Annals of Mathematics and Artificial Intelligence
https://doi.org/10.1007/s10472-026-10020-6

RESEARCH

Towards a structural analysis of interesting geometric theorems: analysis of a survey

Pedro Quaresma[1] · **Pierluigi Graziani**[2]

Received: 14 April 2026 / Accepted: 31 August 2026
© The Author(s) 2026

Abstract
A longstanding challenge in automated theorem generation is not merely to derive new statements, but to identify which of them are mathematically interesting. Motivated by Wos' Problem 31, this paper reports the results of a first survey, within a two-stage expert-survey study on theorem interestingness in geometry, and explores whether some of the dimensions elicited from human judgements admit formally computable structural counterparts. The descriptive analysis identifies six recurring, non-exclusive and non-exhaustive dimensions of interestingness: simplicity and clarity, surprise, conceptual depth, usefulness and applicability, aesthetic value, and historical, pedagogical, or emotional relevance. We then examine these dimensions in relation to a modern geometrographic analysis of area-method proofs, considering the overall complexity of a proof, the number and complexity of its steps, the occurrence and diversity of difficult lemmas, and the *Geometrographic Readability Coefficient of Proofs* (GRCP). The comparison suggests that simplicity, clarity, and certain aspects of conceptual depth may have plausible structural counterparts in proof features. By contrast, surprise, usefulness, aesthetic value, and historical or pedagogical relevance remain strongly dependent on mathematical and human context. The analysis is exploratory and does not establish statistical correlations or validate GRCP as a measure of theorem interestingness. Rather, it provides an empirical and formal basis for hypotheses to be tested in the second survey and in subsequent comparisons with human judgements of proof readability.

Keywords Automated deduction · Geometrography · Interestingness · Wos' Problem 31

Mathematics Subject Classification (2010) 68V15 · 03F07 · 68V30

1 Introduction

In *Automated Reasoning: 33 Basic Research Problems*, Larry Wos identified a collection of foundational challenges for automated reasoning. Among them, Problem 31 remains

✉ Pierluigi Graziani
pierluigi.graziani@uniurb.it

Pedro Quaresma
pedro@mat.uc.pt

[1] CISUC/Department of Mathematics, University of Coimbra, Coimbra, Portugal

[2] Department of Pure and Applied Science, University of Urbino, Urbino, Italy

challenging and continues to motivate active research more than three decades after its formulation [1, 2]:

> *What properties can be identified to permit an automated reasoning program to find new and interesting theorems, as opposed to proving conjectured theorems?*

This question encapsulates two conceptually distinct but tightly intertwined issues: the generation of *new* theorems and the identification of *interesting* ones. The former concerns the capacity of an automated system to produce unknown (new) statements that are derivable from a given corpus or background theory. The latter concerns the far more elusive task of discriminating, among the potentially vast set of generated consequences, those results that possess mathematical relevance or significance.

The automatic generation of new theorems has been addressed in a variety of mathematical domains and through heterogeneous methodologies. Broadly speaking, one line of research adopts a systematic exploration of extensive areas of mathematical knowledge, generating large numbers of candidate theorems with the expectation that some of them will exhibit non-trivial or valuable properties [3–9]. A different approach focuses on automated discovery within narrowly specified domains, such as locus equations [10, 11], angle theorems [12], vector-based geometric theorems [13–15], diagram-based generation [16], or the addition of a new conjectural thesis to problems in elementary geometry [17–20]. In these cases, the objective is not to solve the general problem of automated theorem generation, but to support discovery in specific mathematical settings, for instance by identifying complementary hypotheses that render a given statement true.

By contrast, the problem of *interestingness* has proven considerably more resistant. As Wos already observed, since an automated reasoning program can, in principle, be instructed to generate an arbitrarily large set of conclusions, the central challenge lies in determining which criteria should allow the system to select those results – if any – that may have a possible structural counterpart in mathematically interesting theorems [21–31]. This difficulty has been approached from multiple disciplinary perspectives, including the sociology of proof and demonstration [32–34], educational and cognitive studies on proving [35–40], the semiotics and epistemology of mathematics [41–45], and automated theorem proving itself [5, 7–9, 46–48].

