Informal Logic

Bad Arguments and Objectively Bad Arguments
Mauvais arguments et arguments objectivement mauvais

Michael H.G. Hoffmann et Richard Catrambone

Volume 43, numéro 1, 2023

URI : https://id.erudit.org/iderudit/1099208ar
DOI : https://doi.org/10.22329/il.v43i1.7076

Aller au sommaire du numéro

Éditeur(s)

Informal Logic

ISSN

0824-2577 (imprimé)
2293-734X (numérique)

Découvrir la revue

Citer cet article

Hoffmann, M. & Catrambone, R. (2023). Bad Arguments and Objectively Bad Arguments. *Informal Logic*, *43*(1), 23–90. https://doi.org/10.22329/il.v43i1.7076

Résumé de l'article

Beaucoup ont fait valoir qu'il est impossible de déterminer des critères pour identifier les bons arguments. Dans cette contribution, nous soutenons qu'il est au moins possible d'identifier les caractéristiques des arguments objectivement mauvais. Allant au-delà des critères APS de Blair et Johnson qui stipulent que les raisons doivent être acceptables, pertinentes et suffisantes, nous développons une liste de huit critères avec des instructions sur la façon de les appliquer dans l'évaluation des arguments. Nous concluons en présentant les données de deux études empiriques qui montrent à quelle fréquence les étudiants violent ces critères dans des conditions de laboratoire et « dans la nature ».

Bad Arguments and Objectively Bad Arguments

MICHAEL H. G. HOFFMANN

School of Public Policy
Georgia Institute of Technology
685 Cherry Street, N.W.
Atlanta, GA 30332-0345
USA
m.hoffmann@gatech.edu

RICHARD CATRAMBONE

School of Psychology
Georgia Institute of Technology
654 Cherry Street, N.W.
Atlanta, GA 30332-0170
USA
rc7@gatech.edu

Abstract: Many have argued that it is impossible to determine criteria to identify good arguments. In this contribution, we argue that it is at least possible to identify features of objectively bad arguments. Going beyond Blair and Johnson's ARS criteria, which state that reasons must be acceptable, relevant, and sufficient, we develop a list of eight criteria with instructions for how to apply them to assess arguments. We conclude by presenting data from two empirical studies that show how frequently students violate these criteria in lab conditions and "in the wild."

Résumé: Beaucoup ont fait valoir qu'il est impossible de déterminer des critères pour identifier les bons arguments. Dans cette contribution, nous soutenons qu'il est au moins possible d'identifier les caractéristiques des arguments objectivement mauvais. Allant au-delà des critères APS de Blair et Johnson qui stipulent que les raisons doivent être acceptables, pertinentes et suffisantes, nous développons une liste de huit critères avec des instructions sur la façon de les appliquer dans l'évaluation des arguments. Nous concluons en présentant les données de deux études empiriques qui montrent à quelle fréquence les étudiants violent ces critères dans des conditions de laboratoire et « dans la nature ».

Keywords: argument appraisal, argument assessment, argument evaluation, argument mapping, argument quality, ARS criteria, bad arguments, critical thinking, good arguments, RSA criteria

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

1. Introduction

The ability to assess the quality of arguments is crucial for scientific reasoning, for deliberation in public and private spaces, and for critical thinking in general. People need to know what a good argument is, how to distinguish a good one from a bad one, and how to identify weaknesses in their own arguments so that they can improve their reasoning. Determining criteria that can be used for assessing the quality of arguments is far from trivial. The quality of many arguments that we encounter in academic, public, and private settings cannot be determined objectively based on the fact that there is no common ground, that is, no agreement on the assumptions upon which these arguments are based (Feldman 1994; Hoffmann 2018). There may be conflicting values or background knowledge that is not shared, which makes those assumptions controversial. Or it may just be the case that we are content with the quality of an argument on one day but see something missing on another. When it comes to problems like these, it is probably best to rely on dialogic procedures to assess the quality of an argument, that is, to rely on debates about these assumptions and perceptions.

However, there are situations in which we do not have the luxury of deliberation to determine the quality of arguments—for example, when we grade the work of students or when we consider our own arguments under time constraints. Moreover, when we teach the basics of argument assessment, we should provide a clear set of quality criteria that are useful for learners. We should be clear about which criteria are essential to assess the quality of arguments, and we should make sure that students acquire such a degree of familiarity with these criteria that using them in critical analyses of their own reasoning and that of others becomes second nature. This goal leads to the question that motivates the present contribution: What could be a set of core criteria that can be used to assess particular features of any argument objectively?

This question, though, requires a few clarifications. First, our reference to 'any argument' is limited by a particular definition of 'argument.' The focus of this contribution is on argument as a set of propositions composed of reason(s) and a conclusion where the reason is supposed to justify the conclusion. (What we call 'rea-

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

son' can itself be a set of premises.) There are, of course, other definitions of 'argument.' Some put "convincing a reasonable critic of the acceptability of a standpoint" at the center (van Eemeren and Grootendorst 2004, p. 1), others a multi-party argumentative discussion (Lewinski and Aakhus 2014). An even wider variety of conceptualizations becomes visible when we look at non-Western cultures as was done in a recent special issue of *Argumentation* on "Argumentation Through Languages and Cultures" (2021, vol. 35; see Plantin 2021).

All of this needs to be excluded here because the question of how to assess the quality of arguments depends, of course, on the notion of argument used, and discussing quality criteria for every such notion is not possible given the limited space available for the present contribution. The present contribution focuses on arguments in the sense of a set of propositions composed of reason(s) and a conclusion.[1] Thus, 'argument' refers here to a certain *product*, an artifact, be it a text, a graphical argument map, or a recording; something that is completed and can be studied in the form of a representation. This means that we will neither talk about the quality of arguments or argumentations in the sense of processes that can be good or bad or that can improve the quality of arguments in the sense used here,[2] nor about certain characteristics of the person arguing or virtuous consequences of arguments, as has been discussed in the 'virtue theoretic approach to argument' (Aberdein 2010).

The second necessary clarification refers to our aim of 'objective' assessment. We are using 'objective' here to indicate that the focus of this contribution is on only assessment criteria whose interpretation and application does not depend on conditions that might vary among educated people or that might change over time. So, 'objective' should be read as a short expression of 'inter- and intrasubjectively stable over time,' where the latter refers to stabil-

[1] For more on how argument in this narrow sense can be defined more precisely see Hitchcock (2007); Goddu (2009a, 2018); and Goodman (2018).

[2] Thus, there will neither be discussions of dialogic or dialectical approaches, such as the pragma-dialectical approach, nor reflections on Bayesian approaches that focus on "updating" quality assessments based on new information, such as recently described by Godden and Zenker (2018).

ity within the same subject (the person doing the assessment) over time.

The first challenge is to show whether those objective criteria are possible at all. An example can demonstrate the possibility of at least one of these criteria. It refers to the question of whether the reasons provided in an argument are sufficient to justify the conclusion. Let us take the conclusion of Nicholas Stern's (2006) argument for strong climate policies that is depicted in the form of an argument map in Figure 1: "The benefits of strong and early action to avoid the worst impacts of climate change far outweigh the economic costs of not acting" (p. vi). To justify this claim, what is provided in the premises must be rich enough to address two points: first, the net benefits (i.e., economic benefits minus costs) of doing nothing about climate change and, second, the net benefits of "strong and early" climate action. If one of these points is missing in the set of premises provided, then these cannot be sufficient to justify the conclusion. If you look at Figure 1, the argument would be objectively incomplete if either premise 2.1 or 2.2 were not there. The question of whether both of these points are addressed in a particular argument can be answered "objectively" in the sense that everybody who is able to understand what is claimed in the conclusion and who is not cognitively incapacitated in the moment of assessment will come to the same answer. The answer is not dependent on any particular values, beliefs, attitudes, or background knowledge of the assessor. Any rational agent would provide the same answer, and the answer would not change over time.

As will be discussed in more detail in the fifth section of this paper, in the empirical part of our study, we measured the degree of agreement between two trained coders of argument quality. Intercoder reliability ranged from 'good' to 'very good' for the quality of criteria discussed here.

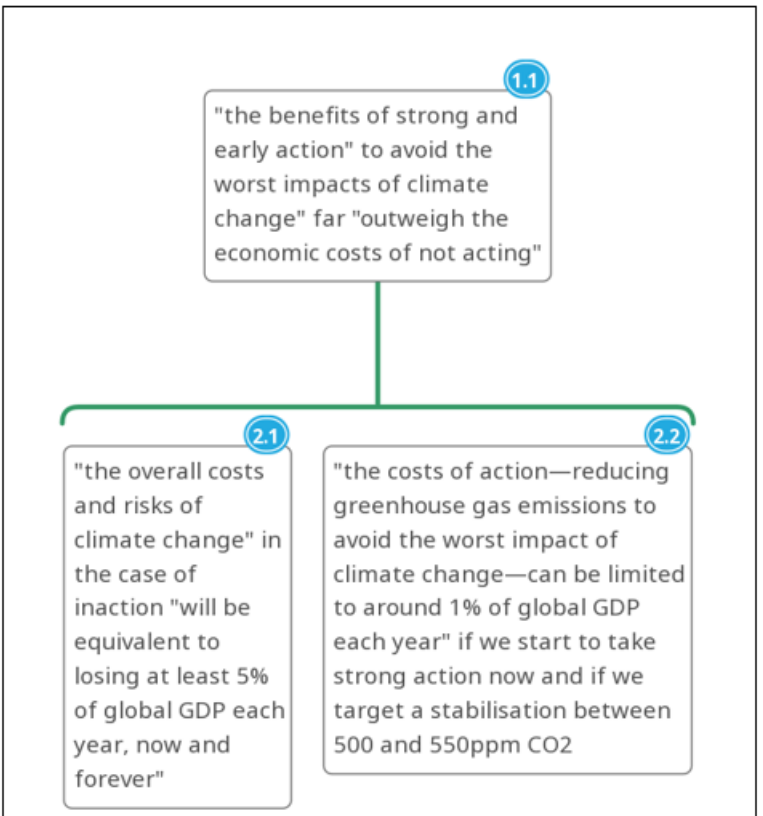


Figure 1: Stern's main argument (Stern 2006, p. vi). Note that lower costs in Premise 2.2 compared to the costs mentioned in 2.1 is logically equivalent to the 'benefits' mentioned in the conclusion. The entire report (Stern 2007) can be reconstructed as an argumentation that aims at justifying premises 2.1 and 2.2 (Hoffmann 2010). All argument maps in this contribution were created with MindMup (https://www.mindmup.com)

Limiting the assessment criteria considered here to those that can be applied "objectively" is intended to exclude two sorts of non-objective criteria. The first group contains empirical criteria, such as the persuasiveness of an argument. As Richard Feldman argued with regard to persuasiveness, the question of whether people can be convinced by an argument or not depends to a certain degree on how stubborn or gullible these people are (1994, p. 168). Variance

in relevant characteristics of an intended audience excludes or limits the possibility of objective assessment according to a persuasiveness criterion.

The second group of non-objective assessment criteria includes those about which there can be reasonable disagreement based on the fact that judging an argument as good or bad depends on conflicting values and background assumptions. For example, the main argument that Nicholas Stern (2006) developed in *The Economics of Climate Change* (Figure 1) has been criticized for diverting "attention away from alternative approaches, away from ethical debates over harming the innocent, the poor and future generations, and away from the fundamental changes needed to tackle the very real and serious problems current economic systems pose for environmental systems" (Spash 2007, p. 704). This controversy, you could say, is about the question of whether climate policies should be justified based on utilitarian arguments (cost-benefit arguments in the language of economics and public policy) or on deontological arguments. This could be considered a question of value, and since people live according to different values, there is no objective way to decide whether a utilitarian argument, like the one in Figure 1, is a bad argument solely due to that fact that it is utilitarian.

Other controversies about the Stern report concern certain background assumptions that his team made to calculate future costs and benefits of the 'business as usual' approach in contrast to 'strong and early action.' For example, the team decided to calculate these based on a discount rate of 1.4%. A discount rate determines, roughly, how much value we assign today to costs and benefits that will occur in the future, over generations to come, given the assumption that these generations will be richer than we are today.[3] The discount rate used in the Stern report has been challenged by other authors (Nordhaus 2007; Spash 2007; Baer and Spash 2010). But again, there can be reasonable, scientific debate about this question, so we should not condemn an argument that is based on a particular discount rate as objectively bad. The assessment of arguments that are based on values or background

[3] A much more detailed explanation can be found in Roberts (2012).

assumptions about which there can be reasonable disagreement are better left to dialogic approaches. A determination of the quality of these arguments is best done by the particular scientific community of relevance or by the entire society.

Even though these examples for objectivity on one side and non-objectivity on the other should not be controversial, it is important to note that they leave a large gray area in between. In our empirical study, we found higher degrees of intercoder reliability for some quality criteria than for others. Although agreement between the coders was generally high, the measurement was done under controlled lab conditions after substantial training.

In the next section, we situate our contribution in the broader research context of argument assessment. Section three describes our methodology, and section four introduces and discusses each of the eight criteria we are proposing. Section five will present some of our empirical work. Figure 16 in the conclusion provides a summary of the suggested assessment procedure.

2. Various approaches to assess the quality of arguments

Our approach to develop objective criteria for the assessment of arguments can be justified by comparing it with other approaches. The one with the longest history is Aristotle's notion of fallacy. The Greek term, 'paralogism,' indicates what he had in mind. There are forms of reasoning that "only seem to be, but are not really" syllogisms (Aristotle *Soph. el.* 164a24). Attempts, though, to define bad arguments as those that deviate from the standard of logical validity are nowadays generally rejected. There are two main reasons. First, there are many good arguments that are not valid (strong inductive arguments, for instance). Second, many valid arguments are epistemically useless; they fail, as Richard Feldman puts it, "to have rational merit" (1994, pp. 161, 165). Fabio Paglieri (2015) introduced the fitting term 'balidity' to capture the latter: valid but bad. "Both Pierre and Marie Curie were physicists. Therefore, Marie Curie was a physicist" is an example of a perfectly valid but utterly useless argument that Paglieri (2015) quotes from Cohen (2013).

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

So, using only validity as a criterion to distinguish good arguments from bad ones forces us to classify many arguments as bad which are good and others as good which are bad. A similar critique can be applied to the claim that there is a clear distinction between fallacious arguments and good ones. Starting from a non-Aristotelian understanding of 'fallacy' that is nowadays more widely shared, Christopher Tindale defines 'fallacy' as "a particular kind of egregious error, one that seriously undermines the power of reason in an argument by diverting it or screening it in some way" (2007, pp. 1-2). However, as he shows in a summary of the literature, we cannot simply ostracize all arguments that can be subsumed under one of the well-known fallacies. As he says,

> [Many] of the fallacies are failed instances of good argument schemes or forms. Hence, we cannot dismiss all *ad hominem* arguments or slippery slopes, for example, because there are circumstances under which such reasoning is appropriate. What is required, then, is a careful review of the differences between good and bad instances of such schemes (Tindale 2007, p. xiv)[4]

For a more careful analysis, Tindale proposes the use of 'critical questions' that can be determined for each pattern of fallacious reasoning. A corresponding method has also been developed in the literature on argumentation schemes. In what is currently the most comprehensive compendium of these schemes, Douglas Walton, Chris Reed, and Fabricio Macagno define argumentation schemes as "forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation" (2008, p. 1). In their work, each scheme comes with a particular set of critical questions. For example, the scheme 'argument from position to know' is presented as follows:

[4] Boudry, Paglieri and Pigliucci (2015) have argued that this approach might not be sufficient.

(1) *Major Premise*: Source *a* is in a position to know about things in a certain subject domain *S* containing proposition *A*.
Minor Premise: *a* asserts that proposition *A* is true (false).
Conclusion: *A* is true (false).

