WHAT COUNTS AS A "GOOD DATA-BASED ARGUMENT"? STUDENTS' EVALUATIVE CRITERIA IN A CLASS DISCUSSION

Carlotta Vielmo[1], Mirko Maracci[2]

[1]University of Trento, [2]University of Pisa

This paper investigates how 10th-grade students conceptualise criteria for evaluating data-based arguments in decision-making contexts. Drawing on Habermas' construct of rational behaviour, we analyse how evaluative criteria are articulated and negotiated in students' talk in a whole-class discussion aimed at co-constructing validity and quality criteria. Findings show diverse students' conceptualisations of prototypical criteria such as correctness, completeness, and clarity, and the emergence of terms as orientation, reasonableness, and significance, whose meanings cut across epistemic, teleological, and communicative dimensions.

INTRODUCTION

Citizens increasingly encounter decisions in several domains, justified through statistical evidence. As a result, engaging in data-based argumentation, i.e., making claims and justifications grounded in data while accounting for uncertainty and variability, is increasingly framed as a key component of statistical literacy (Bolstad, 2025). Although research extensively investigates students' reasoning with data, the construct of data-based argumentation has received comparatively less explicit conceptual development (e.g., Krummenauer et al., 2023). Beyond producing data-based arguments, students also need opportunities to evaluate their quality, especially in decision-making contexts where multiple analytical options and interpretations are available. In this regard, moments of collective evaluation can foreground what the classroom community comes to treat as legitimate evidence, acceptable inferential moves, and appropriate ways of communicating data-based claims. However, how evaluative criteria for "good" data-based arguments are made explicit, discussed, and negotiated in class interactions remains under-explored. This study is part of a broader ongoing project (Vielmo, 2025) investigating how statistical rationality takes shape in mathematics classrooms. Statistical rationality is conceived as a specific form of rational behaviour (Habermas, 2003) that may develop when working with data and justifying and evaluating data-based claims under uncertainty and variability. Within this perspective, this contribution analyses a class discussion situated within a designed teaching sequence and explicitly oriented toward evaluating data-based arguments. The aim is to characterise the forms of statistical rationality enacted as students collectively negotiate criteria for argument quality.

CONCEPTUAL FRAMEWORK

This study draws on Habermas' characterisation of rationality by three intertwined dimensions: an epistemic dimension, involving the validation of claims according to

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 979-986. PME.

shared premises and legitimate modes of argument; a teleological dimension, concerning the adequate selection and use of tools and strategies to achieve the activity's goals; and a communicative dimension, including adherence to rules ensuring both the ability to communicate the steps of argumentation and the conformity of the outcome to the standard or target reference (Habermas, 2003). More specifically, we conceptualise statistical rationality as a domain-specific form of rational behaviour that may develop when working with data and become visible in communicative interaction, when reasons are made public, open to critique, and oriented toward intersubjective validation.

In this view, data-based argumentation is seen as a discursive practice co-constructed within the classroom community, where data works as central support for claims and decisions, requiring engagement with uncertainty and variability (Büscher, 2024). While data-based argumentation is often linked to inferential reasoning (e.g., Gómez-Blancarte & Tobías-Lara, 2018), this study focuses on foundational data-level choices (e.g., selecting variables, grouping cases, or choosing representations) that already entail justification and demand to coordinate statistical knowledge (SK) and context knowledge (CK) (Wild & Pfannkuch, 1999). As SK and CK may draw on distinct rationalities, their coordination may also open tensions around what counts as a reasonable interpretation and a warranted conclusion in stochastic settings.

Since argumentation is framed as rational behaviour emerging in interaction, the construction and negotiation of criteria for argument quality become constitutive of the development of rationality and analytically relevant for investigating how rationality is enacted (Boscolo et al., 2025). Research on formative assessment takes *correctness*, *completeness*, and *clarity* as core evaluative criteria to support teachers' and students' meta-talk about what counts as a good mathematical argument (Cusi et al., 2017) and as discursive mediators that support the articulation and negotiation of rational dimensions in classroom interaction (Boscolo et al., 2025). The dimensions of rationality are interpreted through how evaluative criteria are articulated, justified, and negotiated in classroom discourse.