Despite this extensive body of work, Wos' Problem 31 remains unresolved. Moreover, the difficulty can be sharpened by appealing to results from computability theory. In particular, one can show that there exists no algorithmic procedure capable of deciding, for an arbitrary Turing machine that generates theorems, whether it is able to produce *interesting* theorems [29]. This undecidability result has important conceptual consequences. It implies that the task of determining whether a theorem prover or generator is capable of producing interesting theorems cannot be fully automated. At best, it can be approached through heuristic, non-deterministic procedures that provide partial or context-dependent guidance. Consequently, any serious attempt to understand and evaluate interestingness in automated theorem generation must take into account the irreducible role of human expertise.

For this reason, it was proposed to complement the theoretical analysis with an empirical investigation based on expert surveys [29]. These surveys are designed to elicit criteria that mathematicians implicitly or explicitly employ when judging the interestingness of geometric theorems, and to clarify how such criteria relate to existing approaches and metrics in automated theorem generation and finding. Rather than replacing formal evaluation, this

empirical perspective aims to shed light on the human judgements that currently compensate for the intrinsic limits of algorithmic assessment.

Despite this limitation, mathematical interestingness can be studied systematically by comparing survey-based judgements with formally describable properties of proofs. The present work combines a descriptive analysis of Survey 1 (see Appendix B) with a geometrographic interpretation of selected dimensions elicited from the responses. It identifies candidate relations between those dimensions and structural parameters of area-method proofs, providing a basis for Survey 2 and subsequent theorem-level comparisons. In this way, categories such as simplicity and readability become objects of joint empirical and formal investigation.

Structure of the paper Section 2 presents the surveys. Section 3 discusses the preliminary results from the first survey. Section 4 introduces the geometrographic framework. Section 5 focuses on a geometrographic reading of selected features emerging from the survey that are related to interestingness. In Section 6 conclusions and the outline of directions for future work are presented.

2 Summary of the expert survey framework

In response to the undecidability result, Quaresma et al. [29] propose an explicitly empirical route: to study what experts *mean* by "interesting theorem" and to extract stable characteristics that can later inform evaluation and filtering pipelines for Automated Theorem Generation/Automated Theorem Finding (ATG/ATF) systems. The full survey instruments are provided in the appendices (Appendix B and Appendix C); their structure and intended role in the overall methodology are presented.

2.1 First survey: qualitative elicitation of interestingness criteria

The first survey (Appendix B) is designed as an *open-ended, expert-elicitation* instrument, aimed at collecting: (i) concrete examples of theorems judged interesting or uninteresting, and (ii) the justificatory vocabulary experts spontaneously employ when making such judgements.[1]

Part A: "Interesting and why?" (contextual usage) Participants are asked to describe situations (e.g. lectures, classes, seminars, research discussions) in which they used the adjective *interesting* when referring to a theorem in geometry, and to explain in detail why the adjective was appropriate in that case. The goal is to capture pragmatic and contextual dimensions of interestingness (pedagogical, explanatory, research-driven, historical, etc.), rather than forcing a priori definitions.

Part B: five interesting theorems (example-first elicitation) Participants are asked to list at least five geometric theorems they consider interesting. For each listed theorem, they are

[1] The survey is available as a *Google Form*, https://forms.gle/KNaHGMou9jNczhY48

invited to justify their choice by providing a set of adjectives (at least five) intended to name the characteristics that, in their view, make that theorem interesting. This section is meant to yield both a pool of candidate "interesting" theorems and an initial, bottom-up lexicon of attributed properties (e.g. generality, elegance, surprising connections, explanatory clarity, methodological fruitfulness).

Part C: five uninteresting theorems (contrast set) Symmetrically, participants are asked to list at least five geometric theorems they consider uninteresting, again accompanied by an adjective-based explanation of which properties make them uninteresting. This contrastive component is intended to control for one-sided theorising and to identify negative indicators (e.g. triviality, over-specificity, lack of conceptual insight, purely technical character, redundancy).