Critical Questions

CQ1: Is *a* in a position to know whether *A* is true (false)?
CQ2: Is *a* an honest (trustworthy, reliable) source?
CQ3: Did *a* assert that *A* is true (false)? (Walton, Reed and Macagno 2008, p. 309)

These critical questions can be used to assess the quality of all arguments that can be characterized as 'argument from position to know.' In general, argument assessment can, thus, be realized in three steps:

1. Determine under which argumentation scheme a given argument falls.
2. Use the critical questions that experts formulated for this particular scheme[5] and try to answer them with regard to the given argument.
3. Determine the quality of the argument based on how many of the questions can be answered from the context in which the argument has been developed and the degree to which these answers are satisfying.

This approach, however, faces some serious limitations. Even though Walton, Reed, and Macagno (2008) distinguish sixty argumentation schemes, many with multiple sub-schemes so that the overall number is approximately one hundred, the list is incomplete (Lumer 2011). Missing are ethical argument schemes—for example the distinction between utilitarian and deontological justifications of what should be done—but also structures such as transcendental arguments (Hoffmann 2019; Einstein used those to

[5] For more on how critical questions should be formulated see Baumtrog 2021.

justify crucial assumptions in his theory of relativity). In general, it seems questionable whether there can ever be a complete list of argumentation schemes. Often, scientific disciplines have their own forms of argument, many of which could be formulated in different ways, and it seems possible to divide any particular scheme into more specific forms.

A second limitation of this approach is that there can be reasonable disagreement about the list of critical questions that is assigned to each scheme (Baumtrog 2021). For example, in the argument from position to know (1), we might also ask: Is *a* lying? (Yu and Zenker 2020). A third problem is the existence of cognitive limitations. First, there are so many argumentation schemes that it is difficult to memorize them all and, second, it is not trivial to identify the scheme that fits best to a particular argument.

However, it seems that these three limitations of assessing arguments by means of argumentation schemes can be overcome using an approach that Shiyang Yu and Frank Zenker (2020) recently developed. Even though they focus on argument schemes, their 'meta-level' representation of schemes is so general that an application to all arguments—whether they are realizing a certain scheme or not—is possible. Their approach is important because they claim that it provides a method for "complete argument evaluation" (Yu and Zenker 2020).

Since a discussion of this broad claim requires a more detailed analysis than necessary for present purposes, we put it in the Appendix. As we argue there, Yu and Zenker (2020) fail in their attempt to provide a complete list of critical questions for argument assessment. The reason is that their final list does not cover some of the criteria that we present in this contribution. Another point that we discuss in the Appendix is a general concern about analytical approaches that determine quality criteria based on an analysis of the definition of an argument. Their value is limited if they do not cover what is observable in the practice of constructing arguments. This leads us to a discussion of the methodology that we are using here.

3. Methodology

The proposed list of assessment criteria is designed to satisfy three conditions:

1. Each criterion should be applicable to all arguments in the sense of a reason-conclusion constellation, not only to a subset of arguments such as deductively valid arguments or arguments that can be subsumed under a specific argument scheme or a fallacy.
2. The list of assessment criteria should be cognitively manageable in the assessment practice. This means that instead of attempting to compile the most comprehensive list, it would be better to create a list of criteria that are most often violated by arguers or that are considered to be most important by the community of experts. Figuring out what should be listed is, thus, an empirical question and a question of deliberation among experts.
3. The criteria should be specified to a degree of precision that permits an acceptable level of interrater reliability as discussed above regarding the notion of objectivity.

The foundation for our approach has been laid with the well-known "ARS criteria" formulated by Ralph Johnson and Anthony Blair (2006/1977), which focus on the idea that premises that support the conclusion of an argument should be *acceptable*, *relevant*, and *sufficient*. These three criteria—which are also known as the RSA-conditions for good argument—are widely considered to fulfil the first two conditions listed above.[6]

Johnson and Blair's (2006/1977) ARS criteria are very attractive when it comes to cognitive overload: just three criteria that can be applied to every possible argument. As the authors write in the preface to the textbook's 2nd edition:

[6] As Dove and Nussbaum 2018 showed, the ARS criteria are equivalent to the ARG conditions that Govier (2010/1985) developed. While the 'A' and 'R' refer in both accounts to acceptability and relevance, Govier's 'G' stands for 'grounding': Is the conclusion grounded by the premises? This is equivalent to Johnson and Blair's 'sufficiency.'

> These criteria have the advantage of including deductive validity and inductive strength as special cases, while at the same time they leave open the possibility that there are other legitimate kinds of inference in arguments besides valid deductions and strong inductions" (Johnson and Blair 2006/1977, p. xiii).

This is indeed important because it points to the possibility that the ARS approach can be used not only as an umbrella under which logical validity and the strength of justification can be discussed, as Johnson and Blair suggest, but also what we know about fallacies and argumentation schemes. Many fallacies can be distinguished as *fallacies of acceptability* (such as false dichotomy, which uses an alternative as a premise whose acceptability can be attacked by showing that there is a third option), *fallacies of relevance* (such as attacking a straw person and *ad hominem*), and *fallacies of sufficiency* (such as hasty generalization). Similarly, it seems to be possible to divide all the critical questions that Walton, Reed, and Macagno (2008) assign to specific argumentation schemes into those that question either the acceptability or the sufficiency of premises (irrelevant premises do not seem to play a role in argumentation schemes).

Our contribution proposes one major innovation. Going beyond the ARS criteria introduced by Johnson and Blair (2006/1977), it proposes a list of eight basic criteria that, we will argue, should be taken into account.

Before we go into the details, we would like to make a few points in advance. First, this contribution is limited to the assessment of individual *arguments*, that is, connections of reason and conclusion. Of course, it is possible that a premise of such an argument might itself be justified by another argument or that the same conclusion is justified by more than one independent argument (for more details see section 4.7 below). In the present context, we call any combination of independent arguments—whether they support one of the premises of an argument or the same conclusion independently—an *argumentation*. Such an argumentation can include objections to certain components of arguments as well as possible counterarguments against those objections and so on. To assess the quality of an argumentation would require more than can be provided in the current contribution.

Second, before an argument can be evaluated it needs to be identified as such (Ennis 2001, p. 97; Blair 2019). The argumentative material that we encounter in the real world is often far removed from the clear reason–conclusion standard that is provided by our definition of argument. What we encounter in the real world is often a mess of "ill-organized, incompletely stated, wandering-off-topic arguments" (Johnson 2000, p. 128). But even if we try to reconstruct such high-quality argumentation as we might find in excellent journalism or scientific publications, experience shows that it is often possible to develop various and significantly different representations of its structure if this structure is a bit more complex. Often, it is not easy to identify what exactly the conclusion of such a piece might be. Visual arguments present another challenge. As Leo Groarke points out, what we need for those is "a standard method that can be used in preparing any argument for assessment" (Groarke 2019, p. 351); a method that permits identifying "its premises and conclusions and depicting its structure in the form of a diagram" (p. 342).

For these reasons, we will work only with arguments whose reason and conclusion are already clearly distinguished. Most of the arguments that are used here as examples are presented in the form of an 'argument map.' For their construction, the computer-supported argument visualization tool MindMup has been used.[7] Of course, the assessment method described here can be applied to any presentation of an argument in which reason and conclusion are clearly identifiable.

Finally, a few words about the background of this project. The first author taught argument mapping for more than fifteen years and assessed thousands of argument maps created by students. In later years, standardized training of argument mapping preceded work on so-called wicked problems which students performed in small teams. The teams grappled with questions such as whether and how micro-targeting in social media should be regulated, or facial recognition technologies. The teams in these classes—applied philosophy classes for undergraduate students across

[7] See https://www.mindmup.com. MindMup allows both concept mapping and argument visualization. Only the latter has been used here.

campus at a research-oriented university with a focus on engineering—are first tasked with formulating a problem, then performing a stakeholder analysis, and finally developing a proposal on how we, as a society, should deal with the problem. As an important part of this last step, the teams are asked to provide a justification for each component of their proposal in the form of an argument map (for more detail, see Hoffmann 2020a). This means the argument maps that were assessed were not in any way given or alluded to by an instructor; they are arguments "from the wild."

Besides being informed by the literature, the eight criteria to assess the quality of those argument maps that we are going to discuss below were developed over years based on the first author's assessment practice. An important refinement, though, has been achieved in the context of a National Science Foundation (NSF) project in which we assessed the quality of argument maps that were created, under controlled experimental conditions, with two different computer-supported argument visualization tools.[8] In order to determine assessment criteria that could then be used to train coders, we used Task Analysis by Problem Solving (TAPS; Catrambone 2011). In the TAPS procedure, a subject-matter expert (SME; in this case Hoffmann) identifies a set of problems that the SME thinks learners should be able to solve if they conceptually understand the procedure. The SME solves some of the problems from this set while justifying each step to the knowledge extraction expert (KEE; Catrambone) who is a domain novice. The KEE develops detailed notes based on the solution procedures and justifications (the 'why' for each step) provided by the SME. The SME is not invited to provide abstract theory outside of justification for steps. Eventually the KEE uses the notes to solve the problems that the SME already solved; the KEE can request help from the SME to resolve impasses. Throughout this iterative process, the KEE continuously updates and reorganizes the notes; this reorganization allows the KEE to develop solution procedures that are independent of specific examples. Once the KEE can solve all of the old problems, the KEE then attempts to solve new problems provided by the SME; again, the KEE can get help with impasses.

[8] See Award Abstract (2020).

When the KEE can, using the notes, solve all problems given by the SME without the help of the SME, then these notes represent, as a practical matter, a complete task analysis of the procedure.

TAPS allowed us to determine not only a first list of six assessment criteria, but also their sequencing so that they can be used in a decision-tree procedure. The results of TAPS have been used to develop a curriculum to teach the assessment of arguments, and for the creation of an interactive Argument Assessment Tutor that can be used online for free (Hoffmann 2020b). Since 2016, Hoffmann has taught the *construction* and *assessment* of arguments side by side. In the fall of 2018, he graded all 475 arguments created by student teams in the wicked-problem-context described above, following the sequence of these six criteria. Almost all the criteria were frequently used to identify bad arguments. (How often each has been used will be shown in section 5.) In the process of doing these assessments, he discovered that two criteria should be added, so that the final list of assessment criteria grew to eight. The eight criteria cover all observed problems in the arguments.

This way, the list of eight assessment criteria, as well as frequent typologies of cases in which they might be manifested, is empirically grounded. There is no theoretical foundation from which they are derived. Even though this approach is obviously limited by the fact that the assessment of these 475 arguments has been performed by just one expert, it should be significant that to this assessor, there was no point at which the need for further criteria was made salient by the data.

In the next section, we consider the list of eight criteria that can be used to assess the quality of arguments with various degrees of objectivity. Most of the arguments that we are going to present as examples were created by students in classes taught by Hoffmann. Consent to use them for publication was granted in the context of an IRB-approved research protocol.

4. Eight criteria to assess the quality of any argument

The following criteria are put in a particular sequence. The justification for this particular sequence has both logical and practical

components. We propose that one must start by asking whether the argument's conclusion is formulated appropriately. If the conclusion is not formulated clearly enough or has other problems, the assessment can stop right there. It would be a waste of time trying to analyze reasons for a conclusion that in itself is not formulated clearly enough. If we do not know what exactly an argument is supposed to justify, there is no way that we could say anything about the quality of the argument.

Assuming that the argument's conclusion is formulated clearly enough, the next criterion on which to focus is the quality of each premise's formulation. As with the conclusion, it is impossible to assess the relevance or acceptability of a premise if it is not formulated clearly enough.

To determine the order of applying the ARS criteria—acceptability, relevance, sufficiency—to the reasons, consider the following: Assessing sufficiency is more demanding than the assessment of acceptability because sufficiency might be established by combining a set of premises whereas acceptability always refers to a singular premise. Since a premise cannot help to justify a conclusion if the premise is false, we can, thus, simplify the assessment of sufficiency by removing unacceptable premises from further consideration. Therefore, we suggest assessing the acceptability of all premises as the third step after assessing the quality of the formulation of the conclusion and then all premises.
Regarding relevance, the following should be considered. The fact that a premise is irrelevant implies that it cannot contribute to the sufficiency of all premises. This logical argument suggests that it would make sense to assess relevance before sufficiency, just as it makes sense to assess acceptability before sufficiency. However, in contrast to acceptability, relevance can be established by combining various premises in one reason, as we will discuss in section 4.7. This means that before we assess relevance and sufficiency, we should get a better understanding of the structure of the argument in question.

In order to understand the structure of an argument, one initially might ask whether each significant component of a complex conclusion is addressed by at least one reason. Answering this question helps us to consider the overall structure of the argument:

which premises justify what in a more complex conclusion? This fourth step of the suggested assessment procedures fulfills, thus, two functions: first, we can identify as objectively bad all those arguments in which one or more of the conclusion's components are not justified by any reason; and second, we get a better understanding of the argument's structure. With this we can then assess the relevance of the reasons provided and, finally, their sufficiency. Relevance should be assessed before sufficiency for the logical reason mentioned above.

However, things become more complicated if we take the possibility into account that the argument is poorly structured. Up to now, we considered the argument's structure only with regard to the conclusion's components. In all this, we took the structure as presented by the arguer. But the arguer might have chosen an inappropriate structure. This needs to be assessed in the seventh step of our assessment procedure. The complication that we must discuss in this context results from the fact that there is, as will be discussed in section 4.7, a mutual dependency between assessing relevance and structure.

The last criterion is then the question of whether there are contradictions among the propositions used in an argument. Putting this question at the end is justified by the observation that it is not easy to spot contradictions. We believe the search for contradictions can be done most effectively after all components of an argument are analyzed from the various perspectives that the other criteria provide. As already mentioned, the entire assessment procedure is summarized in Figure 16.

4.1 Is the conclusion formulated appropriately?

If the conclusion of an argument is formulated such that it is not clear *what exactly* should be justified, then the argument does not have any value. Here are types of formulations that cannot be justified because there are problems determining what is claimed in the conclusion:

1. *Questions*. A question cannot be the conclusion of an argument. The only exception to this rule is cases in which a reason for *asking* the question is provided. David

Hitchcock gives the following example from a newspaper headline: 'Your smart phone is making you stupid, antisocial and unhealthy. So why can't you put it down?' (Hitchcock 2020). 'So' indicates an argument, but the only purpose of the three premises is to establish that this question should be answered. A question, in contrast, whose purpose is to obtain a piece of information cannot be the conclusion of an argument. It would probably be best to interpret the example provided by Hitchcock as an abbreviated conclusion. In its full form it would be: 'Therefore, you should answer the question: Why can't you put it down?' In order to decide whether a question is acceptable as a conclusion it is, thus, necessary to take the premises provided into account.

2. *Formulations that do not state anything*. 'Paul is smarter than.' 'Freedom and security.' As formulated, these words do not state anything. The first is an example of an incomplete sentence, and the second is just a sequence of words. Neither can be justified. Every conclusion must either state something or it must be possible to transform it—without adding something that is not there—into a statement. The latter is the case, for example, in commands. A command can be justified because it is equivalent to a normative statement. 'Open the window!' is equivalent to 'You should open the window.' Since the normative statement can be justified, the command can be justified as well. We can say: 'It is hot in here; therefore: Open the window!' There are probably other cases in which the premises of an argument provide information that allows the transformation of a non-statement into a statement without adding something that is not there.
3. *Arguments*. When using argument mapping tools, students frequently put entire arguments into the text box that is supposed to show only the conclusion. 'We should go swimming because it is hot, and all the work is done.' This is an argument. It includes a justification. But as an argument, it cannot itself be justified. It is, of course,

possible to justify the premises provided in this argument ('it is hot'; 'all the work is done'), and it is also possible to justify why we propose this argument, but the argument in itself cannot be justified. If you think it can, try to reconstruct one. The outcome of this thought experiment would be something like: 'The sun shines, therefore, we should go swimming because it is hot and all the work is done.' Premises are now all over the place and it is unclear what to do with them, and how they are related. An argument has a conclusion, but the conclusion cannot itself be an argument.