We draw on this research while acknowledging an important specificity of data-based argumentation. Working with uncertainty and variability implies that argument legitimacy is often tied less to a single necessary deductive chain than to the transparency and coherence of analytical choices. As multiple analytical moves are available, the argumentative product must account not only for the final claim and the supporting evidence but also for the entire inquiry process. In this sense, the teleological dimension, while enacted at the process level, may also surface within the product itself. Within this framework, we interpret the epistemic dimension of statistical rationality as manifesting primarily in how students mobilise correctness and completeness. References to correctness may concern reading and elaborating data without errors, the consistency of data interpretations with activated CK, the reliability of selected data sources, and the warranting of conclusions by the statistics used,

accounting for uncertainty. Completeness, from an epistemic perspective, concerns the relevance of the statistical and contextual resources selection to meaningfully represent the phenomenon and the coherence of the argumentative chain across the SK and CK domains. The teleological dimension is interpreted as manifesting in the goal-orientedness of strategies and the completeness of the argumentative product with respect to the task. It becomes visible when students evaluate the adequacy of selected data, representations, or statistical tools in supporting a given claim, and when completeness is invoked to consider alternative analyses, data sources, or representations before drawing a conclusion. The communicative dimension is interpreted as manifesting in how students mobilise clarity and completeness. Clarity concerns the intelligibility and transparency of the argumentative product for the target community, while completeness refers to providing sufficient textual or visual information for others to understand, assess, and potentially challenge the argument.

DESIGN CONTEXT AND FOCUS OF THE STUDY

The study is situated in a teaching sequence designed to foster data-based decision making and meta-level evaluation of arguments, while creating conditions under which such processes can be observed. Within this sequence, particular attention is given to whole-class discussions explicitly oriented toward evaluating data-based arguments and negotiating what counts as a valid or convincing justification in a stochastic context. The present study focuses on one whole-class discussion in which students collectively examine and assess data-based arguments related to the same decision-making task. Against this background, the study addresses the following research question: *Which criteria emerge from the classroom discussion? In particular, which criteria align with the conceptualization of the a-priori criteria and, if additional ones emerge, which dimensions of rationality do they appear to reflect?*

While the study builds on an a priori framework of evaluative criteria, the analysis remains open to identifying additional criteria emerging from classroom discussion.

METHOD

The study was conducted in a 10th-grade class of an Italian science-oriented high school, involving 21 students. The class participated in a 10-hour teaching sequence designed to foster data-based argumentation and the evaluation of argument quality. Students are here anonymised and referred to as S1-S21 throughout the paper.

The empirical focus of the study is a whole-class discussion centred on a statistical decision-making task on school mobility. Students had to decide whether to support the installation of bicycle maintenance stations (pit-stops) in the school district and to justify their position using a dataset on students' commuting habits, explored through a web-based data analysis tool (for the full task, see: https://tinyurl.com/taskPME49). Prior to the discussion, students produced written responses to the task and examined six pre-constructed exemplars (Grion et al., 2017) representing various levels of argument quality. These activities served as a shared reference for the subsequent

discussion. The whole-class discussion, conducted by the first author, unfolded across two consecutive sessions, was video-recorded, and transcribed verbatim.

Analytical perspective and analysis protocol

The analysis adopts a qualitative, discourse-oriented perspective and follows a hybrid approach. The unit of analysis consists of evaluative contributions, defined as utterances (or clauses within an utterance) in which participants assess the quality of a data-based argument, request clarification or justification, propose revisions, or negotiate what counts as an appropriate evaluative criterion. The a priori criteria outlined in the conceptual framework were used as initial interpretive lenses to examine evaluative contributions in relation to correctness, completeness, and clarity. When contributions could not be adequately captured by such criteria, additional ones were identified inductively through iterative comparison. For each contribution, the analysis considered (a) the aspect of the data-based argument addressed (e.g., data reading, data interpretation, or use of representations), and (b) the role of the contribution within the ongoing discussion (e.g., challenge, agreement, clarification request, or negotiated reformulation), to interpret how evaluative criteria were enacted and negotiated in interaction rather than merely named. In a subsequent step, evaluative contributions and the criteria they mobilised were interpreted in relation to the epistemic, teleological, and communicative dimensions of rationality. Coding and interpretation were iterative and refined across episodes through repeated comparison.

