netherlands. Because they are not specifically taught to handle messy grammatical problems (Van Rijt et al., 2021), they are not systematically trained to reason grammatically, let alone to handle irrelevant or incorrect grammatical arguments in the context of grammatical discussions. Given their experience in grammar education, it is plausible that the student teachers in the present study are better at categorizing relevant arguments than irrelevant arguments. Nevertheless, they should be better at categorizing both types of arguments than they are in the present study.

With regard to question b), different factors explain the Relevant Argument scores on the one hand and the Irrelevant Argument (IR) scores on the other.

Relevant Argument scores are partly accounted for by student teachers' education (i.e., BEd full-time, BEd part-time, MEd) and Need for Cognition, whereas Irrelevant Argument scores do not. It seems plausible that education only affects

Relevant Argument scores and not Irrelevant Argument scores, because student teachers are more likely to use and discuss relevant grammatical arguments in class than irrelevant arguments. BEd full-time students outperform BEd part-time students with respect to Relevant Argument scores. This could be explained by the fact that part-time students typically are expected to work through the subject matter in a more individual manner, receiving less conceptual and didactic guidance by a teacher educator. MEd students, even though engaged in a more advanced track, typically do not receive additional courses in school grammar on top of the ones that they received in the BEd. This might explain why MEd students do not perform better than BEd students. We leave this matter open for future research. Another predictor for the Relevant Argument scores is Need For Cognition. When solving a cognitively demanding problem, student teachers may well find that relevant arguments bring them closer to a satisfying solution than irrelevant arguments, so they may be more focused on those. This might explain why Need For Cognition predicts Relevant Argument scores, but not Irrelevant Argument scores. Future studies might also explore this.

Irrelevant Argument scores were explained by student teachers' age and perceived mental effort (MERS). As far as Mental Effort Rating Scale scores are concerned, it seems that the more difficult students perceive a complete reasoning task to be, the poorer they perform in terms of correctly categorizing irrelevant arguments. This would suggest that it is mostly the irrelevant arguments, and not the relevant arguments, that determine the overall difficulty rating of a task. This aligns with the finding that students perform much worse on Irrelevant Arguments than on Relevant Arguments.

The relationship between age and Irrelevant Argument scores, which is absent between age and Relevant Argument scores, is perhaps the most striking result. It suggests that student teachers' natural cognitive development across the lifespan might play a larger role than their education, as education could not explain student teachers' Irrelevant Argument scores and age did. This raises the question whether the ability to separate irrelevant arguments from relevant ones might—to an extent—be untrainable; rather, students might naturally develop these skills as they grow older because of cognitive maturing. The question remains whether despite this, student teachers' Irrelevant Argument scores could still increase with the right training, regardless of their natural cognitive maturity. While previous studies suggest that students' general grammatical reasoning skills can improve as a result of targeted interventions (cf. Van Rijt et al., 2019b, 2020, 2021), the question remains whether a similar result can be obtained for reasoning about irrelevant arguments specifically.

Another intriguing question is why older student teachers seem to be better at evaluating irrelevant arguments than younger ones. While the current study cannot provide definitive answers to this question, the finding is congruent with findings from neuropsychological studies showing that in certain areas, cognitive ability increases as people age, especially when it comes to skills such as learning ability and pattern recognition (see e.g., Goldberg, 2005). The ability to distinguish relevant information from irrelevant information might be seen as an instantiation of pattern recognition: older students may simply be better at identifying irrelevant arguments because they can more easily recognize them as not being part of a pattern. For instance, certain rules of thumb-based arguments are highly associated with a particular grammatical function (e.g., 'who or what + subject + verbal predicate = direct object'), whereas others are not. Older students have gained more experience with pattern recognition in general, so they might therefore also be more capable of determining quickly which arguments do not fit within a relevant pattern. This merits further research that falls outside the scope of the current study.