4. *Inappropriately nested propositions*. 'Dr. Wiseman claimed that dental hygiene is important.' Even though a statement like this one can be justified, the only thing that can be justified here—from a grammatical point of view—is 'Dr. Wiseman claimed (something).' Either she did claim it or she didn't. Reasons for the proposition 'she claimed it' could refer to why she claimed it—for example, because she conducted a study about dental hygiene—or to the fact that she claimed it; a reason for the latter could be a piece of evidence such as a quote from a book she wrote or a tape on which she can be heard saying so. What we have here can be called a *nested proposition*: the claim 'dental hygiene is important' is embedded into the *main proposition* 'Dr. Wiseman claimed that ...' It is possible to provide reasons for this main proposition, but if the reasons provided justify only the *embedded* proposition, then we have a case of inappropriately nested propositions in the conclusion. This way, the question of whether propositions are appropriately or inappropriately nested can be answered only in view of the reasons provided.
5. *Inconsistent or contradictory statements*. 'This car is too expensive, and it is cheap.' This statement is inconsistent, and it is even a logical contradiction because it can be transformed into 'the car is too expensive and it is not too expensive.' A position that is inconsistent can never be justified.

6. *The meaning of important concepts is not clear*. 'Ogliwoopses are extremely dangerous.' In the case that we are not provided with a definition of what ogliwoopses are, we will never be able to assess the quality of arguments that are intended to justify such a conclusion. Key concepts of a conclusion need to be either clear or defined (Glassner 2017, p. 99). Having an unclear key concept in the conclusion is acceptable only if a clarification or definition is provided at another place in the argument or argumentation.
7. *The conclusion, or an important part of it, is so badly formulated that its meaning is incomprehensible or depends clearly on the assessor's interpretation*. Figure 2 provides an example. Note that in this case, the reason does not provide anything that could be used to clarify the meaning of the conclusion. The argument is clearly bad—not only with regard to the formulation of the conclusion. Focusing here only on the conclusion is justified by the consideration that the purpose of an argument—in the sense that we are using the term here—is to justify or support a claim by reasons. The claim is the starting point. But if the meaning of this claim is incomprehensible, then there is no point in assessing the rest of the argument. It has to be noted, though, that it is often hard to decide whether the formulation of a conclusion is clear enough or not. Figure 3 provides an example. Depending on one's interpretation, 'The state takes priority over the individual' can either mean 'The state should determine what the individual does' or 'The survival of the state is more important than the survival of particular individuals.' The formulation is ambiguous. However, if we take the reason into account, it seems plausible to assume that the latter is meant. The problem is that there are certainly cases in which one interpreter might think that the conclusion is clear enough while another one disagrees.

These seven possibilities for arguments with unclear conclusions—and the fact that Hoffmann encountered all of them in the

practice of argument assessment—show that looking at the question of whether the conclusion of an argument is formulated clearly enough is an important criterion for quality assessment. However, it is doubtful whether this criterion allows for a high enough degree of inter- and intra-reliability of assessment results. There might be clear cases, but there might be more in which an objective assessment is out of reach.

4.2 Are the premises formulated clearly enough?

A premise should be criticized if it is not formulated clearly enough. The typology of the seven cases that we distinguished above can be used here as well, even though the only cases that we encountered with regard to the clarity of premises were those in which the premise meaning was incomprehensible in the context of the argument ('there is no competition') or those in which the premise consisted of incomplete statements ('Bacteria cause illnesses that are more difficult to cure'—more difficult than what?). The same limitations regarding the objectivity of assessment mentioned above apply here.

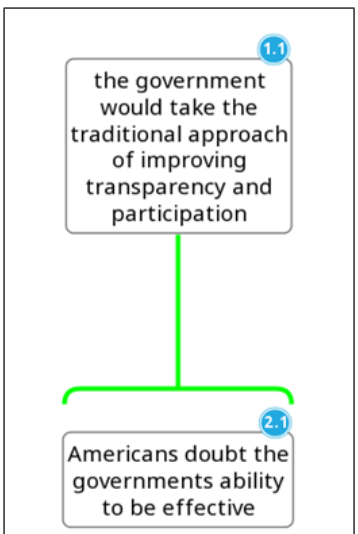


Figure 2: An argument map created by a student

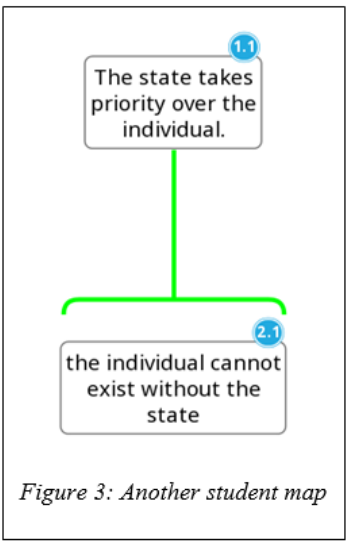


Figure 3: Another student map

4.3 Are the premises obviously unacceptable?

A premise that is obviously false or looks immediately questionable cannot justify a conclusion. It may be the case that a premise that is not immediately convincing is justified by further arguments—which would mean that we need to assess these arguments and not the one in question—but if that is not the case, then there is a serious problem.

For many claims that are used as premises, there can be a legitimate debate about the question of whether they are acceptable. Many claims in science, for example, are controversial, and this is even more prevalent in the political realm. Since we are interested here only in assessment criteria that can be applied more or less objectively, we will not judge premises as unacceptable in cases where this judgment depends on the particular stance of the assessor on a controversial claim. Only the acceptability of those claims should be questioned that seem 'obviously false.' Or, as Leo Groarke and Christopher Tindale characterized the unacceptability of a claim: "The statement conflicts with what is known to be the case such that a reasonable audience (and evaluator) has reason to reject it" (2008, p. 259).

An objective assessment is possible at least in cases where a universal proposition ('all birds fly') can be defeated by a counterexample ('penguins cannot fly'). As soon as we provide the counterexample, nobody would seriously doubt that the premise in which such a universal proposition is used is unacceptable. We encounter those examples frequently. Normative statements are generally universal propositions ('all lying is wrong'), but so are statements such as, 'Each person must be either a theist or an atheist.' This claim is false because the alternative excludes the possibility that someone simply does not care about the existence of God or is agnostic about the question. This example represents the well-known fallacy 'false alternative.'

While there are, thus, examples of cases that allow an objective assessment of premise acceptability, in many other cases the assessment depends on available background knowledge, trust in experts or witnesses, or our own judgment.

4.4 Is each significant component of a complex conclusion addressed by at least one premise?

In the literature, this assessment criterion has been highlighted by scholars who developed argument mapping tools. Tim van Gelder who created ReasonAble and Rationale[9] called it the "Rabbit Rule" because

> you can't pull rabbits out of hats just by magic. If a rabbit appears above the hat, it must have been put in there previously. In argument mapping terms, nothing can magically appear in the contention [i.e., conclusion]; it must have been put in the premises first (The Rabbit Rule n.d.).

More precisely, the rule says "that any significant term or concept which appears in the contention must also appear in one of the premises" (The Rabbit Rule n.d.).

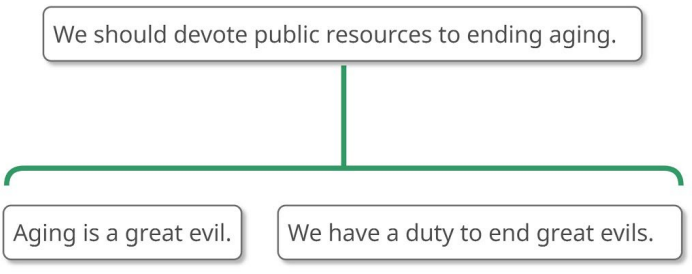


Figure 4: Cullen's example for not addressing significant components of the conclusion by reasons (Cullen 2022).

Simon Cullen, who developed MindMup, formulated the same rule as: "Don't conjure concepts out of thin air" (Cullen 2022). He provides the example in Figure 4 and writes:

> The argument is guilty of conjuring because the conclusion concerns the idea of public resources—an idea which does not appear in either of the claims supporting the conclusion. Therefore, the

[9] Now available at https://www.rationaleonline.com

> conclusion is simply conjured out of thin air. You know that this cannot be a good argument because it is impossible to conclude anything about what public resources should be devoted to something without relying on at least one claim that mentions public resources! Watch out for this common and easy-to-avoid error (Cullen 2022).

The main difficulty in applying what van Gelder called the 'Rabbit Test' is the determination of what exactly the 'significant components' of a conclusion are. While van Gelder and Cullen talk about 'concepts,' it seems more appropriate to focus on the smallest possible parts of a given conclusion that can still be formulated as a complete and grammatically correct sentence without changing the meaning of the original conclusion. It needs to be a complete statement because nothing else can be justified by reasons. Here is an example:

> (2) A person should be allowed to buy drugs that treat anxiety and attention deficit disorder (ADD) over the counter without a prescription once a month, which would be documented by providing passport data for a registration system.[10]

While it might be difficult to identify in (2) the 'significant concepts'—because there are so many concepts—it seems easier to list the greatest number of components that can be formulated as propositions without changing the meaning of the original conclusion. There seem to be just three:

1. Allow the purchase of drugs that treat anxiety and ADD without a prescription.
2. Such a purchase should be allowed only once a month.
3. Documentation with a passport is required for such a purchase.

However, one might ask whether the first component should be further divided into 'Allow the purchase of drugs that treat anxiety

[10] Created from an argument published on Reddit (u/SoftCatsMeow 2018).

and ADD' and 'No prescription should be required for such a purchase.' To answer this question requires background knowledge. One must know that regulations for pharmacies usually distinguish only between getting drugs with or without a prescription. The question of whether 'the purchase' should be allowed does not capture this distinction; taken in isolation, it is too vague. For this reason, the first component should be kept as suggested in the list above. The main point, however, concerns the fact that the possibility of an objective assessment is in doubt if the correct determination of the significant components in a conclusion depends on background knowledge.

An objective assessment, again, is possible when it comes to *limiting conditions* as they are indicated with phrases such as 'unless' and 'under the condition that …' Here is an example:

> (3) We have a moral duty to assist the third world through the distribution of genetically modified (GM) plants as long as the risk to human health of consuming GM crops is measurable and found to be within safe limits.

Applying the rule that the significant components of a conclusion should be determined by dividing this conclusion into the greatest possible number of components that can be formulated as propositions without changing the meaning of the original conclusion, we get to the following two components:

1. We have a moral duty to assist the third world through the distribution of GM plants.
2. This duty is limited to cases in which the risk to human health of consuming GM crops is measurable and found to be within safe limits.

Note that the second component cannot be divided further because it is the *conjunction* of both of these limiting conditions that needs to be justified, not the conditions themselves. It is important to note that not every complex statement can be divided into components:

(4) Crimes in which the damage to the victim is exacerbated by a long-term loss of trust in other people should be punished more severely than similar ones without.[11]

This statement cannot be divided into components in a way that satisfies the conditions discussed above. The reason is that any such division would either lead to sentences that are not grammatically complete or to statements in which the meaning of the original statement is changed significantly. This can be seen in the following divisions:

- Crimes should be punished more severely: More than what?
- Crimes should be punished more severely than similar ones: This is just incomprehensible.
- There are crimes in which the damage to the victim is exacerbated by a long-term loss of trust in other people: This is not what the original statement claims. It is not a claim about existence; rather, it is claiming that a very specific kind of crime should be punished more severely than another kind of crime independently from the question of whether there are such crimes or not.

Another example of a conclusion that cannot be divided into components is the one in Stern's (2006) argument for strong climate policies displayed in Figure 1: 'The benefits of strong and early action to avoid the worst impacts of climate change far outweigh the economic costs of not acting' (p. vi). This is one statement about 'outweighing.'

While arguments like these can be used to justify an objective assessment of the question of whether the Rabbit Rule has been violated in particular cases, there are other cases that require further rules. For example, it is not clear how to apply this criterion to conclusions that include qualifiers such as 'probably,' 'likely,' 'perhaps,' and so on. If we only add 'probably' to (4), we get:

[11] This is a modified version of a claim for which someone argued on Reddit (u/metheist 2018).

(5) Crimes in which the damage to the victim is exacerbated by a long-term loss of trust in other people should *probably* be punished more severely than similar ones without.

Should we identify here the following two components?

1. Crimes in which the damage to the victim is exacerbated by a long-term loss of trust in other people should be punished more severely than similar ones without.
2. This should probably be done.

Even though this distinction should be required according to our formulation of the rule that describes the determination of significant components, this is probably not a good idea. The reason for not counting qualifiers as components is that they usually weaken the conclusion and, thus, do not require their own justification by reasons. Such a qualifier represents an acknowledgment that the premises provided for a conclusion are not strong enough to justify what is claimed without a weakening qualification. The only case in which qualifiers should be counted as components is when they strengthen the conclusion: 'I am 100% certain that x is the case.' 'It is absolutely true that y.' In general, we could say, *something should be counted as a component of a conclusion if the expectation is that it should be justified*. If it makes a difference that something is claimed as part of a conclusion, then it should be counted.

However, if we engage in formulating additional rules to get to a more precise formulation of our assessment criterion that can capture a larger number of problematic cases, then we run into a problem. The problem is that since it is probably impossible to develop a complete list of such rules, the possibility of an objective application of this criterion is limited.

Overall, however, the criterion that all significant components of a complex conclusion need to be justified by at least one premise seems to allow more objectively decidable cases than we see

with some of the other criteria. Many of the examples discussed in this section are objectively decidable.

4.5 Are the premises relevant to the conclusion?

To capture the broad spectrum of discussions about relevance in argumentation theory, Fabio Paglieri and Cristiano Castelfranchi proposed, following prior suggestions in the literature, the distinction between

> . . . internal relevance, i.e. the extent to which a premise has a bearing on its purported conclusion, and external relevance, i.e. a measure of how much a whole argument is pertinent to the matter under discussion, in the broader dialogical context where it is proposed (2014, p. 216).

An example for external irrelevance is the fallacy of *ignoratio elenchi*, or irrelevant conclusion. Hans Hansen provides the example of the claim that "Calgary is the *fastest growing* city in Canada." If somebody tries to refute this claim with an argument "showing that it is not the *biggest* city in Canada," then this person argues against a *different* conclusion (2020; our italics).[12] Since the current contribution is limited to arguments in the sense of reason-conclusion constellations, we will focus only on internal relevance.[13]

---

[12] Paglieri and Castelfranchi (2014, p. 218) claim that *ignoratio elenchi* is a case of internal, not external relevance. But that is not convincing. What has later been called *ignoratio elenchi* is in Greek ἤ τοῦ ἐλέγχου ἄγνοια, which means literally 'ignorance regarding refutation.' For Aristotle, this refers to counterarguments that are fallacious because they violate certain principles that are determined by his definition of refutation (Aristotle *Soph. el.* 167a22-36). A refutation, however, is always external to an argument in the sense of a reason-conclusion constellation because it is not part of this constellation but instead attacks one of its elements from the outside. Their only evidence for their claim is from Aristotle (*Top.* 162a13–16) where the term is not mentioned.