Outcome of survey 1 The intended output of the first survey is a structured dataset pairing *theorem examples* with *expert-elicited descriptors and rationales*. These data provide the empirical basis for: (a) clustering and consolidating descriptors into candidate dimensions of interestingness, and (b) selecting representative theorems to be used in the second survey.

2.2 Second survey: quantitative assessment of extracted characteristics

The second survey (Appendix C) is explicitly framed as a *follow-up* instrument to be designed *after* analysing Survey 1. Its purpose is to test, in a more controlled and quantifiable way, whether the candidate characteristics distilled from Survey 1 reliably track expert judgements across a set of representative theorems.

At the time of writing, Survey 2 has not yet been administered, and the present paper reports no data from it. It is included here to make explicit the planned second stage of the research design, through which the candidate dimensions identified in Survey 1 will be tested in a more controlled and quantitative setting.

To address the limitations of Survey 1, the design of Survey 2 will deliberately include contrastive profiles, such as theorems with comparatively readable proofs that may nevertheless be judged routine or uninteresting, and theorems regarded as conceptually significant despite proofs of lower geometrographic readability. Participants will be required to assess both positive and negative cases. This design is intended to clarify whether readability contributes to interestingness and, if so, under which conditions.

Item structure: theorem + candidate characteristics Each item presents a theorem (selected from Survey 1 responses) together with a list of candidate characteristics (also extracted from Survey 1). Participants are first asked for a coarse-grained judgement (e.g. whether the theorem is interesting), and are then asked to rate the relevance of each proposed characteristic via Likert-type agreement scales (e.g. from strongly disagree to strongly agree) that link interestingness to specific features.

Rationale Once administered, Survey 2 will operationalise the transition from a qualitative space of reasons and descriptors to a set of testable and repeatable associations between theorem stimuli and characteristic attributions. It will enable analyses of inter-rater agreement, the relative weight of different characteristics, their stability across sub-communities, and their possible association with geometrographic parameters.

3 Analytical report on the first survey responses

This section presents an analytical account of the 20 responses collected[2] through the first expert survey administered via Google Forms. The analysis is grounded in the raw textual answers, and a summary of the responses is available on the project website. Given the qualitative and exploratory nature of the survey, the goal of this section is not to introduce an operational definition of interestingness, but to describe how respondents articulate and motivate their judgements when referring to geometric theorems as *interesting*.

3.1 Participants and recruitment

Participation was voluntary and responses were anonymous. The survey was circulated through seminars, conferences, and academic networks connected with automated deduction, especially automated deduction in geometry, as well as through Master's programmes in mathematics education. No formal inclusion or exclusion criteria were applied. Because the questionnaire did not collect respondent-level demographic or professional information, the exact composition of the sample cannot be reconstructed. On the basis of the recruitment channels, the respondent pool likely included researchers working in automated deduction and geometry, together with Master's students in mathematics education. This characterisation should therefore be regarded as approximate. The resulting group is a small, self-selected convenience sample. It is not intended to be statistically representative of mathematicians, geometry researchers, or mathematics educators. The purpose of the survey is exploratory: to elicit examples, descriptors, and reasons that may provide a preliminary basis for subsequent, more controlled investigations.

3.2 Structure and coverage of the responses

The survey consists of two main types of open-ended prompts. The first type concerns situations in which participants use the adjective *interesting* in connection with geometric theorems. The second type asks participants to list geometric theorems they consider interesting, often accompanied by short explanations.

All 20 participants answered the first situational question and provided at least one example of an interesting theorem. Subsequent situational questions were answered less frequently, whereas the theorem-listing prompts exhibit almost complete coverage across all five items. This distribution suggests that respondents were more systematic and confident when providing examples of interesting theorems than when providing multiple contextual descriptions of usage.