4.2 Limitations

The current study also has a few limitations that might influence the interpretation of the results. First, while the current study provides some insights into student teachers' grammatical reasoning scores, these cannot be seen as exemplary for students' overall grammatical competence. The student teachers might have performed better if there was no time pressure, or worse if there was more time pressure (e.g., in a real classroom situation). In addition, the effect of time pressure may vary depending on argument type (Relevant Arguments versus Irrelevant Arguments): given student teachers' difficulties with Irrelevant Arguments, these might take longer processing time than Relevant Arguments. Future research should therefore examine student teachers' grammatical reasoning processes more elaborately, both with and without time constraints, and both within and outside of the secondary school classroom. Such research should not merely focus on quantitative aspects of reasoning, but also on qualitative aspects, so a more complete picture about grammatical reasoning emerges. For instance, think-aloud protocols could be used to examine *why* student teachers would consider certain arguments to be relevant or irrelevant, shedding light on sources of confusion or misconception, which can then be specifically targeted in teacher education.

Another limitation is that the results we obtained may be task dependent. While the effects we found were measured across the board (i.e., for both reasoning tasks) and therefore have some internal validity, Appendix A shows there are some important differences in how different arguments belonging to different reasonings tasks are scored. This suggests that task type and content, e.g., specific concept use or the familiarity of the problems, can have a profound influence on student teachers' Relevant Argument and Irrelevant Argument scores.

A final limitation is that the results from this study cannot be fully generalized to teacher training institutions in other educational jurisdictions, as there are substantial international differences in terms of how teachers are being prepared to teach grammar (Boivin et al., 2018).

4.3 Implications

Despite the abovementioned limitations, the study does have some implications for (Dutch) teacher education. First, as the study shows how student teachers struggle with grammatical reasoning, it is recommended that teacher trainer institutes move beyond traditional parsing exercises, and that they teach students how to reason grammatically. Rather than avoiding the complexities of grammatical discussions, student teachers should be taught how to engage in them (Wijnands et al., 2021; Van Rijt et al., 2021). A crucial aspect in this is that they need to be equipped to evaluate grammatical arguments, both relevant and irrelevant ones. This way, student teachers can grow towards being more confident in grammar, being able to teach their pupils how to reason about grammar as well (Wijnands et al., 2022). This would enable them to reason more like linguistic experts do, as separating relevant from irrelevant information is a hallmark of critical thinking. A second implication is then that student teachers should also be taught how to stimulate grammatical reasoning in the classroom, for which Dielemans and Coppen's (2021) framework can serve as a blueprint.

REFERENCES

Baumberger, C. (2019). Explicating Objectual Understanding: Taking Degrees Seriously. *Journal for General Philosophy of Science, 50*, 367–388. https://doi.org/10.1007/s10838-019-09474-6

Boivin, M.-C., Fontich, X., Funke, R., García-Folgado, M.-J., & Myhill, D. (2018). Working on grammar at school in L1 education: Empirical research across linguistic regions. *L1-Educational Studies in Language and Literature*, *18*(3), 1–6. https://doi.org/10.17239/L1ESLL-2018.18.04.01

Brøseth, H., & Nygård, M. (2023). Norwegian first-year student teachers' knowledge of L1 grammar. *L1-Educational Studies in Language and Literature*, *23*(1), 1–30. https://doi.org/10.21248/l1esll.2023.23.1.411

Cacioppo, J., Petty, R., & Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306-307. https://doi.org/10.1207/s15327752jpa4803_13

Cajkler, W., & Hislam, J. (2002). Trainee Teachers' Grammatical Knowledge: The Tension Between Public Expectation and Individual Competence. *Language Awareness, 11*(3), 161-177. https://doi.org/10.1080/09658410208667054

Camps, A., & Fontich, X. (2019). Teachers' concepts on the teaching of grammar in relation to the teaching of writing in Spain: a case study. *L1-Educational Studies in Language and Literature, 19*, 1-36. https://doi.org/10.17239/L1ESLL-2019.19.02.02

Coppen, P.-A. (2009). *Leren tasten in het duister* [Learning to grope in the dark]. Inaugural address. Nijmegen: Radboud University.

Denham, K. (2020). Positioning students as linguistic and social experts. Teaching grammar and linguistics in the United States. *L1-Educational Studies in Language and Literature, 20*(3), 1-16. https://doi.org/10.17239/L1ESLL-2020.20.03.02

De Regt, H. (2009). The epistemic value of understanding. *Philosophy of Science, 76*(5), 585–597. https://doi.org/10.1086/605795

Dielemans, R., & Coppen, P.-A. (2021). Defining linguistic reasoning. Transposing and grounding a model for historical reasoning to the linguistic domain. *Dutch Journal of Applied Linguistics, 9*(1/2), 182-206. https://doi.org/10.1075/dujal.19038.die

Elsner, D. (2021). Knowledge about grammar and the role of epistemological beliefs. *Pedagogical Linguistics 2*(2), 107-128. https://doi.org/10.1075/pl.21003.els

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. Retrieved from https://gamlj.github.io/.