[13] Walton's (2008) 'criticism of irrelevance' focuses only on the relevance of arguments in debates and, thus, on external relevance, so it is irrelevant in our context. The same is the case in his book *Relevance in Argumentation* (Walton 2004). The fact that Paglieri and Castelfranchi justify their thesis of a "deep" connection between trust and relevance with the reason "trust in relevance is an essential ingredient of everyday communication right now" indicates that they

A premise is irrelevant with regard to a particular conclusion if it does not contribute anything to the justification of what is claimed in the conclusion. A special case of this situation is a premise that simply repeats the conclusion or a part of it. This includes cases such as 'Paul is a bachelor because he is an unmarried man' where the terms used in reason and conclusion are synonyms or refer to the same thing by definition. To define internal irrelevance in general, we can follow a definition provided by Groarke and Tindale:

> If a premise increases the likelihood of the conclusion it is intended to support, or if it decreases the likelihood of that conclusion, then the premise is relevant to the conclusion. If neither of these conditions holds, then the premise is not relevant. (2008, p. 280)

A test for internal relevance could be as follows: Does your assessment regarding the truth of the conclusion change if you change the truth value of the premise from true to false? If there is no change regarding your assessment of the conclusion, then the truth or falsity of the premise does not have any bearing on the truth or falsity of the conclusion, and the premise is irrelevant. Note, however, that this test cannot be applied to irrelevance in the sense of repetition. If a premise just repeats a part of the conclusion (as in 2.3 in Figure 12 below), then its negation will create a contradiction to the conclusion. Overall, relevance is an epistemic criterion, not a formal one (Hitchcock 1992). Determining relevance requires an answer to the question: Does the truth or falsity of the premise change the likelihood of the conclusion?[14]

---

focus on external relevance as well (2014, p. 220). To be more precise, they argue convincingly that "trust in internal relevance indeed occurs within argumentative exchanges" (Paglieri and Castelfranchi 2014 p. 223). However, trust 'in' something means that the trust is given from the outside to something. Thus, the trust in the relevance of a premise for a conclusion is external to the reason-conclusion relation. Our argument here is that the relevance of a reason for a conclusion can be determined—at least in some cases—objectively so that no recourse to trust is needed.

[14] David Botting reaches the same result when he argues that a "theory of relevance is neither desirable nor possible" (2013, p. 1). Even though he titles his article "The irrelevance of relevance," he clarifies at the end: "When I speak

As Paglieri and Castelfranchi (2014) write in a rephrasing of a rule from the U.S. *Federal Rules of Evidence*: a premise is relevant only if it has "any tendency to make its conclusion more or less probable than it would be without" it. (p. 219). Similarly, Deidre Wilson and Dan Sperber measure relevance in terms of cognitive effects (2006, p. 609). However, whereas their approach allows for the determination of degrees of relevance—which requires the measurement of cognitive effects on a continuum—we prefer a binary understanding of relevance and irrelevance because that increases the chances of an objective determination. Either a premise increases the likelihood that the conclusion is true or it does not.

Paglieri and Castelfranchi (2014) provide a summary of the literature on both internal and external relevance referring, for example, to Aristotle's remark in the *Topoi* that "no inference will be drawn from" a premise that "has nothing to do with the conclusion" (Aristotle *Top.* 162a12–15), to Anthony Blair's notion of "premissary relevance" to describe the "idea of a premise's 'lending support' to a conclusion" (1992, p. 207), and to Scott Jacobs and Sally Jackson's "informational relevance of propositions to the truth value of a conclusion" (1992, p. 161). A clear example for irrelevant premises is provided in Figure 5. It is hard to imagine that anyone would disagree with the claim that 2.1 does not

---

of the irrelevance of relevance I do not wish to be taken as implying that it is not necessary to make relevance judgments; what is irrelevant is a theory or formal analysis of relevance" (p. 17). Such a theory, he claims, should be able to provide "a set of formal conditions that are individually necessary and jointly sufficient to establish that a premise is relevant to the conclusion" (p. 15). Formal conditions can never be sufficient if relevance is a question of epistemic judgments. This insight can also be used—in contrast to the argument that Botting (2013) develops—to reject an idea developed by John Woods: that the epistemic notion of relevance is "analyzable in terms of contextual implication" (1992, p. 190). For Botting (2013), contextual implication means that premise "P is relevant to Q in context C if P and C together non-trivially imply Q but neither P nor C on their own non-trivially implies Q" (p. 19). By focusing on implication, the discussion of relevance is transformed from an epistemological or cognitive problem to a formal or logical one. But why should we perform this shift? Woods (1992) does not provide a sufficient justification for this move. An important disadvantage of reducing relevance to a problem of implication is that it makes it virtually impossible to distinguish relevance from sufficiency.

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

contribute anything to increasing the likelihood of 1.1, and the same applies to 3.1. and 3.2 as premises for 2.1. Thus, at least this case allows an objective determination of irrelevance.

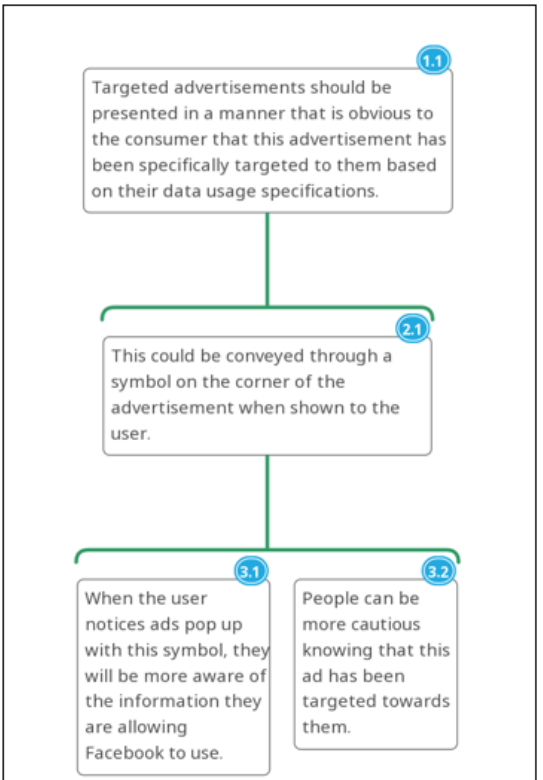


Figure 5: Premise 2.1 is obviously irrelevant for the conclusion in 1.1. It provides information about a detail that should itself be part of the conclusion. Additionally, 3.1 and 3.2 are not relevant for 2.1. However, since they would have been relevant as premises for 1.1, we can assume that the problem here is more missing familiarity with the requirements of argument mapping than with an understanding of relevance.

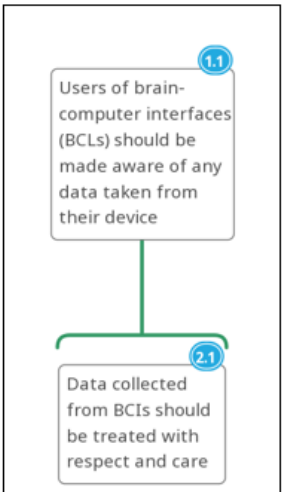


Figure 6: The reason is about the treatment of data, whereas the conclusion is about providing information that data were collected. But does that mean that the truth of the reason does not increase the likelihood that the conclusion is true?

An objective determination of irrelevance, however, cannot always be achieved that easily. If we look at the example in Figure 6, it could be argued that the premise in 2.1 does indeed increase the likelihood of 1.1 being true because it emphasizes the value of the data in question, and this value is relevant to the conclusion. This

example shows that additional rules are needed to increase the objectivity of the assessment. Such a rule could be: To determine whether the truth of the premise would increase the likelihood that the conclusion is true, do not consider anything that is not explicitly stated in the argument. This rule is important because it covers cases in which a premise becomes relevant in combination with one or more other premises, even though it might not be relevant in itself (more on this in section 4.7). However, what about applying this rule to an argument like the one in Figure 7 but without premise 2.2? It could be argued that an argument composed of 1.1 and 2.1 is just an enthymeme; the implicit assumption that a life with dignity requires that household chores are done on a regular basis is widely shared and intuitively clear. Applying the rule mentioned above would prohibit the acceptance of enthymemes, which seems too harsh. Of course, this problem could be resolved again with an additional rule, or with the refinement of the rule mentioned, but this only highlights a more fundamental problem. We might never be able to formulate a set of assessment rules that cover all possible cases of arguments whose premises should be classified as internally irrelevant. This leaves us again with a conclusion that is similar to the one drawn in the discussion on the appropriate formulation of the conclusion. While there are cases that can be decided with a high degree of inter- and intra-stability of assessment results, there are others that remain problematic.

4.6 Are the premises provided to justify a particular component of the conclusion sufficient to justify this component?

The sufficiency of premises is certainly a crucial assessment criterion, but also a very problematic one when it comes to objective application. Johnson and Blair proposed to define sufficiency as "the property of an argument's premises of supplying all the grounds that are needed to make it reasonable to believe its conclusion" (2006/1977, p. xv). The question, then, is obviously what providing "all the grounds" means.

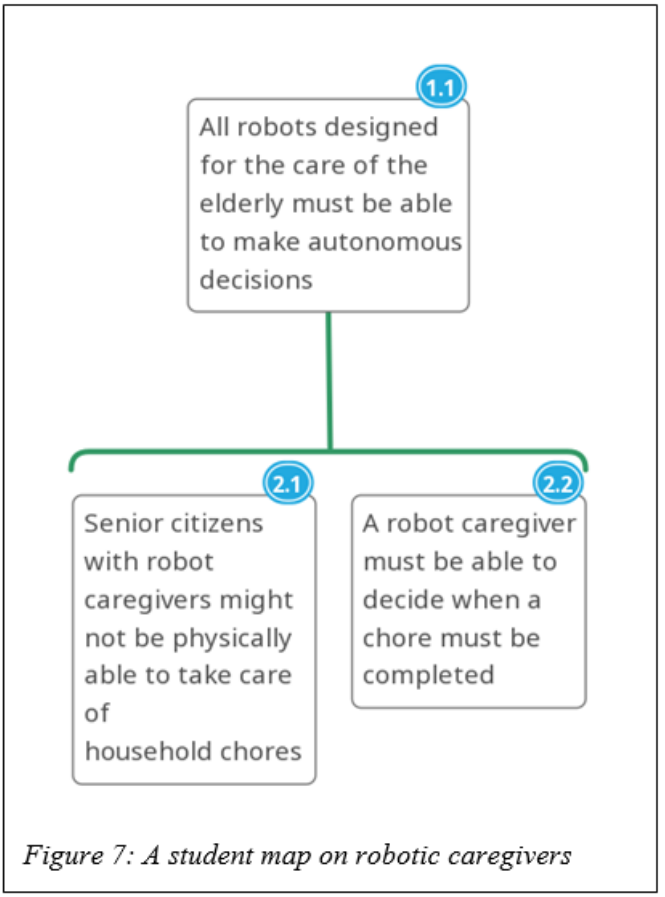


Figure 7: A student map on robotic caregivers

Before we discuss the question of whether, or to what degree, the sufficiency of premises can be determined objectively, two general remarks are in order. First, sufficiency can be defined differently in different contexts. For example, sufficiency in the context of logically valid arguments can be objectively determined by using the well-known formal criteria that determine logical validity in various logical systems. In legal contexts, the notion of 'beyond a reasonable doubt' and similar concepts seem to be well-defined so that they can provide at least a certain standard of sufficiency; whether this standard allows for an objective assessment of arguments is a question that goes beyond our expertise. For all such contextually defined standards of sufficiency, it is the respective

community of experts alone that will be able to determine whether standards can be applied objectively or based only on deliberation.

The second remark refers to the distinction between 'sufficient to *justify*' and 'sufficient to *support*' a conclusion. It is important to note that the latter condition is already satisfied by premises that are relevant, at least according to the definition of relevance suggested above. If a premise is relevant in case it increases the likelihood that the conclusion is true, then every relevant premise supports the conclusion. This way, however, there would be no point in distinguishing relevance from sufficiency as assessment criteria. As Leo Groarke and Christopher Tindale (2008) point out, the main difficulty in determining sufficiency is answering the question "how much is enough...? Experience tells us that this will vary from argument to argument. There are no precise rules for determining when enough evidence has been put forward" (Groarke and Tindale 2008, p. 281). However, the examples that we are going to discuss below show that there are indeed cases that seem to allow for an objective determination of insufficient premises. These examples do not only counter Groarke and Tindale's (2008) relativism, but also Blair's (2012) thesis that all norms of sufficiency are domain-dependent and therefore relative to collective assessment by specialists of the respective field.

Given our definition of an argument as a reason-conclusion constellation, we can distinguish three potential gaps between reason and conclusion regarding sufficiency. One concerns the *scope* of what is claimed in both—that is, what is 'covered' by the claims—and the second refers to the *degree of certainty* with which reason and conclusion are expressed. A third gap concerns generally accepted standards that certain kinds of claims or arguments require certain kinds of premises. Thus, we can distinguish a 'scope gap,' a 'certainty gap,' and an 'expectation gap.' With regard to the first two, it should be noted that the notion of 'gap' implies that the scope of the conclusion is broader than is justifiable by the premises or its degree of certainty is higher than can be justified by the reason. If it were the other way around, there would be no gap but an oversupply of certainty and scope provided by the reasons. Let us turn now to the question of whether there are types of cases for each gap that can be assessed objectively.

1. The scope gap: The scope of the conclusion is broader than justifiable by the premises provided

Figure 8 provides an example of the scope gap. The critical point in this argument is that premise 2.3 formulates a *limiting condition*. The moral duty to help is limited to cases in which we can be certain that our help reaches those in need.

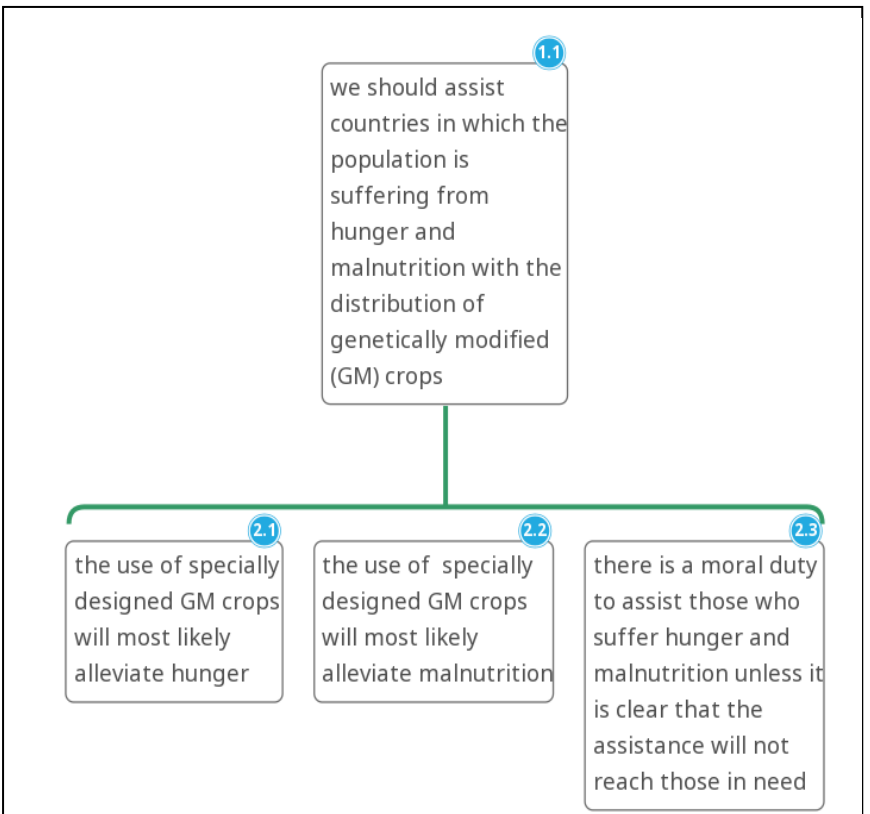


Figure 8: A modified version of an argument created by students. Are the premises sufficient to justify the conclusion?

However, if there is such a limitation in one of the premises, then the same limitation must also be included in the conclusion: We should assist these countries unless it is clear that the assistance

would not reach those in need. If the scope of a premise is limited in a certain way, then the scope of the conclusion needs to be limited in the same way. Otherwise, we get a scope gap: the conclusion claims more than can be justified by the reason. Whether or not this is the case can be assessed objectively; we only need to look for limiting conditions.

The best-known cases of arguments in which the scope of the conclusion is broader than allowed by the reason are, of course, inductive arguments in which a universal proposition is inferred from a limited sample of particular cases. Established scientific practices in empirical research require things like confidence level and margin of error to be added to any conclusion; this reflects an astute awareness of the problem of sufficiency. We should teach our students to do something similar by formulating generalizations carefully: data *indicate* a certain general conclusion, or a conclusion is *probably* true. An inductive argument that does not include an explicit acknowledgment of its limitations should not count as a good argument. Thus, we get again an objective assessment criterion: We need only to check whether the conclusions of inductive arguments are qualified or not.