A significant asymmetry emerged between the positive and contrastive parts of the instrument. Although participants supplied examples of interesting theorems, the section asking for theorems considered *uninteresting* yielded no usable theorem entries and accompanying justifications. Survey 1 therefore does not provide a contrast class against which the positive cases can be systematically compared. This absence constitutes an important limitation. In particular, the present data cannot establish which properties distinguish an interesting theorem from one that is merely simple or readable. The analysis reported below is consequently restricted to the reasons given in support of positive judgements.

[2] On 2026/03/24. The survey data, in CSV format, and an analytical summary are available on the project webpage, https://geotiles.mat.uc.pt/GeoTiles/Projects/SimpleProofs/simpleProofs.php.

3.3 Situational use of the adjective *interesting*

The situational responses describe concrete episodes drawn from professional practice. A substantial portion of the answers explicitly refers to educational contexts, such as school lessons or university courses, in which a theorem is described as interesting because it supports explanation, visualisation, or conceptual understanding. Other responses describe moments of personal discovery or study, often linked to surprise, unexpected connections, or a deeper appreciation of a familiar result.

Across the responses, interestingness is consistently framed as something that emerges in use: a theorem is called interesting when it plays a specific role in explanation, learning, or insight, rather than in virtue of an abstract property attributed independently of context. Later situational questions contribute additional individual cases, but their limited response rate and heterogeneity do not support further aggregation.

3.4 Theorems cited as interesting

Across the theorem-related prompts, participants listed a wide and heterogeneous set of geometric theorems. Certain classical results recur under different linguistic formulations, most notably variants of the Pythagorean theorem and of Thales' theorem. Beyond these recurrent cases, the distribution of cited theorems is highly dispersed. Most theorems are mentioned by a single participant and range from elementary geometry to configuration theorems, projective results, and, in a few instances, well-known theorems outside geometry.

This dispersion indicates that the survey does not single out a canonical set of interesting geometric theorems. Rather, it reflects the diversity of mathematical backgrounds and perspectives represented in the respondent pool, while still exhibiting common modes of justification.

Among the 47 individual theorem entries analysed, the most frequently mentioned were:

- Pythagorean Theorem (13 mentions): associated with simplicity and clarity, usefulness and applicability, historical, pedagogical, or emotional relevance, and the availability of multiple proofs;
- Thales' Theorem (7 mentions): associated with usefulness and applicability;
- Euler's line (3 mentions): with no specific characteristics associated;
- Ptolemy's Theorem (3 mentions): associated with capacity to surprise and usefulness and applicability;
- Ceva's Theorem (2 mentions): associated with simplicity and clarity, capacity to surprise, conceptual depth, and usefulness and applicability.

3.5 Qualitative coding procedure

The six categories discussed below were obtained through an exploratory consolidation of the descriptors and explanations contained in the raw responses. An initial draft of the analytical summary and a provisional grouping of semantically related descriptors were produced with the assistance of GPT-5.4 Thinking. The authors subsequently compared the proposed groupings with the raw responses, corrected misclassifications, revised the boundaries between categories, and consolidated the final descriptive scheme. No category or individual coding decision was accepted automatically, and responsibility for the final analysis and interpretation remains entirely with the authors.

This procedure should be regarded as exploratory rather than as a fully developed qualitative-coding protocol. The responses were not coded independently by multiple human

Table 1 Descriptive frequencies of the categories identified in the justifications provided for interesting theorems. The denominator is the total number of codable descriptors (88), from the 47 individual theorem entries analysed. Categories are non-exclusive and non-exhaustive

Category	Illustrative descriptors	*n*	%
Simplicity and clarity	simple, clear, accessible, easy to state or prove	11	12.50%
Capacity to surprise	surprising, unexpected, counter-intuitive, non-obvious	7	7.95%
Conceptual depth	deep, revealing, connecting different concepts or domains	10	11.36%
Usefulness and applicability	useful, applicable, fruitful in problem solving or teaching	21	23.86%
Aesthetic value	beautiful, elegant, symmetric, harmonious	5	5.68%
Historical, pedagogical, or emotional relevance	historically important, memorable, educationally significant, personally meaningful	8	9.09%

coders, and no measure of inter-coder agreement was calculated. Future analyses should use a predefined coding manual, independent coders, and an assessment of coding agreement.