Gauvin, I. (2016). *Savoirs en grammaire et en didactique de la grammaire chez des étudiants en enseignement du français au secondaire*. Action concertée sur l'écriture, rapport no 2013-ER-164739. Fonds de recherche du Québec – Société et culture.

Glassner, A. (2017). Evaluating arguments in instruction: Theoretical and practical directions. *Thinking Skills and Creativity, 24*, 95-103. https://doi.org/10.1016/j.tsc.2017.02.013

Goldberg, E. (2005). *The wisdom paradox. How Your Mind Can Grow Stronger As Your Brain Grows Older.* Gotham Books.

Haeseryn, W., Romijn, K., Geerts, G., Rooij, J., & Van den Toorn, M.C. (1997). *Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Martinus Nijhoff Uitgevers.

Honda, M., & O'Neil, W. (2007). *Thinking Linguistically. A Scientific Approach to Language*. Wiley-Blackwell.

King, P., & Kitchener, K. (1994). *Developing reflective judgement*. Jossey-Bass.

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Leenders, G., De Graaff, R., & Van Koppen, M. (2021). Hoe meet je bewuste taalvaardigheid? Grammaticaal redeneren in de vakken Nederlands, Engels en Duits. *Pedagogische Studiën, 98*(1), 67-93.

Li, L. (2011). Obstacles and opportunities for developing thinking through interaction in language classrooms. *Thinking Skills and Creativity, 6*(3), 146-158. https://doi.org/10.1016/j.tsc.2011.05.001

Macken-Horarik, M., K. Love, & S. Horarik (2018). Rethinking Grammar in Language Arts: Insights from an Australian Survey of Teachers' Subject Knowledge. *Research in the Teaching of English, 52*(3), 288–316.

Myhill, D. (2018). Grammar as a meaning-making resource for improving writing. *L1-Educational Studies in Language and Literature, 18*(3), 1–21. https://doi.org/10.17239/L1ESLL-2018.18.04.04

Myhill, D. (2000). Misconceptions and Difficulties in the Acquisition of Metalinguistic Knowledge. *Language and Education, 14*(3), 151-163. https://doi.org/10.1080/09500780008666787

Myhill, D. (2021). Grammar re-imagined: foregrounding understanding of language choice in writing. *English in Education, 55*(3), 265-278. https://doi.org/10.1080/04250494.2021.1885975

Myhill, D., Jones, S., & Watson, A. (2013). Grammar matters: How teachers' grammatical knowledge impacts on the teaching of writing. *Teaching and Teacher Education, 36*, 77-91. https://doi.org/10.1016/j.tate.2013.07.005

Nezlek, J. B. (2008). An introduction to multilevel modeling for Social and Personality Psychology. *Social and Personality Psychology Compass*, *2*(2), 842–860. https://doi.org/10.1111/j.1751-9004.2007.00059.x

Nussbaum, E. M. (2005). The effect of goal instructions and need for cognition on interactive argumentation. *Contemporary Educational Psychology*, *30*(3), 286-313. https://doi.org/10.1016/j.cedpsych.2004.11.002

Nygård, M., & Brøseth, H. (2021). Norwegian teacher students' conceptions of grammar, *Pedagogical Linguistics, 2*(2), 129-152. https://doi.org/10.1075/pl.21005.nyg

Paas, F. (1992).Training strategies for attaining transfer of problem-solving skills in statistics: a cognitive load approach. *Journal of Educational Psychology, 84*(4): 429-434. https://doi.org/10.1037/0022-0663.84.4.429

Paulsen, V. & Kolstø, S. (2022). Students' reasoning when faced with test items of challenging aspects of critical thinking. *Thinking Skills and Creativity, 43*, 1-13. https://doi.org/10.1016/j.tsc.2021.100969

Pillen, M., Beijaard, D., & Den Brok, P. (2013). Tensions in beginning teachers' professional identity development, accompanying feelings and coping strategies. *European Journal of Teacher Education, 36*(3), 240-260. https://doi.org/10.1080/02619768.2012.696192

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296. https://doi.org/10.1023/A:1022193728205

Van Drie, J., & Van Boxtel, C. (2008). Historical reasoning: towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review, 20*(2), 87-110. https://doi.org/10.1007/s10648-007-9056-1

Van Rijt, J. (2020). *Understanding grammar: the impact of linguistic metaconcepts on L1 grammar education*. Docotoral dissertation. Nijmegen: Radboud University.