Cases in which an objective assessment is more problematic include those in which the identification of a scope gap depends on background knowledge. An example that seems to allow for an objective assessment has been provided by Groarke and Tindale (2008). A disgruntled resident claims, based on her personal experience, that the postal service in the United States is inadequate (Groarke and Tindale 2008, p. 282). Obviously, any conclusion with such a broad scope would require at least something like a national poll. If the claim is justified only by observations in one's neighborhood, then there is a scope gap that can be identified objectively. However, there are certainly also cases in which differences in background knowledge preclude an objective assessment (Hoffmann 2018).

2. The certainty gap: The certainty of the conclusion is expressed in stronger terms than is justifiable by the premises provided

Here is an example of the certainty gap: 'Frodo will be late because he will probably take the route through downtown.' Given this qualified reason, the conclusion can only be: 'Frodo will probably be late.' An objective assessment is possible because we only need to check whether the qualification of a premise is followed by a qualified conclusion. Another example is a modified version of the argument in Figure 8: 'We have a moral duty to assist Third World countries through the distribution of genetically modified crops because the use of GM crops can potentially alleviate hunger in situations where crops with special features such as drought-resistance are needed.' The qualification in the reason ('potentially') might be needed because it might not be known whether those specially designed crops will actually grow in a larger variety of these situations. But if a reason is qualified in this way, then the conclusion needs to be qualified as well to avoid a certainty gap. As long as those qualifications are easily identifiable, the assessment can be objective.

3. The expectation gap: A certain type of claim or argument requires a certain kind of premises but these are not provided

Amnon Glassner points out that certain scientific claims or theoretical explanations require supporting evidence to verify them. If we justify the claim 'the earth orbits the sun' with reasons that refer to the 'gravitational force' exerted by bodies with a mass, then there is an expectation that we can provide evidence for the existence of such a force (Glassner 2017, p. 100). We can find universally accepted expectations regarding the kind of premises that need to be provided to achieve sufficiency in other areas as well. Our discussion of Stern's (2006) argument for strong climate policies, displayed in Figure 1, provides an example. If the conclusion claims that benefits outweigh costs, then the premises need to talk about both costs and benefits. Every cost-benefit argument must have a premise about costs and another one about benefits; if

one of these premises is missing, then the argument is objectively bad. The same applies to the following argument: 'The apples are cheaper than the oranges because the apples cost only $2 a pound.' It is clear that the argument can only be good if the price of the oranges is provided as well. If a corresponding premise is missing, then the reason given is objectively insufficient.

Maybe the most interesting feature of argumentation schemes is that they provide a standard that can be used to assess arguments with regard to the expectation gap. Since work on these schemes determines, for a large set of argument types, the kinds of premises required for a particular type of argument, the schemes—as long as they are generally accepted as standards for particular types of arguments—represent general expectations for sufficiency.

Objectivity regarding the expectation gap requires that expectations are generally shared. This is not always the case. Figure 9 provides the case of an argument with a normative claim (indicated by 'should') in the conclusion. The premises given in this argument are all factual claims. If we accept that the so-called naturalistic fallacy is indeed a fallacy, then we would assess the argument as bad because it infers an 'ought' from an 'is.' But not everyone accepts this assumption so that the assessment depends on one's stance regarding the 'naturalistic fallacy.'

Another problem regarding shared expectations refers to enthymemes. With regard to Figure 9, it can be argued that a premise such as: 'We should remove carbon dioxide from the atmosphere' is *implicitly* given. If we accept this, then the argument is a good one even for those who think there are naturalistic fallacies. Again, the assessment depends on one's stance.

The presented examples for arguments whose premises are insufficient because they do not bridge the scope, certainty, or expectation gaps show that an objective application of the sufficiency criterion is possible. But that does not mean that it is always possible. In cases of what Sally Jackson and Jodi Schneider (2018) called 'field dependency'—where standards are broadly accepted in one scientific discipline but not in others—or in those where the assessment depends on background knowledge or shared assumptions, objectivity is limited.

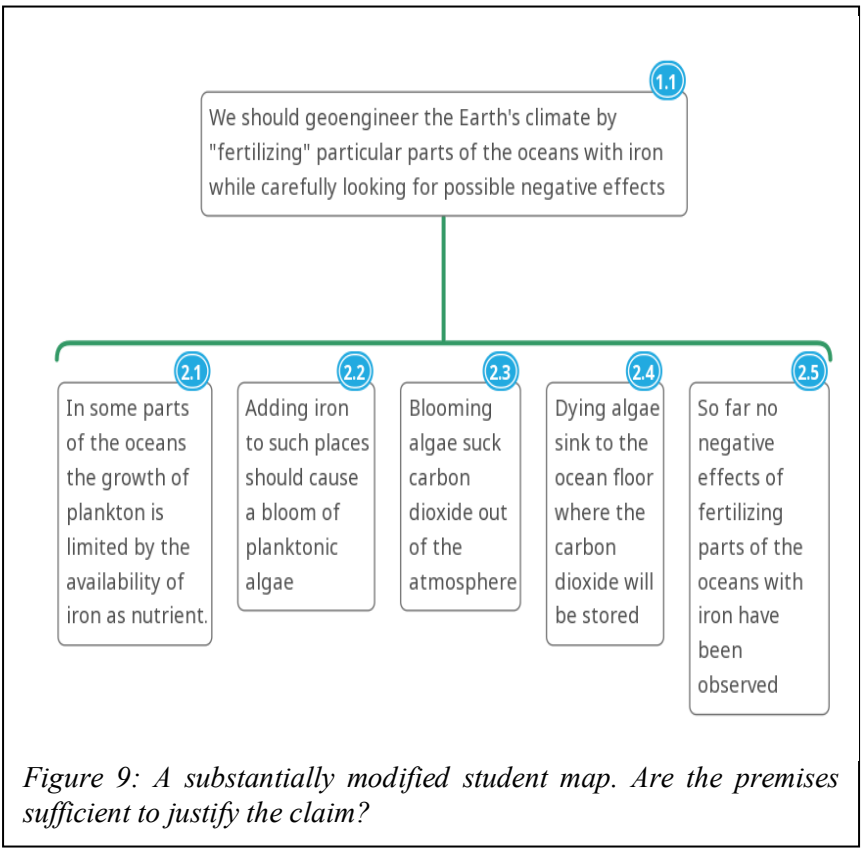


Figure 9: A substantially modified student map. Are the premises sufficient to justify the claim?

4.7 Have independent sets of premises for the same conclusion been correctly identified?

We saw in our discussion of Stern's cost-benefit argument for climate policies that the argument presented in Figure 1 would be objectively bad if one of its premises were missing. Since both premises are required, they are presented in the argument map as being connected. This way, we get a set of premises that is sufficient to justify the conclusion. In addition to this argument structure, it is also possible that there are two or more sets of premises, each of which is sufficient to justify the same conclusion. An example for this case is shown in Figure 10. In this representation, each set of premises forms, together with the conclusion, what we call an independent argument—independent because each set can

justify the conclusion on its own; it is not dependent on the other set. The premises in the second independent argument, by contrast, can be described as 'co-dependent' or 'mutually dependent' because both are needed to justify the conclusion.

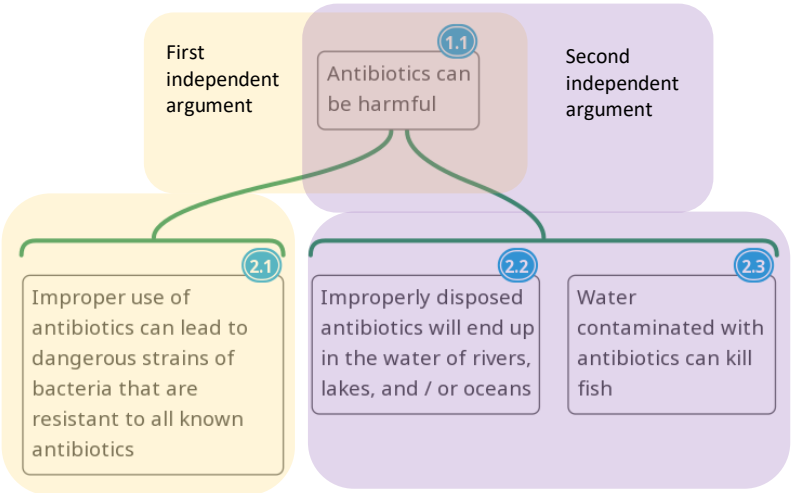


Figure 10: Premise 2.1 joins with the conclusion 1.1 to form a first independent argument for the conclusion. This argument is 'independent' from the one formed by 2.2, 2.3, and 1.1, which is formed by two co-dependent premises. In MindMup, this distinction is expressed by separated 'umbrellas' above sets of premises.

In the literature on argument structure, the situation depicted in Figure 10 is usually called a convergent argument because both independent arguments 'converge' on the same conclusion (see Yu and Zenker 2022, pp. 367-368, for a summary of the literature). We call it a convergent *argumentation* in the sense that it is a combination of two independent arguments. With regard to the second independent argument in Figure 10, its structure is usually called a 'linked argument' or an argument with linked premises (2.2 and 2.3). In Goddu's (2009b, p. 182) and Yu and Zenker's (2022, pp. 365-67) summaries of the literature, a structure is linked if premises provide 'inter-dependent support,' when 'the premisses must work together to support their conclusion,' and when each

premise 'needs the other to support the conclusion.' 'Taken independently, they do not support the argument's conclusion.' According to these formulations, it seems that 'linked' means exactly the same as 'mutually dependent.'

The literature about the linked-convergent distinction (and related distinctions) is too rich, too complicated, and too controversial to be discussed here. Since we are interested only in what we call 'objective' assessment criteria, the following is limited to what we need for just two specific rules. In order to avoid any confusion with the existing literature, we do not use the term 'linked' but only 'mutually dependent' or 'co-dependent,' which we define as follows: A set of premises consists of mutually dependent (or co-dependent) premises if and only if one or both of the following two conditions is fulfilled: either the defeat of one of these premises is sufficient to render at least one of the others irrelevant to the conclusion, or each of these premises is needed to justify a particular component of the conclusion. An example for a set of co-dependent premises that satisfies the first condition is the combination of premises 2.2 and 2.3 in Figure 10. Note that defeating premise 2.2 would not make 2.3 irrelevant. But since losing 2.3 would turn 2.2 into an irrelevant premise, both are co-dependent according to our definition. An example for the second condition is premise 2.3 in Figure 11. This premise justifies the component 'public resources should be devoted for doing this' (that is, 'ending aging,' the other component of the conclusion). Since all components of a conclusion need to be justified, as we argued in section 4.4, all the premises that are required for this justification are co-dependent.

The argument in Figure 9 presents an interesting case. Whereas premises 2.1 through 2.4 satisfy the first condition—losing one of them would render all the others irrelevant—premise 2.5 needs to be added to this set of co-dependent premises because it justifies a component in the conclusion that is only implicitly there: 'We should fertilize the oceans only if there are no negative effects of doing so.'

We continue using the term 'convergent' but define sets of premises as convergent if they—the sets—are not mutually dependent. The main function of convergent argumentations is to

increase the support provided for the conclusion. This way, we distinguish *arguments with mutually dependent premises* and *support-increasing convergent argumentations*.[15]

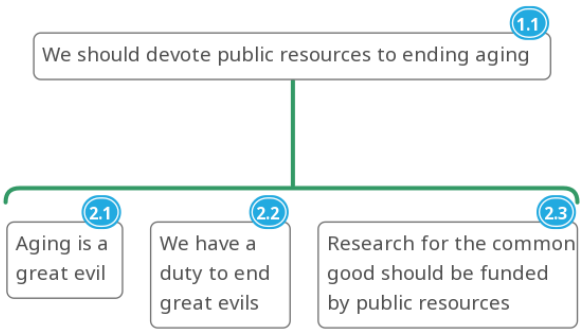


Figure 11: A modified version of the argument in Figure 4 above.

Why is it necessary to talk here about the distinction between arguments with mutually dependent premises and convergent argumentations? The distinction is important for two reasons.

[15] This distinction corresponds to James Freeman's distinction between 'relevance combination' and 'modal combination' (2011, pp. vii-viii). We acknowledge that there are cases in which it is difficult to decide which structure is given. For Freeman (2011), the following argument quoted from Stephen Thomas is an example of modal combination: [P1] His swimming suit is wet. [P2] His hair is plastered down. Therefore [C] He's been swimming (p. viii). However, let's imagine that it turns out that his swimming suit is completely dry. This does *not* mean that 'his hair is plastered' still provides some support for the conclusion. If the suit is dry, that statement about the hair *must* be connected to something other than swimming. From this point of view, [P1] and [P2] must be determined to be co-dependent because the negation of [P1] renders [P2] irrelevant. However, let us imagine that [P2] is false, not [P1]. In this case (hair dry and suit wet), [P1] is still relevant because he could have swum without getting his hair wet. This analysis has a more general implication. Yu and Zenker (2022) attempt to distinguish linked and convergent structures by purely analytical means so that the determination is not dependent on an "analysts' evaluative judgements" (p. 385). The example above indicates that his attempt is hopeless. Any decision about the relation between the two premises depends on an analyst's background knowledge about the world.

First, all arguments with co-dependent premises can be defeated by a counterargument that undermines the acceptability of just one of these premises—because these premises are, by definition, mutually dependent. If someone shows that just one of the four premises (2.1 through 2.4) of the argument in Figure 9 is false, then the conclusion is no longer justified; the argument breaks down. However, if a conclusion is justified by several independent arguments, losing one means that the conclusion is still justified by the others.[16] Thus, it is always better to have as many independent arguments as possible. However, since the premises of these arguments need to be sufficient to justify the conclusion, we need to check which premises should be connected as mutually dependent and which sets of premises can justify the conclusion independently.

For this reason, it is crucial to be familiar with the distinction between arguments with co-dependent premises and convergent argumentations and to construct justifications in a way that independent sets of premises are clearly and correctly identified. At this point, it should be noted that one of the major advantages of argument mapping software is that more sophisticated systems challenge users to reflect on this distinction by offering two display options—like the ones shown in Figure 10—to choose from. In natural language, by contrast, specific efforts are required to clarify this distinction.

The second reason for the importance of the distinction between arguments with co-dependent premises and convergent argumentations is that we need to be able to distinguish them correctly to prepare an argument for evaluation (Freeman 2011, p. 89). In more complex arguments, both the relevance and the sufficiency of premises often depend on their relation to other premises. This can be demonstrated with the argumentation in Figure 10. If you look at premise 2.2 in isolation, it is not relevant to the conclusion. Applying the test for relevance that we suggested in section 4.5, changing the truth value of 2.2 from true to false will not have any effect on the likelihood of the conclusion being true or false. However, premise 2.2 becomes obviously relevant in

[16] See Walton (1996, p. 175) and Freeman (2001, pp. 405, 413).

combination with 2.3. Since premise 2.2 would be, again, irrelevant in combination with 2.1, it is clear that the application of the relevance criterion depends on a prior structuring of an argumentation.[17] We need to know which premises are co-dependent and which can justify the conclusion independently before we can assess the relevance of particular premises. The same argument applies with regard to sufficiency. Since sufficiency is increased by combining a set of premises, we need to know which premises are combined in a particular argument representation before we can assess the sufficiency of the reason provided.

For both of these reasons, we should not give up the distinction between independent arguments for the same conclusion and co-dependent premises, as Goddu (2009b) suggested.[18] If we do not

[17] This in contrast to what David Hitchcock claims when he writes that "one has to do the evaluation first in order to classify the argument [as linked or convergent] in a way that indicates how one is to do the evaluation. Better just to do the evaluation and forget about the classification" (2015b, p. 90).