FINDINGS

The discussion reveals *correctness* in ways that align, to varying degrees, with the a priori conceptualization, manifesting both epistemic and teleological dimensions. Correctness is first framed in general terms by S1, saying that the first exemplar “has no weaknesses, because it says something concise and correct”. S1 then articulates correctness more precisely as warranting the claim by the statistics used: “the important thing is to use the data correctly” and an incorrect conclusion is one that “contradicts itself with the data”. Within this framing, S1 also notes that data can only support an “expectation”, not a certain conclusion, indicating some awareness of the epistemic status of inferential claims, even though qualifiers are not explicitly voiced. A second framing of correctness regards the sensibility of contextual interpretations. For instance, S2 and S3 call a conclusion “wrong”, saying “because almost 9 out of 100 is too small”, referring to the 8.8% biking rate as insufficient to conclude installing bike pit-stops. Here, correctness refers to the plausibility of the interpretation and its fit in the argument, rather than misuses of statistical outputs. A third focus concerns correctness understood as reading and elaborating data without errors. This aspect becomes salient when the researcher prompts students to consider the appropriateness of computing the mean of the “class” variable, thereby foregrounding correctness in relation to data handling at the statistical level. However, this attempt does not lead to a stable shared understanding. At a later point in the discussion, S4 argues that one “cannot talk about correctness in reading data, because that is personal interpretation,”

thus collapsing reading into interpretation within the domain of CK. While initially apparently accepted by students, the distinction between reading (in SK terms) and interpreting does not hold across the discussion. When students are asked to judge the correctness of a mean computation, the class does not converge, indicating that the very possibility of assessing correctness at this level remains contested. This is well illustrated by two contrasting contributions: S3 states that the mean "is correct, but wrong to use it," while S5 justifies by referring to the suitability of quantitative data for computing a mean. Another epistemic aspect related to correctness - namely, the reliability of data sources- is never raised, even though one exemplar relies on a newspaper article, which could have invited such reflection.

The discussion foregrounds *completeness* at various times. Students often use the term *complete* to evaluate whether the evidence or elaboration is sufficient and relevant to represent the phenomenon. For instance, an exemplar justifies installing pit-stops by checking biking rates within a limited distance, and S3 and S6 criticise it, as the argument "is not complete, because it considers just the range under 4km". This narrowing of the dataset is seen as undermining the argument's legitimacy, as the conclusion should draw on the whole dataset. Similarly, S7 and S8 note that the same argument "does not take into account how bikers are distributed across schools" and one should "check how many people attend each school", pointing to overlooked sources of variation that may weaken the argument's epistemic validity. Across both sessions, students propose diverse ways to tackle incompleteness: considering other variables (e.g., S2 mentions commute time), gaining more contextual information (e.g., S4 suggests considering possible pit-stops locations) or collecting new data from a survey (e.g., S9 suggests investigating willingness to change mobility habits). When S7 and S8 suggest completing the exemplar by including more data, S10 states that "it is also important to understand what the purpose is" and contrast two installation aims (supporting an existing ecological behaviour or targeting the majority), arguing that what counts as relevant evidence (only the rate of current bikers or also non-bikers) shifts accordingly. S10 highlights the importance of the intended goal within CK and opens space for intending completeness as accounting for the evidence's epistemic legitimacy or as ensuring that, once the goal is fixed, all relevant data and elaborations have been considered, enacting the teleological dimension. The role of the goal and the related adequacy of strategies is made explicit by S1, noting the inquiry process requires to first formulate a thesis to reach a conclusion, otherwise one "looks at the data in vain". Moreover, S4 says that the author of one exemplar "could have written that, in their view, installing them [pit-stops] could potentially increase the percentage [of bikers]; otherwise, it does not make sense", linking goal-stating to argument's quality. To some extent, S4's comment may also point to completeness related to the communicative dimension, focusing on what must be communicated for the argument to be understood.