3.6 Justifications and descriptive synthesis

Many theorem entries are accompanied by explicit justifications, which vary in style and emphasis. For the descriptive coding reported below, the unit of analysis was a theorem entry accompanied by a non-empty justification. Semantically related descriptors were consolidated into six non-exclusive categories. Since a single entry could be assigned to more than one category, the frequencies reported in Table 1 are not mutually exclusive and their percentages do not sum to 100%. More detailed information, including the raw responses and the analytical summary, is available on the project website.

These frequencies add a descriptive quantitative layer to the qualitative analysis and characterise the responses collected in this small, self-selected sample. The six categories provide a preliminary vocabulary for the ways in which participants justified their judgements; they are not presented as a population-level or generalisable structure of mathematical interestingness.

This vocabulary provides the empirical starting point for Section 5, where selected categories are compared with structural features of geometric proofs. Section 4 first introduces the geometrographic framework used for this comparison by Quaresma and Graziani [30].

4 Geometrography, simplicity and readability of synthetic geometric proofs

Geometrography is a mathematical framework originally introduced by Lemoine [25] to analyse and compare ruler-and-compass geometric constructions. Its central idea is that a construction can be decomposed into a finite sequence of elementary operations, and that the difficulty of a construction can be quantified by counting and classifying such operations. Mackay [26] later refined this approach by systematising the set of primitive ruler and compass

actions and clarifying their respective roles in determining the complexity and accuracy of a construction.

4.1 Geometrography of geometric constructions

Following Lemoine's original proposal and Mackay's refinements, the elementary operations considered in ruler-and-compass geometry are the following:[3]

To place the edge of the ruler in coincidence with one point ..... R_1
To place the edge of the ruler in coincidence with two points ..... $2R_1$
To draw a straight line ..... R_2
To put one point of the compass on a determinate point ..... C_1
To put the points of the compass on two determinate points ..... $2C_1$
To describe a circle ..... C_2

A geometric construction can therefore be represented as a linear combination

$$l_1R_1 + l_2R_2 + m_1C_1 + m_2C_2,$$

where the coefficients count the number of occurrences of the corresponding elementary operations.

From this representation, two fundamental coefficients are defined. The *coefficient of simplicity*, given by $l_1 + l_2 + m_1 + m_2$, measures the total number of elementary operations required to perform the construction. The *coefficient of exactitude*, given by $l_1 + m_1$, counts the preparatory positioning operations on which the accuracy of the construction depends when it is carried out with physical instruments. Geometric simplicity is thus understood as procedural economy.

4.2 A modernisation of geometrography

Quaresma and Graziani [30] extended this geometrographic framework from informal geometric construction to entire synthetic geometric construction and proofs, produced by a geometric automated theorem prover (GATP), *GCLC*, using the area method [49].

Classical geometrography applies to geometric constructions made with the help of a ruler and a compass. Its modernisation uses the tools of the dynamic geometry systems (DGS). In [27], it was shown how to modernise geometrography using *GCLC*; in [50], the generality of the approach is shown using *GeoGebra* [51].

Consider two basic types of operation: defining a free point anywhere in the plane, denoted by D, and defining a new object from previously defined objects, denoted by C. Each occurrence of C represents one previously defined object used as input. On this basis, the following values are assigned to the basic *GCLC* constructions:

The value assigned to a construction is determined by the number of previously defined objects required as inputs, rather than by the geometric type of the construction. Consequently, geometrically different constructions may receive the same value when they depend on the same number of input objects. For example, a line determined by two points and an intersection point determined by two previously defined lines are both assigned the value $2C$. If, by contrast, the intersection is specified by four points — two points determining each of the two lines — the corresponding value is $4C$.

[3] A compass defines a circle by using two points: the needle point (the centre of the circle) and the pencil point (the radius).