[18] Goddu 2009b claims to provide an argument that the linked-convergent distinction is useless; "there is no utility in making the distinction" (p. 183). However, he does not address our second reason for making the distinction, the one that Freeman puts into the rhetorical question: "How can one properly evaluate an argument unless one sees what supports what in that argument?" (2011, p. vii). Instead, he argues against the claim that "rejecting a single premise from each reason subset [is] sufficient" to refute an argument (p. 183). So, he basically argues that it is impossible to determine whether premises are mutually dependent. But the only thing he shows is that there are *examples* of arguments that *look like* arguments with co-dependent premises (in our terminology) but that can *also* be interpreted as convergent argumentations in the sense that each premise "provides some support for the conclusion" independently from the others (p. 184). But, the fact that there are those examples does not show that it is *always* impossible to refute an argument by "rejecting a single premise."—Yu and Zenker (2022) do not suggest giving up the linked-convergent distinction, but they argue that its justification requires better theories than are currently available. However, there is a fundamental difference between their way of approaching the problem and ours. For them, the goal is to find a 'test' that allows one to determine whether "a linked argument structure is distinguished from convergent structure" (p. 385). Such a test needs to be applicable to all those structures. Our goal is similarly broad when we ask whether in a given structure independent sets of premises for the same conclusion have been correctly identified. However, our burden of providing the means that are required to answer this question is substantially lower because—

give up the distinction, we must then ask how we can know whether an arguer correctly identified the independent sets of premises for the same conclusion in a particular representation like the one in Figure 10. Let us look at Figure 12 as an example. Based on the fact that the argument's conclusion has multiple components, the conclusion is sufficiently justified only if each independent argument—each set of co-dependent premises—includes everything that is needed to justify the combination of all components. This cannot be achieved by presenting premises 2.1 and 2.2 as being independently able to justify the conclusion. (Note that premises 2.3 and 2.4 are irrelevant. 2.3 repeats a component of the conclusion, and changing the truth value of 2.4 does not have any effect on the likelihood that the conclusion is true.)

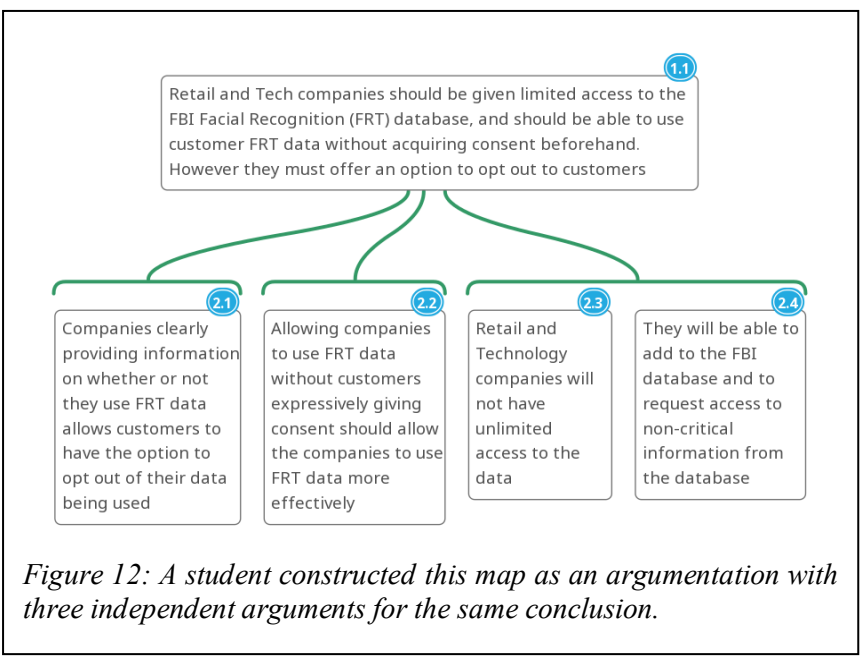


Figure 12: A student constructed this map as an argumentation with three independent arguments for the same conclusion.

Thus, we can formulate the following first rule to assess whether independent sets of premises for the same conclusion have been correctly identified:

---

as will be clear soon—we aim at just two specific cases in which, as we will argue, it is possible to answer this question with a high degree of confidence.

> *1. All those premises that are required to justify all components of a complex conclusion need to be connected in one reason for this conclusion.*

Whereas this rule is important only for arguments with more complex conclusions, the argument depicted in Figure 13 needs to be criticized based on the criterion of relevance (see section 4.5). This example justifies and illustrates our second rule:

> *2. If a premise is relevant only in combination with other premises, then the set of all those premises that are relevant only in combination need to be presented as co-dependent.*[19]

This second rule not only allows us to determine the appropriateness of the structure in Figure 10, but also that of the structure in Figure 9: premises 2.1 through 2.4 need to be presented here as co-dependent (premise 2.5 needs to be added to this structure according to our first rule). If you look at each one of these four premises in isolation, it is clear that none of them is relevant to the conclusion. They are relevant only in combination, which means they are co-dependent. The argumentation presented in Figure 13 clearly violates this second rule.

Even though it seems that these two rules can be applied objectively, it has to be noted that this objectivity is limited by the same limitations that we discussed with regard to the possibility of determining (1) relevance correctly and (2) all components of a conclusion. If it is not clear what the components of the conclusion are, then it also might not be clear which premises are required for a particular conclusion. If the relevance of a certain premise—be it in isolation or in various combinations with other premises—cannot be determined or is controversial, then it might not be possible to assign it to a particular set of other premises.

---

[19] Freeman formulated a similar criterion: "We hold that when two or more premises must be taken together to form a relevant reason for the conclusion, the structure is linked, while the structure is convergent when the premises are independently relevant to the conclusion" (2011, p. 89).

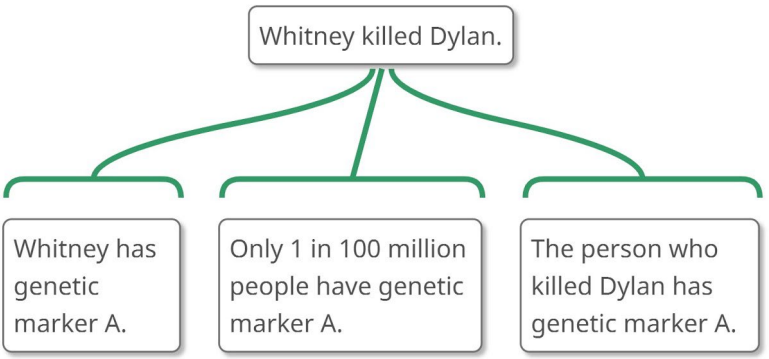


Figure 13: Since none of the premises, as presented here, increases the likelihood that the conclusion is true, they are all irrelevant. They would be relevant if combined. This example has been published by Simon Cullen (2022). Cullen points out that not realizing the linked-convergent distinction correctly "is one of the most common" mistakes in argument construction. We can confirm that, but we have to note that this happens much more frequently when students are using MindMup than when using the AGORA system, as we found out in the NSF-funded study mentioned above.

With regard to relevance, however, we now face a problem of circularity. Above we argued that assessing relevance sometimes requires knowing the structure of the argument, whereas now we are saying that assessing the structure requires a prior assessment of relevance. This circularity is an important outcome of our analysis. Whereas the possibility of clearly distinguishing arguments with linked premises from convergent argumentations has frequently been questioned (Goddu 2009b; Yu and Zenker 2022), what we are questioning here is the possibility of assessing *relevance* in all those cases where the structure is controversial and a premise is relevant only in combination with other premises. (To be clear: there is no circularity if the premise can be assessed as relevant without looking at other premises. The problem occurs only if there is more than one premise and one of them is relevant only in combination with others.)

In these cases, the observed circularity can be overcome only if a particular structure is clearly given. Relevance always must be

assessed, so it can never simply be assumed as given. Only if an argument or argumentation is presented in a clearly identifiable structure—as is the case with all the argument maps used in this contribution—are we able to assess relevance and the adequacy of the chosen structure in *one* process. If we do not know how the creator of an argument or argumentation intended to structure a set of premises, we cannot determine whether such a premise is relevant without imposing a structure ourselves—which should not be done if we want students to learn about the linked-convergent distinction. This consideration provides a powerful argument for the use of argument mapping software in education because with these tools—if they are well designed—it is impossible to present an argument or argumentation without a particular structure. If the structure is given, the assessment of relevance and structure can move forward together. For example, the assessment of the structure in Figure 13 should start with asking if each of the three premises is relevant in isolation—because the map presents a convergent argumentation. The result would be no, none is relevant in isolation. The next step should then be to ask, which of the premises would be relevant if linked to another one? At this point it becomes clear that the problem of the argumentation is the structure not the irrelevance of the premises. Independent sets of premises for the same conclusion have not been correctly identified.

4.8 Are there contradictions among the propositions used in an argument?

We mentioned in section 4.1 that an inconsistent conclusion cannot be justified. If there is a contradiction in the conclusion, the assessment can stop right there. Contradictions can also occur among the premises provided, or among any propositions in the set used in an argument. If one justifies the claim 'We should go swimming' not only by the observation 'it is hot,' but also by the claim 'we should not go swimming,' then the argument cannot be good. That the identification of contradictions or inconsistencies might not always be that easy is shown in the example in Figure 14.

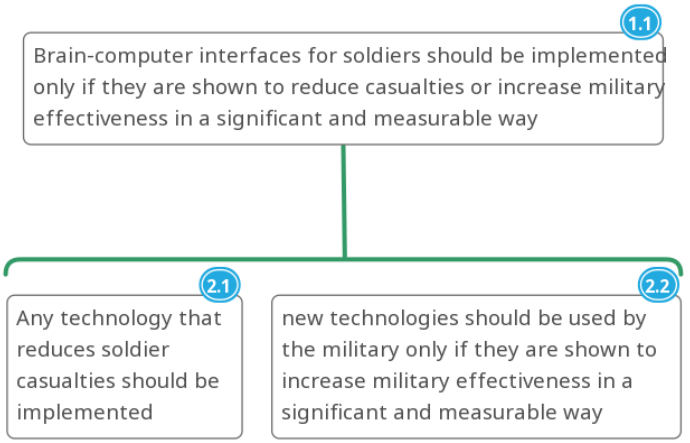


Figure 14: According to the conclusion of this argument, only one of two conditions must be fulfilled for the implementation of brain–computer interfaces (BCIs) ('only if'), and it does not matter which one: either they are shown to reduce casualties or they significantly increase military effectiveness. The second premise, however, claims that any new technology—which includes BCIs—should be implemented 'only' if the second of the two conditions in the conclusion is fulfilled. Thus, the argument contains a logical contradiction. 2.2 claims that increasing military effectiveness is a necessary condition for using new technologies, and 1.1 denies this.

Once it is discovered, however, the presence of a contradiction can be determined objectively as a question of logical consistency. Note that one type of argumentation, *reductio ad absurdum*, uses contradictions intentionally. This does not violate the eighth criterion because our criterion should be applied to arguments only, not to argumentations. This is important here as can be seen with the first known *reductio ad absurdum*: the proof that the diagonal of a square and its side are incommensurable. The proof starts with the assumption that they are commensurable. Then a series of arguments demonstrates that this assumption leads to a logical contradiction, an 'absurdity.' The final argument uses this contradiction and the fact that commensurable and incommensurable are mutually exclusive to show that the two lengths are indeed incommensurable.

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

5. How frequently are the eight criteria violated?

In addition to what has been developed so far in this paper, we would like to provide some data from two different assessment contexts demonstrating how often these criteria are violated. Both assessments were performed in philosophy classes that offered ethics education for students across campus at a research-oriented university in the United States. The first class was in the fall of 2018 entitled "Science, Technology, and Human Values" (enrollment: 162), and the second class—with the same title as the 2018 class—was offered in the spring of 2019 in two independently taught sections (enrollment was 35 in each section). All three classes implemented almost the same curriculum, which included four class meetings on argument mapping and argument assessment. All were taught by the first author of this paper. The assessment in 2019 also included 33 students from a class on "Philosophy of Food" that was taught by a colleague. Students in this class did not receive any systematic training in argument construction and assessment. Across the three classes taught in 2019, 95 students participated in the experiment.

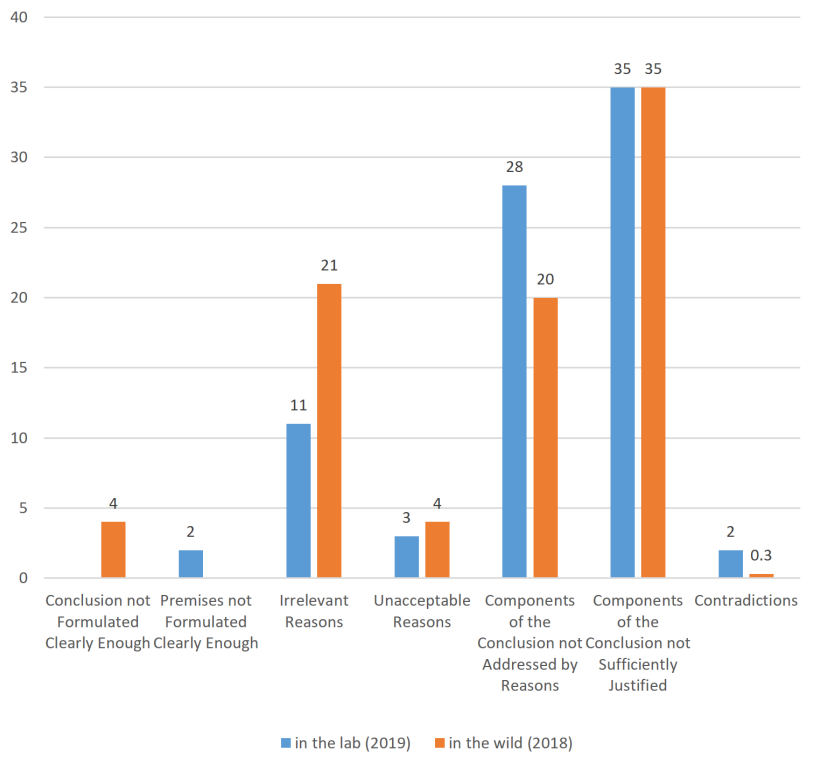
The main difference between the 2018 and the 2019 assessment settings was that the latter was performed under quasi-lab conditions: students worked independently using their own laptops on prepared argument construction tasks during a class meeting. In 2018, by contrast, argument maps that teams of about four students created at the end of a semester-long project were assessed. These were the projects on 'wicked problems' that we briefly described in section 3 above. Since the argument maps that the teams created in 2018 provided justifications for multiple components of team-generated proposals, they were all different. For this reason, the arguments created in 2018 were referred to as being 'from the wild' whereas the 2019 maps were created under more controlled lab-like conditions.

In spite of this 'in the wild' versus 'lab' difference, the proportions of particular criteria-violations that are depicted in Figure 15 are quite similar. The 'in the wild' observations have to be taken with a grain of salt. Whereas the 2019 assessment was performed by two coders after substantial training, the 2018 assessment was conducted as a pre-study to test a first draft of our assessment

criteria, and it was performed by just the first author. The discrepancies for various criteria in Figure 15 can be explained by the fact that we developed the detailed *interpretations* for each of the eight criteria only in the context of the training for the coders in 2019.

To determine possible limitations of the 2019 study, we need to say more about the methodology used. Participating students volunteered for extra credit in their courses. Participation occurred during time set aside in a one-hour class meeting for those who chose to be in the study. The experiment was entirely web-based and, as previously mentioned, participants used their own laptops.

In the lab-like study, participants did some tasks with provided argumentative texts of about half a page or a bit longer together with a more or less complex conclusion that was extracted from the texts. (These texts were slightly modified versions of arguments published on the subreddit "Change my view" on Reddit.) These tasks were part of a larger study and are not discussed here. The two tasks of interest were ones in which participants were given only a conclusion and were asked to construct an argument using their own reasons. Participants coming from the "Science, Technology, and Human Values" courses created their arguments using argument mapping software (Agora or MindMup); they were trained in the use of one of these software tools during the argument mapping section of the curriculum mentioned above. Students from the other course did not have experience with any of these software tools and typed their arguments into a text document using Microsoft Word. A total of 95 students participated, divided roughly equally among the three conditions (Agora, MindMup, and text document). Participants' data on some tasks were removed from analysis if the participant did not follow the instructions; for example, some participants altered the provided conclusion.