Van Rijt, J., Wijnands, A., & Coppen, P.-A. (2019a). Dutch teachers' beliefs on linguistic concepts and reflective judgement in grammar teaching. Contribution to a special issue What is Grammar in L1 Education Today?, edited by Kaisu Rättyä, Elżbieta Awramiuk, and Xavier Fontich. *L1-Educational Studies in Language and Literature, 19*(2), 1-28. https://doi.org/10.17239/L1ESLL-2019.19.02.03 .

Van Rijt, J., De Swart, P., Wijnands, A., & Coppen, P.-A. (2019b). When students tackle grammatical problems: Exploring linguistic reasoning with linguistic metaconcepts. *Linguistics and Education, 52*, 78-88. https://doi.org/10.1016/j.linged.2019.06.004

Van Rijt, J., Wijnands, A., & Coppen, P.-A. (2020). How secondary school students may benefit from linguistic metaconcepts to reason about L1 grammatical problems. *Language and Education, 34*(3), 231-248. https://doi.org/10.1080/09500782.2019.1690503

Van Rijt, J., Hulshof, H., & Coppen, P.-A. (2021). 'X is the odd one out, because the other two are more about the farmland'—Dutch L1 student teachers' struggles to reason about grammar in odd one out tasks. *Studies in Educational Evaluation, 70*, 1-13. https://doi.org/10.1016/j.stueduc.2021.101007

Van Rijt, J., Myhill, D., De Maeyer, S., & Coppen, P.-A. (2022). Linguistic metaconcepts can improve grammatical understanding in L1 education: evidence from a Dutch quasi-experimental study. *PLOS ONE 17*(2), 1-25. https://doi.org/10.1371/journal.pone.0263123

Van Rijt, J., Banga, A., & Goudbeek, M. (2023). Getting a load of linguistic reasoning: how L1 student teachers process rules of thumb and linguistic manipulations in discussions about grammar. *Applied Linguistics*, 1-26. https://doi.org/10.1093/applin/amad011

Watson, A. (2012). Navigating 'the pit of doom': Affective responses to teaching 'grammar'. *English in Education, 46*(1), 22-37. https://doi.org/10.1111/j.1754-8845.2011.01113.x

Weijun Liang, T., & Fung, D. (2021). Fostering critical thinking in English-as-a-second-language classrooms: Challenges and opportunities. *Thinking Skills and Creativity, 39*, 1-12. https://doi.org/10.1016/j.tsc.2020.100769

Wijnands, A., Van Rijt, J., & Coppen, P.-A. (2021). Learning to think about language step by step: A pedagogical template for the development of cognitive and reflective thinking skills in L1 grammar education. *Language Awareness 30*(4), 317-335. https://doi.org/10.1080/09658416.2021.1871911

Wijnands, A., Van Rijt, J., & Coppen, P.-A. (2022). Measuring epistemic beliefs about grammar. *L1-Educational Studies in Language and Literature, 22*(1), 1–29. https://doi.org/10.21248/l1esll.2022.22.1.362

APPENDIX A

Case A: Afgelopen winter hebben we voor het eerst in tijden weer de 400 meter geschaatst. (‘Last winter, we ice-skated the 400 meters again for the first time in ages’). Teacher A argues that *de 400 meter* is an adverbial; teacher B argues that it is a direct object. Which arguments are in favor of which position?

Mean % correct for all 40 arguments: 57.04, SD = 18.83, *N* = 299 student teachers. The arguments appear in abbreviated form in the Tables below (A1-A4); full arguments can be found in Table A5.