Clarity is not explicitly addressed in the first session, thus the researcher directly questions the clarity of an exemplar in the second one, eliciting multiple views. Some

students treat it as comprehensibility (e.g., S5 says "[the argument is] written in a simple way, without beating around the bush"; S11 claims that "data are explained well"). Others link it to structural aspects: for S1, a clear exemplar combines "a text that explains the reasoning and the considered data (...) and a graph that shows the reasoning". S1 later argues that "two or three lines that only make the data explicit" are not enough, and S11 counters that a text "can be clear, even if short", revealing a tension between brevity and sufficient explication. An analogous tension surfaces in talking about graphs. Some students (e.g., S7, S9, S10) consider the graph needed because it allows for explanation "beyond words"; S3 replies that it "must be a support instead of the whole answer", calling for an accompanying text. When prompted to compare graphs with different stylistic features, students treat clarity as readability (e.g., "one is less intuitive"; "I find it difficult, because it is not easy to read"). S10 adds that "if the graph includes too much information, it can also be confusing; here information is smaller and, let's say, disorganized", whereas S1 values information-rich graphs because they "show the data clearly", attributing to graphs a role in making data accessible to the reader. The importance of accounting for the target community to evaluate communicative quality remains implicit in students' talk and is made visible only by the researcher.

A new criterion emerging early in the discussion is that of *orientation*, used in at least two distinct ways. First, as goal-alignment: some students stress that the argument should provide data supporting the argument's aim. This can be seen when S10 states that "data must be oriented toward the goal," echoed by the researcher's call for data collection to be "oriented and complete." Second, in epistemic terms, as the provision of meaningful and valid data: S1 explains that "oriented means selecting data that we know are meaningful and cannot be disproved," implying that a not-oriented argument lacks legitimacy. S3 adds that orientation should not be confused with correctness. Later, S5 refers to an "oriented and significant elaboration", possibly invoking both purpose and evidential strength.

Reasonableness is often invoked as a criterion assessing whether an interpretation of statistical evidence is acceptable. S10 triggers its meaning negotiation, noting that evidence may be reasonable for some and not for others. S4 and S1 react by framing reasonable evidence as "proved by data", emphasising a data-driven perspective. In contrast, S9 argues that personal interpretation is still needed, otherwise data may legitimise unrealistic claims (e.g., "if the maximum distance that people were travelling was 30 km, (...) people should be willing to walk 30 km"). Reasonableness thus becomes a site of tension between data-based and context-based conceptions, sometimes blending both (e.g., S10 closes with "2 km is reasonable, because it is quite objective"), swinging between epistemic legitimacy (proved by data) and teleological concern, judging plausibility in light of action implications.

The term *significant* appears in two exemplars and was reprised and used by students in different moments. For instance, S1 argues that the exemplar "likely referred to 2.11

km as a significant datum because (...) if we looked at the distance travelled, say, across Italy, we might see that 2.11 km is significant", and adds that "if I bring a datum into my argument, I bring it because, statistically, it's a datum that holds up; it should be shown that it is a significant datum", then noting which data properties make a mean a "significant index". The term also occasionally appears to be used interchangeably with, or closely connected to, other terms such as "oriented", leaving open whether the term may refer to statistical legitimacy or goal-oriented adequacy. S5's call for an "oriented and significant" elaboration exemplifies this ambiguity.

DISCUSSION AND CONCLUSION

The study investigates which criteria emerge from the class discussion and which dimensions of rationality they appear to reflect. The analysis shows that students do not rely on a single, stable set of criteria, but rather on a constellation of evaluative criteria whose meanings are continuously re-shaped in interaction.