`point` — fix a point in the plane ..... D
`line` — uses two points ..... $2C$
`circle` — uses two points ..... $2C$
`intersec` — uses two lines ..... $2C$
`intersec` — uses four points ..... $4C$
`intersec2` — uses a circle and a circle or line ..... $2C$
`midpoint` — uses two points ..... $2C$
`med` — uses two points ..... $2C$
`bis` — uses three points ..... $3C$
`perp` — uses a point and a line ..... $2C$
`foot` — uses a point and a line ..... $2C$
`parallel` — uses a point and a line ..... $2C$
`onsegment` — uses two points ..... $2C$
`online` — uses two points ..... $2C$
`oncircle` — uses two points ..... $2C$

In the modernisation of Geometrography, the tools provided by dynamic geometry systems replace physical ruler-and-compass operations. Consequently, the *coefficient of exactitude* loses its original meaning, since the constructions are executed computationally by the DGS and are therefore exact. The *coefficient of simplicity* (CS_{gcl}), however, remains useful for classifying constructions according to their relative levels of simplicity. A further dimension is provided by the *coefficient of freedom* (CF_{gcl}), which records the number of degrees of freedom of a geometric object. For example, a point constrained to a line has one degree of freedom, whereas a free point in the plane has two. This coefficient provides a quantitative description of the dynamism of the geometric construction. Degrees of freedom are measured with reference to point definitions. The `point` construction defines a point with two degrees of freedom, whereas the `onsegment`, `online`, and `oncircle` constructions define points with one degree of freedom.

For the *GCLC* constructions contained in the *Thousand of Geometric problems for Geometric Theorem Provers* (*TGTP*) repository,[4] an average value of simplicity (CS_{gcl}) of 20.8 was obtained. Using the k-means clustering function implemented in the statistics package of *Octave*,[5] three classes of geometric constructions describing an increasing level of complexity were defined: simple constructions, $1 \leq CS_{gcl} \leq 18$; average complexity constructions, $18 < CS_{gcl} \leq 28$; complex constructions, $CS_{gcl} > 28$. *TGTP* contains 71 simple constructions; 81 average complexity constructions; 28 complex constructions.

From measuring the simplicity of geometric constructions, Quaresma and Graziani [30] moved on to measuring proofs. Given the availability of an extensive study of the area-method, including rigorous proofs of all its lemmas [52], that method and its implementation by *GCLC* were chosen.[6] An area-method proof consists of several types of steps: geometric constructions, elementary algebraic simplifications, elementary geometric simplifications, and applications of the area-method lemmas.

[4] TGTP server at: https://geotiles.mat.uc.pt/GeoTiles/Projects/TGTP/

[5] GNU Octave, version 6.1.1, package octave-statistics, function kmeans https://octave.sourceforge.io/statistics/function/kmeans.html

[6] The proofs developed by GATPs based on the area-method are formal proofs. The method itself was formalised, and proved sound, using the *Coq* proof assistant. The GATP developed by J. Narboux, as a *Coq* tactic, can have the proofs verified by *Coq*. The *GCLC* implementation of the area method does not explicitly provide that possibility; however, it could be achieved by developing a translator from the *GCLC* language to the *Coq* language.

Apart from the geometric constructions on which the proof is based (with coefficient of simplicity $n\mathbf{Cnst}$), there are other steps to be considered.

(Elementary) Algebraic Simplification ..... (**AS**)
(Elementary) Geometric Simplification ..... (**GS**)
Application of the Area-method Lemma n ..... ($\mathbf{AML_n}$)

Elementary algebraic and geometric simplifications are treated as trivial steps and are counted with unit weight. Applications of lemmas, by contrast, are weighted according to the geometrographic simplicity coefficient associated with each lemma.[7]

A proof can thus be measured by counting the occurrences of the different types of steps. Let m be the total number of lemma applications in the proof. For each $\ell \in \{1, \ldots, m\}$, let j_ℓ identify the lemma used in the ℓ-th application and let r_ℓ denote its occurrence number among applications of the same lemma. The proof can then be represented by an expression of the form

$$n_1\mathbf{Cnst} + n_2\mathbf{AS} + n_3\mathbf{GS} + \sum_{\ell=1}^{m} \mathbf{AML}_{(\mathbf{j}_\ell, \mathbf{r}_\ell)}.$$