Table A1. Argument set 1 Case A (Rules of thumb based)

Argument	% pro A	% pro B	% irrelevant	% correct	% incorrect
1	**59.87**	12.37	27.76	59.87	40.13
2	**61.87**	22.41	15.72	61.87	38.13
3	**66.56**	10.70	22.74	66.56	33.44
4	**57.19**	19.73	23.08	57.19	42.81
5	7.02	**88.96**	4.01	88.96	11.04
6	17.39	**48.49**	34.11	48.49	51.51
7	11.71	39.80	**48.49**	48.49	51.51
8	11.71	28.09	**60.20**	60.20	39.80
9	20.74	9.36	**69.90**	69.90	30.10
10	7.36	64.21	**28.43**	28.43	71.57

Note. Dark green markings: % correct >= + 1SD above the mean; light green markings: % correct between mean- + 1SD, yellow: % correct = between -1SD below the mean and the mean; red: % correct >= -1SD below the mean. Bold face = correct categorizations

Table A2. Argument set 2 Case A (Linguistic manipulations based)

Argument	% pro A	% pro B	% irrelevant	% correct	% incorrect
1	**60.54**	17.06	22.41	60.54	39.46
2	**57.19**	12.37	30.43	57.19	42.81
3	**46.49**	30.10	23.41	46.49	53.51
4	7.02	**84.95**	8.03	84.95	15.05
5	29.77	**44.48**	25.75	44.48	55.52
6	17.39	**40.13**	42.47	40.13	59.87
7	14.05	**64.21**	21.74	64.21	35.79
8	27.09	37.46	**35.45**	35.45	64.55
9	31.77	13.71	**54.52**	54.52	45.48
10	12.04	10.03	**77.93**	77.93	22.07

Note Dark green markings: % correct >= + 1SD above the mean; light green markings: % correct between mean- + 1SD, yellow: % correct = between -1SD below the mean and the mean; red: % correct >= -1SD below the mean. Bold face = correct categorizations

Case B: Arie is na het puzzelen met plezier aan het koken. (‘Arie is happy to cook after puzzling’). Teacher A argues that *aan het koken* is a verbal predicate; teacher B argues that it is a subject complement. Which arguments are in favor of which position?

Table A3. Argument set 1 case B (Rules of thumb based)

Argument	% pro A	% pro B	% irrelevant	% correct	% incorrect
1	**80.60**	11.71	7.69	80.60	19.40
2	**90.64**	6.35	3.01	90.64	9.36
3	**76.92**	12.71	10.37	76.92	23.08
4	13.04	**69.90**	17.06	69.90	30.10
5	12.04	**35.45**	52.51	35.45	64.55
6	13.04	**77.26**	9.70	77.26	22.74
7	27.42	32.11	**40.47**	40.47	59.53
8	14.05	13.71	**72.24**	72.24	27.76
9	18.39	20.07	**61.54**	61.54	38.46
10	20.40	5.69	**73.91**	73.91	26.09

Note Dark green markings: % correct >= + 1SD above the mean; light green markings: % correct between mean- + 1SD, yellow: % correct = between -1SD below the mean and the mean; red: % correct >= -1SD below the mean. Bold face = correct categorizations

Table A4. Argument set 2 case B (Linguistic manipulations based)

Argument	% pro A	% pro B	% irrelevant	% correct	% incorrect
1	**73.58**	6.02	20.40	73.58	26.42
2	**62.54**	18.39	19.06	62.54	37.46
3	**39.46**	22.07	38.46	39.46	60.54
4	30.43	**23.41**	46.15	23.41	76.59
5	15.72	**70.23**	14.05	70.23	29.77
6	38.46	**28.09**	33.44	28.09	71.91
7	24.08	25.08	**50.84**	50.84	49.16
8	51.17	26.76	**22.07**	22.07	49.16
9	46.49	32.44	**21.07**	21.07	78.93
10	23.08	21.40	**55.52**	55.52	44.48

Note Dark green markings: % correct >= + 1SD above the mean; light green markings: % correct between mean- + 1SD, yellow: % correct = between -1SD below the mean and the mean; red: % correct >= -1SD below the mean. Bold face = correct categorizations

Table A5. Table A5: Full arguments per case (A/B) divided over Rules of Thumb (A1, B1) and Linguistic Manipulations (A2, B2), in Dutch and in English translation