With respect to the a priori criteria, correctness, completeness, and clarity all emerge as salient evaluative resources. However, they are mobilised in multiple ways and negotiated, with shifts that blur the boundaries between epistemic, teleological, and communicative concerns. Correctness is associated with warranting the claim by the statistics used and with concerns about reading and elaborating data. While contextual interpretation is mainly discussed in terms of plausibility and decision relevance, the risk of misreading statistical outputs within CK remains largely implicit, as does data-source reliability. Completeness becomes a central site of negotiation, alternately referring to evidential sufficiency, goal-oriented adequacy and explanatory sufficiency. Clarity is elaborated in relation to comprehensibility and structural organisation, revealing tensions between brevity and explication and between simplicity and informational richness.

Alongside these criteria, new ones emerge. Terms such as orientation, significance, and reasonableness appear to serve to more precisely convey concerns about purpose, evidential strength, and actionability. These findings suggest that traditional criteria may work as only partially adequate mediators of students' meta-talk on argument quality in stochastic contexts. Students' introduction of new terms can be interpreted as an empirical response to the limits of existing criteria in capturing the complexity of data-based inquiry and decision making. Notably, these new criteria lack stable definitions and their relation to epistemic and teleological dimensions remains fluid.

This fluidity should not be read as a deficit, but as evidence of students' engagement with the inherent complexity of statistical decision making. Moreover, the observed instability in the use of evaluative criteria might point to a broader tension between acting with criteria and reflecting on criteria as objects of discourse. The design choice to centre the discussion on the evaluation of exemplars can be seen as an attempt to prevent and mitigate this tension, by grounding evaluative talk in concrete argumentative products rather than abstract definitions. As a result, although the

meaning of evaluative criteria remains partially unstable, throughout the discussion, students frequently positioned themselves with respect to the analysed exemplars.

The study focuses on a single classroom discussion, and not all potential evaluative criteria are explored in depth, partly due to time constraints and the specific features of the exemplars discussed (e.g., none contain errors in data reading or elaboration). Moreover, the researcher's role in steering the discussion is not analysed here, even though it likely contributed to which criteria were foregrounded. Future research could examine how sustained, structured work on evaluative criteria, across multiple tasks and discussions, fosters the development of statistical rationality and students' ability to engage in explicit meta-talk about argument quality.

References

Bolstad, O. H. (2025). Mathematical literacy-What is missing?. In Bosch Casabò, M., Carreira, S., Bolondi, G., Gaidoschick, M., & Spagnolo, C. (Eds.), Proc. of the 14th Congress of the European Society for Research in Mathematics Education. ERME.

Boscolo, A., Morselli, F., & Robutti, E. (2025). Implementation of formative assessment in argumentative activities for the development of algebraic thinking: indicators and where to find them. *La Matematica e la sua Didattica 33*(2), 75–113.

Büscher, C. (2024). Adapting Habermas' construct of communicative rationality into a framework for analyzing students' statistical literacy. *Educational Studies in Mathematics, 117*(1), 121–141.

Cusi, A., Morselli, F., & Sabena, C. (2017). Promoting formative assessment in a connected classroom environment: design and implementation of digital resources. ZDM Mathematics Education, 49(5), 755–767.

Gómez-Blancarte, A., & Tobías-Lara, M. G. (2018). Using the Toulmin model of argumentation to validate students' inferential reasoning. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Proc. of the 10th International Conference on Teaching Statistics. ISI*.

Grion, V., Serbati, A., Tino, C, & Nicol, D. (2017). Ripensare la teoria della valutazione e dell'apprendimento all'università: un modello per implementare pratiche di peer review. *Giornale Italiano della Ricerca Educativa, 10*(19), 209–226.

Habermas (2003). Truth and Justification. MIT Press.

Krummenauer, J., Gutensohn, F., Aichele, J., Emhart, M., & Kuntze, S. (2023). Young primary students' difficulties related to data-based argumentation–a qualitative analysis. In P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, & E. Kónya (Eds.), *Proc. of the 13th Congress of the European Society for Research in Mathematics Education. ERME*.

Vielmo, C. (2025). Investigating students' co-construction of statistical argumentation. In Bosch Casabò, M., Carreira, S., Bolondi, G., Gaidoschick, M., & Spagnolo, C. (Eds.), *Proc. of the 14th Congress of the European Society for Research in Mathematics Education. ERME*.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.