The *simplicity coefficient* of the proof is defined as

$$\mathrm{CS}_{\mathrm{proof}} = n_1 + n_2 + n_3 + \sum_{l=1}^{m} w_{(j_\ell, r_\ell)},$$

where the weight $w_{(j,r)}$ assigned to the r-th application of lemma j, is given by

$$w_{(j,r)} = \begin{cases} \mathrm{CS}_{\mathrm{proof}}(\mathbf{AML_j}), & r = 1 \\ \mathrm{CS}_{\mathrm{gcl}}(\mathbf{AML_j}), & r > 1 \end{cases}$$

$\mathrm{CS}_{\mathrm{proof}}(\mathbf{AML_j})$ denotes the full proof-simplicity coefficient of the lemma, whereas $\mathrm{CS}_{\mathrm{gcl}}(\mathbf{AML_j})$ denotes the constructional coefficient associated with its geometric configuration. For the purposes of the present complexity measure, the first occurrence of a lemma is assigned its full proof-simplicity coefficient. This does not imply that the lemma is literally reproved within the object-level derivation. Rather, the full proof-theoretic cost associated with that lemma type is charged when it first enters the complexity account of the proof. If the same lemma is applied again later in the proof, its full proof-simplicity coefficient is not counted a second time. At that stage, the inferential schema supplied by the lemma is treated as already available. What remains to be accounted for is its instantiation in the new geometric configuration, namely the matching of the variables and geometric objects occurring in the lemma with the objects of the current proof. Consequently, from the second application onward, only the corresponding constructional coefficient $\mathrm{CS}_{\mathrm{gcl}}$ is counted.

In this proof-theoretic setting, the coefficient of exactitude has no meaning, since proofs are symbolic objects and do not depend on the physical accuracy of drawings.

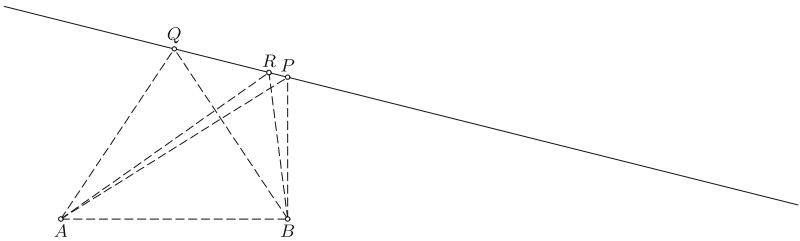
For example, the proof of Lemma 9 will have the following coefficient of simplicity, $\mathrm{CS}_{\mathrm{proof}}(\mathbf{AML_9}) = 74$.

[7] By an elementary algebraic simplification we mean an application of the basic operations of addition, subtraction, multiplication, or division, together with standard algebraic laws such as commutativity, associativity, and distributivity. By an elementary geometric simplification we mean a direct application of the definitions of the area-method quantities. Both kinds of simplification are counted as trivial steps.

Lemma 9 (**AML$_9$**) *Let R be a point on the line PQ. Then for any two points A and B it holds that $\mathcal{S}_{RAB} = \frac{\overline{PR}}{\overline{PQ}}\mathcal{S}_{QAB} + \frac{\overline{RQ}}{\overline{PQ}}\mathcal{S}_{PAB}$.*

The following is a shorter version of its proof with the elementary algebraic and geometric simplifications condensed (the expanded version can be seen in [28]).[8]

Geometrography of Lemma 9 (**AML$_9$**)



$\text{CS}_{\text{gcl}} = 22 = 4\mathbf{D} + 18\mathbf{C}$ (= 4, free points + 18, constructed points and lines)

$\text{CF}_{\text{gcl}} = 8$ (= 4, free points × 2, in the plane)