.

Figure 15: The percentages of mistakes observed for seven assessment criteria. There are no data for 'conclusion not formulated clearly enough' in the 2019 lab study because here all the conclusions were provided. There are no data for 'premises not formulated clearly enough' in 2018 (in the wild) because this criterion was not yet identified as significant at the time. There are no data regarding the question of whether students identified independent sets of premises for the same conclusion correctly because this criterion turned out to be problematic with regard to the main goal of both studies: to determine differences in the use of various argument mapping tools.

The two conclusions for which participants developed their own reasons and constructed arguments concerned stadiums and music. The provided conclusions were:

- Cities should never subsidize the construction of stadiums for professional sports teams, but they should help to provide the infrastructure around stadiums (roads, parking, etc.).
- All students in elementary school should be required to take a course each year in which they are taught to play a musical instrument.

For the stadiums task, data from all 95 participants were used; for the music task data from 91 of the participants were used (some participants were excluded due to incomplete responses or other data collection issues). Using a scoring rubric based on the criteria discussed above except the first and the seventh,[20] two coders independently assessed the quality of the arguments participants created. The intercoder reliability for the different criteria ranged from 0.7 to 0.9, which is considered 'good' to 'very good' (Landis and Koch 1977). The overall quality in the three conditions did not differ for the stadiums task and showed a tendency towards worse performance for Agora participants in the music task. Since our focus is on the frequency of certain types of errors, and given the overall similarity of the groups' performances on the tasks, we collapsed the data over the three conditions in order to focus on the errors.

The coders analyzed the quality of 215 arguments. This means that each participant created 1.20 arguments on average for each task (i.e., the 'stadiums' task and the 'music' task). The reason there is more than one argument per task is that some participants created sub-arguments that justified reasons of the main argument. These 215 arguments had 628 premises, 616 of which were formu-

[20] As already mentioned, the first criterion was not used because the conclusion was provided so that it did not make sense to assess the quality of its formulation. The seventh criterion—referring to the correct application of the distinction between independent arguments for the same conclusion and arguments with co-dependent premises—was not used because we realized in 2018 that users of MindMup created about 17% more arguments that failed with regard to this criterion than users of Agora. We assumed that this might be caused by MindMup's user interface. Since our research hypothesis was not about user interfaces, we decided to avoid the risk of confounding our research results with extraneous factors by not using this criterion.

lated clearly enough. The percentages depicted in Figure 15, however, do not represent the frequencies of mistakes for all these arguments and premises. We learned early on that we had to use our assessment criteria in a particular sequence. For example, assessing the relevance of a premise was almost impossible if this premise was formulated too poorly. Looking at the formulation first allowed us to remove a particular premise from all following assessment steps. This way, the number of assessed premises became smaller and smaller through the assessment process. Additionally, coders were instructed to terminate the assessment in cases in which none of the premises provided were relevant or acceptable. Table 1 shows the ratios from which we calculated the percentages depicted in Figure 15.

*Table 1: The ratios underlying the percentages in Figure 15**

	Criterion / assessment step	Ratios	
		Lab study	In the wild
1	Conclusion not clear enough		18/475 (3.8%)
2	Premises not clear enough	12/628 (1.9%)	
3	Irrelevant premises	68/616 (11.0%)	152/735 (20.7%)
4	Unacceptable premises	14/548 (2.6%)	25/583 (4.3%)
5	Components of conclusion not addressed	106/380 (27.9%)	111/553 (20.1%)
6	Insufficiently justified components	95/274 (34.7%)	153/442 (34.6%)
7	Contradictions in argument	4/202 (2.0%)	1/362 (0.3%)

* *The order of the rows (numbered 1-7) represents the sequence of assessment steps. (Note that we implemented, at that time, a different sequence than the one we justify in the beginning of section 4.) The denominator in line 3 is smaller than the denominator in line 2 because a premise could not be assessed for relevance unless it was judged to have been formulated clearly enough. Twelve of the 628 premises in line 2 were judged 'not clear,' thus 616 premises were left to be considered for relevance in line 3. Similarly, 68 of the 616 premises from line 3 were judged not relevant, thus there were 548 premises left to be considered for acceptability in line 4. Line 5 shows the ratio of those compo-*

nents of conclusions that were not addressed by reasons that were still assessed at this point (each of the conclusions cited above from the lab study has two components). The denominator in line 6 represents all the components of conclusions that were addressed by reasons (380 minus 106, from line 5). Line 7 displays the contradictions across all arguments whose assessment was not terminated before this last assessment step.

The frequencies with which participants made mistakes could have been influenced by a variety of factors in addition to the participants' skill in argument construction and analysis. These factors include the particular topics in the arguments (e.g., the participants' depth of knowledge and emotion about a topic might have affected performance), time pressure, and a number of individual differences. It is important to keep in mind that the participants were students at university with stringent entry requirements and thus generalizations to the larger population must be done cautiously. Our speculation is that the frequencies of mistakes would likely go up with a more general sample.

6. Conclusion

The goal of this contribution was to determine a set of criteria that can be used to assess particular features of any argument objectively. 'Objective' was defined as being applicable with a high degree of inter- and intrasubjective stability. The search for these criteria was driven by three practical requirements. First, each criterion should be applicable to all arguments in the sense of a reason-conclusion constellation and not just to a subset of arguments such as deductively valid arguments or arguments that can be subsumed under a specific argument scheme or a fallacy. Second, the list of assessment criteria should be cognitively manageable in the assessment practice; this means their number should be limited to those that are most often violated by arguers or that are considered to be most important by our community of experts. Third, the criteria should be specified to a degree of precision that permits an acceptable level of objectivity in the sense of interrater reliability.

 Informal Logic, Vol. 43, No. 1 (2023), pp. 23–90.

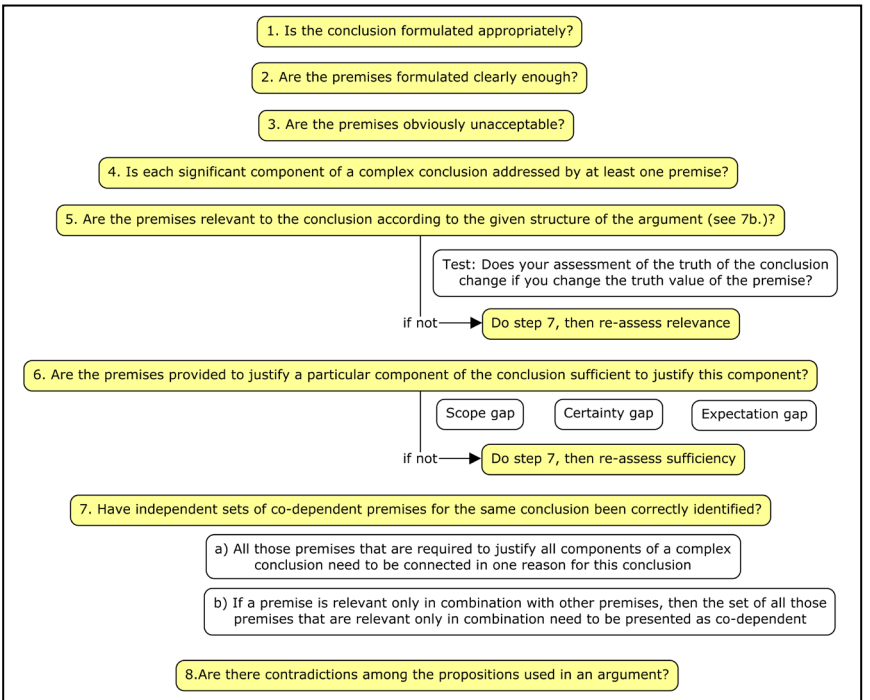


Figure 16: Summary of the assessment procedure

The result is a list of eight assessment criteria that we presented in the form of eight questions. These criteria are applicable to all arguments and should be cognitively manageable; they can be used in the form of a checklist like the one depicted in Figure 16. Still, answering these questions *objectively* is possible only in a limited sense. For example, the first question 'Is the conclusion of an argument formulated appropriately?' usually does not allow an answer that is inter- and intrasubjectively stable, as we discussed. We believe that it is possible to answer objectively only questions such as: Does the conclusion state anything? Is it itself an argument? Is it an inappropriately nested proposition? or: Is it so badly formulated that its meaning is incomprehensible or depends clearly on the assessor's interpretation? Under the umbrella of an 'appropriately formulated conclusion,' only these sub-questions can be answered objectively. The same limitation applies to the next

seven criteria as well: a higher level of confidence that an objective assessment is possible requires the use of the more specific determinations that we discussed for almost all of these criteria.

Based on these limitations, the proposed approach cannot answer the question of what constitutes a good argument; instead, it is limited to the evaluation of specific features of arguments that can be described as objectively bad. For an objective determination of a good argument, it would be necessary for a reasonable audience to agree that there is *nothing* wrong with this argument. However, the only assertion that our approach can justify objectively is that a particular argument passes assessment with regard to a list of very specific points such as all components of the conclusion have been addressed by at least one reason; there is no scope, certainty, or expectation gap between the reason provided and the conclusion; and so on.

Besides developing these eight criteria, the paper also contributes to ongoing debates in argumentation theory and informal logic. We believe two contributions in particular are significant. The first is the two tests that we are proposing: one for determining whether a certain premise is relevant or not and one for identifying the significant components of a given conclusion. The second is our analysis of the vexed problem of the linked-convergent distinction in the debate about argument structure.

As we wrote in the beginning of this paper, the ability to assess the quality of arguments is crucial for scientific reasoning, for deliberation in public and private spaces, and for critical thinking in general. Efforts to improve argument construction and assessment can impact the world on multiple fronts.

Acknowledgements: We are thankful for the contributions that Dania Ibrahim and Typhanie Hall made to the project. Not only did they assess the quality of argument maps created by students for our frequency analysis of certain mistakes, but their feedback during their training as quality assessors was invaluable for the formulation of the argument quality criteria that we are presenting here. We thank our colleague Jeremy Lingle with whom we collaborated on a related project; our discussions on that project improved the present one. We also want to thank John Walsh,

Daniel S. Schiff, and Bryan Norton for feedback on earlier versions of this paper and Justin Biddle for allowing us to measure the argument skills of the students in his class. Thanks also to Kamal Korrapati for motivating us to create a summary of the entire assessment procedure (Figure 16). Finally, we wish to thank the two anonymous reviewers who provided many constructive and thought-provoking comments that motivated significant improvements. This research has been supported by a grant from the National Science Foundation (Award 1623419). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aberdein, Andrew. 2010. Virtue in argument. *Argumentation* 24(2): 165-179.

Aristotle. 1955. Soph. el. Sophistici elenchi. In *On sophistical refutations. On coming to be and passing-away. On the Cosmos*, ed. E. S. Forster, 10-155. Cambridge, Mass.; London: Harvard Univ. Pr.

Aristotle. 1966. Top. Topica. In *Aristotle. Posterior analytics. Topica*, ed. E.S. Forster, 108a18–164b20. Cambridge, Mass: Harvard Univ. Pr.; London: Heinemann.

Award Abstract # 1623419; EXP: Fostering Self-Correcting Reasoning with Reflection Systems. 2020, November 30. National Science Foundation. URL accessed 17 March 2023: <https://www.nsf.gov/awardsearch/showAward?AWD_ID=1623419&HistoricalAwards=false>.

Baer, Paul and Clive L. Spash. 2010. Is climate change cost-benefit analysis defensible? A critique of the Stern Report. In *Science for Policy*, eds. Angela Guimaraes Pereira and Silvio. Functowicz, 167-192. Oxford University Press.

Baumtrog, Michael D. 2021. Designing critical questions for argumentation schemes. *Argumentation* 35(4): 629-643.

Blair, J Anthony. 2012. What is the right amount of support for a conclusion? In *Groundwork in the Theory of Argumentation*, ed. J Anthony Blair, 51-59. Springer.

Blair, J. Anthony. 1992. Premissary relevance. *Argumentation* 6(2): 203-217.

Blair, J. Anthony. 2019. Judging arguments. In *Studies in Critical Thinking*, ed. J. Anthony Blair, 225-247. URL accessed 19 December

2021: <https://windsor.scholarsportal.info/omp/index.php/wsia/catalog/book/106>.

Botting, David. 2013. The irrelevance of relevance. *Informal Logic* 31(1): 1-21.

Boudry, Maarten, Fabio Paglieri and Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4): 431-456.

Catrambone, Richard. 2011. *Task analysis by problem solving (TAPS): Uncovering expert knowledge to develop high-quality instructional materials and training* [conference presentation]. Learning and Technology Symposium, Columbus, GA.

Cohen, Daniel H. 2013. Virtue, in context. *Informal Logic* 33(4): 471-485.

Cullen, Simon. 2022. *Handy hints for making arguments.* Philmaps.com. URL accessed 17 March 2023: <https://maps.simoncullen.org/hints>.

Dove, Ian J. and E. Michael Nussbaum. 2018. The critical question model of argument assessment. In *Argumentation and Inference: Proceedings of the 2nd European Conference on Argumentation, Fribourg 2017, Vol. 1*, eds. Steve Oswald and Didier Maillat. College Publications: London.

Eemeren, Frans H. van and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach.* Cambridge University Press. Cambridge.

Ennis, Robert H. 2001. Argument appraisal strategy: A comprehensive approach. *Informal Logic* 21(2): 97-140.

Feldman, Richard. 1994. Good arguments. In *socializing epistemology. the social dimensions of knowledge*, ed. Frederick F. Schmitt, 159-188. Lanham, MD: Rowman & Littlefield Publishers:

Freeman, James B. 2001. Argument structure and disciplinary perspective. *Argumentation* 15(4): 397-423.

Freeman, James B. 2011. *Argument structure: Representation and theory.* Dordrecht; New York: Springer.

Glassner, Amnon. 2017. Evaluating arguments in instruction: Theoretical and practical directions. *Thinking Skills and Creativity* 24: 95-103.

Godden, David and Frank Zenker. 2018. A probabilistic analysis of argument cogency. *Synthese* 195(4): 1715-1740.

Goddu, G. C. 2009a. Refining Hitchcock's definition of 'Argument'. In *Proceedings of the 8th International Conference of the Ontario Society for the Study of Argumentation (OSSA),* 1-12. Accessed 15 March 2019:

<http://scholar.uwindsor.ca/ossaarchive/OSSA8/papersandcommentaries/55/>.

Goddu, G. C. 2009b. Against making the linked-convergent distinction. In *Pondering on problems of argumentation: Twenty essays on theoretical issues,* eds. Frans H. van Eemeren and Bart Garssen, 181-189. Springer Netherlands: Dordrecht.

Goddu, G. C. 2018. Against the intentional definition of argument. In *Argumentation and Inference: Proceedings of the 2nd European Conference on Argumentation,* Fribourg, Switzerland 2017, eds. Steve Oswald and Didier Maillat. London: College Publications.

Goodman, Jeffrey. 2018. On defining 'Argument'. *Argumentation* 32(4): 589-602.

Govier, Trudy. 2010. A practical study of argument (7th ed.). Belmont, CA: Cengage Learning. (Original work published in 1985)

Groarke, Leo. 2019. Depicting visual arguments: An ART approach. In *Informal logic: A 'Canadian' approach to argument,* ed. Federico Puppo, 332-374. URL accessed 19 December 2021: <https://windsor.scholarsportal.info/omp/index.php/wsia/catalog/download/123/303/1653-1?inline=1>.

Groarke, Leo and Christopher W. Tindale. 2008. *Good reasoning matters! A constructive approach to critical thinking* (4th ed.). Don Mills, Ont.; New York: Oxford University Press.