Case	Arg. number	Argument in Dutch	English translation
A1	1	Als je de zin ontleed hebt tot en met het voorzetselvoorwerp, is *de 400 meter* een van de zinsdelen die overblijven.	If you have parsed the sentence as far as the prepositional object, *de 400 meter* is one of the parts that are left over.
A1	2	Als je de vraag stelt 'Hoe lang / hoe ver hebben we geschaatst?' levert dat als antwoord *de 400 meter* op.	If you ask the audit question 'How long / how far have we ice-skated?' the answer is *de 400 meter*.
A1	3	Je kunt '*de 400 meter'* zonder enig probleem weglaten in deze zin.	You can leave out *de 400 meter* without any trouble in this sentence.
A1	4	*De 400 meter* geeft in deze zin een nadere omschrijving van het werkwoordelijk gezegde doordat het de afstand van het schaatsen aangeeft.	*De 400 meter* provides a description of the verbal predicate because it denotes the distance of ice-skating.
A1	5	Als je de vraag stelt 'Wat hebben we geschaatst?' dan levert dat als antwoord '*de 400 meter'* op.	If you ask the audit question 'What have we ice-skated?', the answer is *de 400 meter*.
A1	6	*De 400 meter* is een woordgroep met een zelfstandig naamwoord als kern.	*De 400 meter* is a phrase with a noun at its core.
A1	7	Je kunt nog minimaal een bijwoordelijke bepaling toevoegen aan deze zin zonder dat je een ongrammaticale zin krijgt.	You can at least add another adverbial to this sentence without it becoming ungrammatical.
A1	8	Het zinsdeel waar het om gaat, zou eigenlijk moeten zijn *weer de 400 meter* in plaats van *de 400 meter*.	The relevant phrase should actually be *weer de 400 meter* instead of *de 400 meter*.
A1	9	Als je de zin in de verleden tijd zet, wordt duidelijk dat 'hebben' de persoonsvorm is, terwijl de rest van het werkwoordelijk gezegde niet van vorm verandert.	If you put this sentence in the past tense, it becomes clear that '*hebben*' is the finite verb, while the rest of the verbal predicate remains unchanged.
A1	10	*400 meter* is een zinsdeel waarmee een hoeveelheid uitgedrukt wordt.	*400 meter* is a phrase which denotes a quantity.
A2	1	Het is 'iemand schaatst', niet 'iemand schaatst iets'.	You can say 'someone ice-skates', but not 'someone ice-skates something'.
A2	2	Als je het zinsdeel *de 400 meter* wilt vervangen, dan kun je het vervangen door het zinsdeel *400 meter ver.*	If you want to replace the phrase *de 400 meter*, you can replace it by the phrase *400 meter ver*.
A2	3	Je kunt de zin herformuleren tot 'We hebben afgelopen winter voor het eerst sinds tijden weer geschaatst, en wel de 400 meter.'	You can paraphrase the sentence to ''*We hebben afgelopen winter voor het eerst sinds tijden weer geschaatst, en wel de 400 meter."*
A2	4	Als je deze zin in de lijdende vorm zet, krijg je zoiets als 'De 400 meter wordt door ons geschaatst'.	If you passivize the sentence, the result is *'De 400 meter wordt door ons geschaatst'*