Hansen, Hans V. 2020. Fallacies. In *The Stanford encyclopedia of philosophy* (summer 2020 edition), ed. Edward N. Zalta. URL accessed 19 December 2021: <https://plato.stanford.edu/archives/sum2020/entries/fallacies/>.

Hitchcock, David. 1992. Relevance. *Argumentation* 6(2): 251-270.

Hitchcock, David. 2007. Informal logic and the concept of argument. In *Philosophy of logic*, ed. Dale Jaquette, 101-129. Elsevier: Amsterdam.

Hitchcock, David. 2015a. Freeman's syntactic criterion for linkage. *Informal logic* (35)1: 1-31.

Hitchcock, David. 2015b. The linked-convergent distinction. In *Reflections on theoretical issues in argumentation theory,* eds. Frans H. van Eemeren and Bart Garssen, 83-91. Springer International Publishing.

Hitchcock, David. 2020. Arguing for questions. In *From argument schemes to argumentative relations in the wild,* eds. Frans H. van Eemeren and Bart Garssen, 167-184. Springer.

Hoffmann, Michael H. G. 2010. The debate about the Stern-Review and the economics of climate change. Argument map. In *SMARTech, ed*. Georgia Institute of Technology. URL accessed 14 January 2023: <http://hdl.handle.net/1853/46190>.

Hoffmann, Michael H. G. 2018. The elusive notion of "argument quality". *Argumentation* 32(2): 213-240.

Hoffmann, Michael H. G. 2019. Transcendental arguments in scientific reasoning. *Erkenntnis* 84(6): 1387-1407.

Hoffmann, Michael H. G. 2020a. Reflective consensus building on wicked problems with the Reflect! platform. *Science and Engineering Ethics* 26: 793-819.

Hoffmann, Michael H. G. 2020b. The argument assessment tutor (AAT). In *Reason to dissent: Proceedings of the 3rd European conference on argumentation* (vol. I,), eds. Catarina Dutilh Novaes, Henrike Jansen, Jan Albert van Laar and Bart Verheij, 289-303. London: College Publications.

Jackson, Sally and Jodi Schneider. 2018. Cochrane review as a "warranting device" for reasoning about health. *Argumentation* 32(2): 241-272.

Jacobs, Scott and Sally Jackson. 1992. Relevance and digressions in argumentative discussion: A pragmatic approach. *Argumentation* 6(2): 161-176.

Johnson, Ralph H. 2000. *Manifest rationality: A pragmatic theory of argument.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Johnson, Ralph H. and J. Anthony Blair. 2006. *Logical self-defense*. New York: International Debate Education Association. (Original work published in 1977)

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159-174.

Lewinski, M. and M. Aakhus. 2014. Argumentative polylogues in a dialectical framework: A Methodological Inquiry. *Argumentation* 28(2): 161-185.

Lumer, Christoph. 2011. Argument schemes – An epistemological aproach. In *Argumentation. Cognition and community. Proceedings of the 9th international conference of the Ontario Society for the Study of Argumentation (OSSA),* ed. F. Zenker, 1-32. Accessed 18 December 2018: <http://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1016&context=ossaarchive>.

Nordhaus, William D. 2007. A review of the Stern review on the economics of climate change. *Journal of Economic Literature* 45(3): 686-702.

Paglieri, Fabio. 2015. Bogency and goodacies: On argument quality in virtue argumentation theory. *Informal Logic* 35(1): 65-87.

Paglieri, Fabio and Cristiano Castelfranchi. 2014. Trust, relevance, and arguments. *Argument & Computation* 5(2-3): 216-236.

Plantin, Christian. 2021. Argumentation through languages and cultures. *Argumentation* 35(1): 1-7.

Prakken, Henry. 2010. On the nature of argument schemes. In *Dialectics, dialogue and argumentation. An examination of Douglas Walton's theories of reasoning*, eds. Chris Reed and Christopher W. Tindale, 167-185. London: College Publications.

Roberts, David. 2012. *Discount rates: A boring thing you should know about (with otters!)*. Grist. URL accessed 17 March 2023: <https://grist.org/article/discount-rates-a-boring-thing-you-should-know-about-with-otters/>.

Spash, Clive L. 2007. The economics of climate change impacts a la Stern: Novel and nuanced or rhetorically restricted? *Ecological Economics* 63(4): 706-713.

Stern, Nicholas Herbert. 2006. *The economics of climate change. The Stern review: Summary of conclusions*. URL accessed 16 August 2010: <http://webarchive.nationalarchives.gov.uk/+/http://www.hm-treasury.gov.uk/d/Summary_of_Conclusions.pdf>.

Stern, Nicholas Herbert. 2007. *The economics of climate change. The Stern review.* Cambridge, UK: Cambridge University Press.

The Rabbit Rule. (n.d.). Reasoninglab; Tools for critical thinking, writing and decision making. <https://www.rationaleonline.com/explore/en/tutorials/tutorials/Tutorial_2/6_Rabbit_Rule/rabbit_rule.htm>.

Tindale, Christopher W. 2007. *Fallacies and argument appraisal*. Cambridge; New York: Cambridge University Press.

u/metheist. (2018, April 6). *CMV: Crimes involving breach of trust should have more severe legal repercussions than the similar ones without* [Online forum post]. Reddit. <https://www.reddit.com/r/changemyview/comments/8a9nt2/cmv_crimes_involving_breach_of_trust_should_have/>.

u/SoftCatsMeow. (2018, March 2). *CMV: Some psychiatric drugs should be legalised (be allowed to buy over the counter) to receive without a prescription* [Online forum post]. Reddit. <https://www.reddit.com/r/changemyview/comments/81d2fc/cmv_some_psychiatric_drugs_should_be_legalised_be/>.

Walton, Douglas N. 1996. *Argument structure: A pragmatic theory*. Toronto; Buffalo: University of Toronto Press.

Walton, Douglas. 2008. *Informal logic: A pragmatic approach* (2d ed.). Cambridge: Cambridge University Press.

Walton, Douglas N. 2004. *Relevance in argumentation.* L. Mahwah, NJ: Erlbaum Associates.

Walton, Douglas N., Chris Reed and Fabrizio Macagno. 2008. *Argumentation schemes.* Cambridge; New York: Cambridge University Press.
Wilson, Deirdre and Dan Sperber. 2006. Relevance theory. In *The handbook of pragmatics,* eds. Laurence R. Horn and Gregory L. Ward. Malden, MA: Blackwell Pub:
Woods, John. 1992. Apocalyptic relevance. *Argumentation* 6(2): 189-202.
Yu, Shiyang and Frank Zenker. 2020. Schemes, critical questions, and complete argument evaluation. *Argumentation* 34(4): 469-498.
Yu, Shiyang and Frank Zenker. 2022. Identifying linked and convergent argument Structures: A problem unsolved. *Informal Logic* 42(2): 363-387.

Appendix: Yu and Zenker's 'Complete Argument Evaluation'

The approach by Yu and Zenker (2020) is based on the literature on argument schemes. Based on the limitations of argument schemes for argument assessment that we discussed in section 2, we will focus here only on the more general part of Yu and Zenker's (2020) approach.

Yu and Zenker (2020) start from a distinction between a 'logical' and a 'substantial' meta-level representation of argument schemes. Combining both, they claim, is sufficient to get completeness in argument assessment. Focusing only on the general meta-level structure, what is relevant here regarding the *logical* representation is the distinction of three possible 'instances' of this structure: [1] 'premise–conclusion'; [2] the same as [1] but with the additional inference link 'If premise(s), then conclusion,' and [3] the same as [1] but with one or more additional premises describing "Absence(s) of exception(s)" (Yu and Zenker 2020, p. 473). Whereas the first instance is obviously so general that it indeed covers all arguments—simply by definition if we define, as suggested in the introduction, an argument as a reason-conclusion constellation—the second and third instances are still committed to argument schemes: [2] by focusing on *modus ponens* as a logical scheme, and [3] by turning critical questions into additional premises—as in turning the question 'Is *a* in a position to know whether *A* is true (false)?' in the 'argument from position to know'

into the premise '*a* is in a position to know whether *A* is true' (p. 472).

For Yu and Zenker (2020), the 'meta-level logical forms' that are represented in these three instances are 'evidently' reasonable (p. 473) so that their normativity, it seems, should be used to assess the reasonableness of any argument scheme and any of their instantiations in concrete arguments. This is certainly convincing for instance [1] even though not much is gained for argument assessment because any 'argument' that does not express a premise-conclusion relation will not be counted as an argument anyway based on the definition of argument that we are using here. But for instances [2] and [3], it is less convincing to use their normativity for argument assessment. As already discussed, using logical validity as a criterion for argument quality excludes too many arguments as bad that are generally considered to be good; and even though listing 'absences of exceptions' as additional premises is an excellent idea to improve the quality of an argument, the fact that we do not have a complete list of argument schemes limits this approach.

What about the authors' 'substantial' meta-level representation of argument schemes? Yu and Zenker (2020) start by claiming that the 'if–then' proposition in instance [2] above—that is, in the logical structure "Premise(s); If premise(s), then conclusion; Therefore, conclusion" (pp. 473 and 476)—"is central to both deductive inferences such as *modus ponens*, as well as to inductive ones such as the statistical syllogism" (p. 477). Moreover, they claim that the 'if–then' premise also specifies a *substantive* relation between premise and conclusion, not only a logical one. This way they feel justified to extend the 'meaning' of the 'if–then' premise to "The relation *R* holds between the referents of premise(s) and conclusion" (p. 477). This relation is considered to be substantive because it covers relations such as symptomaticity, similarity, and causality (p. 476). With this extension, they get to the following meta-level structure that 'combines' a logical with a substantial representation:

(2) Premise(s).

> The relation *R* holds between the referents of premise(s) and conclusion.
> Therefore, conclusion. (Yu and Zenker 2020, p. 477)

In a somewhat surprising move, Yu and Zenker (2020) claim that Toulmin's well-known model of argument "proves sufficient" to develop an account of argument schemes that is both logical (regarding the 'premise–conclusion' relation) and substantive—with regard to the fact that the "relation *R*" mentioned in (2) represents substantive relations between premise(s) and conclusion (p. 479). This way, the argument goes, a modified version of the Toulmin model should be sufficient for a complete determination of argument quality criteria (Yu and Zenker 2020, p. 481). But why the Toulmin model? The authors provide the following argument for choosing Toulmin:

> Because the model prescribes how an argument's structural parts should function, it offers a useful sense in which the model is normatively complete: an argument is good (or valid) *if, and only if,* all model components are fully explicit and fulfill their functions. Thus, to evaluate an argument completely is to evaluate whether all its components function well. (Yu and Zenker 2020, p. 481)

Despite using Toulmin extensively, Yu and Zenker (2020) end up deriving criteria for argument assessment just from the basic definition of an argument as a constellation of reason and conclusion. This is similar to our position. If we focus on the basic structure 'reason–conclusion,' then a complete list of critical questions can be derived from what can be attacked in any argument. If there are just premises and a conclusion, then a premise can be attacked, or the inferential relation between reason and conclusion, or the conclusion itself (by, for example, showing that another argument contradicts this conclusion; see Prakken 2010). Since there is nothing in a basic argument besides reason, conclusion, and an inferential relation, the list of three questions that can be derived from the three possibilities of attack is complete. Accordingly, there are just three basic critical questions for Yu and Zenker (2020). However, they argue that two of them should be divided

into ‘sub-CQs’ so that their complete list contains seven critical questions (Yu and Zenker 2020, pp. 488-490).

In our context, only two considerations are important with regard to Yu and Zenker’s (2020) approach. First, since their list of seven critical question does not provide anything that can be used to determine whether all components of a more complex conclusion have been addressed by premises (our fourth criterion), it cannot be complete. Having a set of premises that is rich enough to address all components of a conclusion is certainly an important criterion to assess the quality of arguments. Thus, any list of criteria that is supposed to be complete should include it.

However, and this is the second consideration, now we have to deal with the obvious contradiction between claiming that the list of just three critical questions is complete because, as argued above, there are exactly three possible attack relations and the argument that Yu and Zenker’s list—which seems to include these three critical questions—is incomplete. A solution for this problem can start with the observation that something important is happening between Yu and Zenker’s (2020) list of three critical question and their final list of seven.

Before we discuss this observation, one thing needs to be clarified. The three CQs that they derive from the three possible attack-relations are committed to the Toulmin model in so far as they use ‘data’ instead of reason or premise in their formulation of the questions. As Yu and Zenker (2020) point out, Toulmin treats ‘data’ and ‘backing’ as facts so that “the D-to-C inference is *fact-to-claim*” and the backing-to-warrant “inference is *fact-to-rule*” (p. 482). However, there are many arguments in which the conclusion or claim is not justified by factual data but only by a general rule. Consider ethical arguments in which a general rule like ‘do not break a promise’ is justified by another general rule like Kant’s categorical imperative. These arguments are not covered by Toulmin’s model. For this reason, we think that Yu and Zenker’s (2020) three CQs should be formulated differently as follows:

Table 2: Complete lists of critical questions?

	Yu and Zenker 2020, pp. 488-489	*Hoffmann and Catrambone*
CQ-1	Are the data correct?	Are the premises acceptable?
CQ-2	Is 'If D, then C' correct?	Is the inferential relation between reason and conclusion acceptable?
CQ-3	Is the claim correct?	Is the conclusion acceptable?

Based on the argument that there are only three attack relations, and given the fact that our list on the right allows us to assess arguments with both particular (factual) and universal propositions (rules) in the positions of reason and conclusion, whereas the Toulmin-based reference to 'data' in CQ-1 and CQ-2 excludes the assessment of arguments in which premises are not factual statements, our list on the right is complete whereas theirs is not. For this reason, we will use in the following only our own list of three critical questions as a set that is complete.

Since we used in our argument against the completeness of the Yu and Zenker (2020) list the criterion that all components of a conclusion need to be addressed by the premises provided, we should ask ourselves whether this criterion—although it is not explicitly mentioned in our list—can be subsumed under one of the three critical questions. The best fit is obviously CQ-2 which refers to the quality of the inferential relation between reason and conclusion. The premises provided cannot be sufficient if they leave a component of the conclusion unaddressed, and sufficiency refers to the inferential relation.

The reason why the criterion relating to the components of the conclusion is not covered by Yu and Zenker's (2020) seven criteria—even though it could be covered by CQ-2—is that they dissolve CQ-2 into four sub-CQs, none of which covers it:

CQ-2.1 What is the intended category of D's subject?
CQ-2.2 What is the content of the 'D therefore C' relation?
CQ-2.3 Does the relation between D's intended category and C's predicate hold necessarily?

CQ-2.4 Does D's subject belong to an exception-class of its intended category (as per CQ-2.1)? (Yu and Zenker 2020, pp. 489-490)

If our fourth criterion can be subsumed in their short but not in their long list, then there is obviously something wrong. However, there is another, more general lesson that can be drawn from a comparison between Yu and Zenker's (2020) three or seven critical questions and our list of eight assessment criteria. There is a difference between *analytical* approaches that determine criteria based on an analysis of the definition of an argument and those that are looking at *observable argument practice*. Even though Yu and Zenker's (2020) three critical questions are complete from an analytical point of view, they do not explicitly cover our first, second, and eighth assessment criteria which are derived from observations. Two of these criteria are that the conclusion (1) or the reasons (2) are so badly formulated that any assessment can stop right there because it is not really clear what exactly the argument is. The other criterion is that there might be a contradiction among the components of an argument (8). It could be argued that those three criteria are implicitly covered by Yu and Zenker's (2020) criteria because, for example, an evaluator cannot satisfactorily answer CQ-2.2 (What is the content of the 'D therefore C' relation?) without considering whether the conclusion and the premises are formulated appropriately.[21] However, a key feature of our criteria is to direct the user's attention explicitly to certain points that should be taken into account. It is not sufficient to hope that people who do not have much experience with argument assessment ask the right questions. For instance, the user's attention is not necessarily directed to our criterion 1 when looking at Yu and Zenker's question.

[21] We thank an anonymous reviewer for raising this issue.