A2	5	Als je van het werkwoord *schaatsen* een zelfstandig naamwoord maakt, krijg je ‘het schaatsen VAN de 400 meter’.	If you turn the verb *to ice-skate* into a noun, you get ‘the ice-skating of the 400 meters’.
A2	6	Zinnen met sportwerkwoorden bevatten vaker een voorwerp in de context van wedstrijden (iets voetballen, iets zwemmen).	Sentences with verbs about sports sometimes contain objects in the context of matches (*iets voetballen, iets zwemmen).*
A2	7	Het lidwoord ‘de’ wijst erop dat het hier niet om een normale afstand gaat maar om een voorwerp.	The article *de* indicates that we are not dealing with a normal distance, but with an object.
A2	8	Als je de zinsdelen in een andere volgorde zet, kun je het zinsdeel *de 400 meter* op de eerste zinsplaats zetten.	If you put the phrases in a different order, *de 400 meter* can be placed sentence-first.
A2	9	Uit de eenzinsdeelproef blijkt dat het relevante zinsdeel ‘weer de 400 meter’ moet zijn, en niet ‘de 400 meter’.	The topicalization test shows that the relevant phrase is ‘*weer de 400 meter*’ and not *de 400 meter*.
A2	10	‘De 400 meter schaatsen’ is een handeling die al ten einde is, dus de persoonsvorm moet eigenlijk *hadden* zijn.	‘Ice-skating the 400 meters’ is an act that has already ended, so the finite verb should actually be ‘had’.
B1	1	*Is* is de persoonsvorm en *aan het koken* bevat het andere werkwoord uit de zin.	*Is* is the finite verb and *aan het koken* contains the other verb in this sentence.
B1	2	Met het werkwoord *koken* wordt in dit geval een handeling aangeduid.	In this case, the verb *koken* denotes an action.
B1	3	In de zin in kwestie staat op het niveau van het gezegde een activiteit centraal.	In the sentence in question an activity is central at the level of the phrase.
B1	4	*Aan het koken* duidt in deze context de toestand van Arie aan.	*Aan het koken* in this context denotes Arie’s condition.
B1	5	Als je kijkt naar het lidwoord dat ervoor staat, moet *koken* wel een zelfstandig naamwoord zijn.	If you look at the article in front of it, *koken* must be a noun.
B1	6	Er staat *is* in de zin, wat een vorm is van *zijn* en *zijn* staat in het rijtje koppelwerkwoorden: *zijn, worden, blijven, blijken, lijken, schijnen, heten, dunken* en *voorkomen*.	The sentence contains *is*, which is a form of *zijn* and *zijn* is in the series of the linking verbs: *zijn, worden, blijven, blijken, lijken, schijnen, heten, dunken* en *voorkomen*.
B1	7	Als je de zin in de verleden tijd zet, verandert het werkwoord *koken* in dat geval niet mee.	If you put this sentence in the past tense, the verb *koken* does not change.
B1	8	De woorden *na*, *met* en *aan* zijn in deze zin alle drie voorzetsels.	The words *na*, *met* and *aan* are all three prepostions in this sentence.
B1	9	In dit voorbeeld is *het* in *aan het koken* geen lidwoord, maar een onbepaald voornaamwoord.	In this example, *het* in *aan het koken* is not an article but an indefinite pronoun.
B1	10	Het werkwoord *koken* staat in deze voorbeeldzin in het meervoud.	The verb *koken* is plural in this sentence.

B2	1	Als je de persoonsvorm weghaalt, komt er op die plek het werkwoord *kookt* te staan.	If you remove the finite verb, the verb *kookt* will appear in that place.
B2	2	Je kunt een lijdend voorwerp toevoegen: *Arie is na het puzzelen met plezier pasta aan het koken.*	You can add a direct object: *Arie is net het puzzelen met plezier pasta aan het koken*.
B2	3	Het is *iemand kookt iets*, niet *iemand kookt*.	It must be *iemand kookt iets*, not *iemand kookt*.
B2	4	Als je er een bijzin van maakt, krijg je … *dat Arie aan het koken is* en niet … *dat Arie is aan het koken*.	If you turn it into a subordinate clause, you get … *dat Arie aan het koken is* and not … *dat Arie is aan het koken*.
B2	5	Je kunt *is* vervangen door andere koppelwerkwoorden, bijvoorbeeld *blijft, lijkt* en *raakt.*	You can replace *is* with other linking verbs, for example *blijft*, *lijkt* and *raakt*.
B2	6	Je kunt *aan het koken* vervangen door het tegenwoordig deelwoord *kokende*.	You can replace *aan het koken* by the present participle *kokende*.
B2	7	Je kunt *aan het koken* op de eerste zinsplaats zetten, waardoor je krijgt: *aan het koken is Arie na het puzzelen met plezier*.	You can put *aan het koken* sentence-first, which gives you: *aan het koken is Arie na het puzzelen met plezier*.
B2	8	Je kunt een voltooid deelwoord toevoegen, waarmee er zoiets ontstaat als: *Arie is aan het koken geslagen*.	You can add a past participle to create something like: *Arie is aan het koken geslagen*.
B2	9	Je kunt het hulpwerkwoord *is* in deze zin niet door *wordt* vervangen.	You cannot replace the auxiliary verb *is* by *wordt* in this sentence.
B2	10	In deze zin is *met plezier* te beschrijven als ‘terwijl hij plezier had’.	In this sentence, *met plezier* can be described as ‘terwijl hij plezier had’.