- $s = \mathcal{S}_{ABPQ}$, initial construction;
- $1 \times$ **GS**, areas of triangles with the same orientation, $\mathcal{S}_{RAB} = s - \mathcal{S}_{ARQ} - \mathcal{S}_{BPR}$;
- $1 \times$ **AML**$_{(14,1)}$, lemma 14, $\frac{\overline{PR}}{\overline{PQ}} = r$ (CS_{proof}(**AML$_{14}$**) = 8);
- $1 \times$ **AML**$_{(5,1)}$, lemma 5, $\frac{\mathcal{S}_{ARQ}}{\mathcal{S}_{APQ}} = \frac{\overline{RQ}}{\overline{PQ}}$ (CS_{proof}(**AML$_5$**) = 18);
- $1 \times$ **GS**, segments with the same orientation, $\frac{\overline{RQ}}{\overline{PQ}} = \frac{\overline{PQ}-\overline{PR}}{\overline{PQ}}$;
- $2 \times$ **AS**, algebraic simplifications, $\frac{\overline{PQ}-\overline{PR}}{\overline{PQ}} = (1-r)$ and $\mathcal{S}_{ARQ} = (1-r)\mathcal{S}_{APQ}$;
- $1 \times$ **AML**$_{(5,2)}$, lemma 5, $\frac{\mathcal{S}_{BPR}}{\mathcal{S}_{BPQ}} = \frac{\overline{PR}}{\overline{PQ}}$ (CS_{gcl}(**AML$_5$**) = 11);
- $2 \times$ **AS**, algebraic simplifications, $\mathcal{S}_{BPR} = r\mathcal{S}_{BPQ}$ and $\mathcal{S}_{RAB} = s - (1-r)\mathcal{S}_{APQ} - r\mathcal{S}_{BPQ}$;
- $2 \times$ **GS**, areas of triangles with the same orientation, $\mathcal{S}_{RAB} = s - (1-r)(s - \mathcal{S}_{PAB}) - r(s - \mathcal{S}_{QAB})$;
- $7 \times$ **AS**, algebraic simplification, $\mathcal{S}_{RAB} = s - s + rs + \mathcal{S}_{PAB} - r\mathcal{S}_{PAB} - rs + r\mathcal{S}_{QAB}$, $\mathcal{S}_{RAB} = r\mathcal{S}_{QAB} + (1-r)\mathcal{S}_{PAB}$ and $\mathcal{S}_{RAB} = \frac{\overline{PR}}{\overline{PQ}}\mathcal{S}_{QAB} + \frac{\overline{RQ}}{\overline{PQ}}\mathcal{S}_{PAB}$;

Geometrography for the demonstration: $4\mathbf{D} + 18\mathbf{C} + 4\mathbf{GS} + 11\mathbf{AS} + 1\mathbf{AML_{14}} + 2\mathbf{AML_5}$

$$\mathbf{AML_9} \begin{cases} \text{CS}_{\text{proof}} = 74 = 22 + 4 + 11 + 8 + (18 + 11) \\ \text{CS}_{\text{gcl}} = 22 \end{cases}$$

with $w_{14,1} = \text{CS}_{\text{proof}}(\mathbf{AML_{14}}) = 8$, $w_{5,1} = \text{CS}_{\text{proof}}(\mathbf{AML_5}) = 18$, and $w_{5,2} = \text{CS}_{\text{gcl}}(\mathbf{AML_5}) = 11$.

In this proof, **AML$_5$** is applied twice. Its first occurrence is assigned the full proof-simplicity coefficient $\text{CS}_{\text{proof}}(\mathbf{AML_5}) = 18$. At its second occurrence, the lemma's

[8] $\mathcal{S}_{ABC}$, the *Signed Area* for a triangle ABC, $\frac{\overline{AB}}{\overline{CD}}$, the *Ratio of Parallel Directed Segments*, and $\mathcal{P}_{ABC}$, the *Pythagoras Difference* for the points A, B, C, are specific geometric quantities of the area method, for treating arrangement relations.

inferential schema is already available, and only the constructional cost of instantiating it in the new configuration is counted: $\mathrm{CS}_{\mathrm{gcl}}(\mathbf{AML_5}) = 11$. This convention explains why the two applications of $\mathbf{AML_5}$ contribute different values to the calculation of $\mathrm{CS}_{\mathrm{proof}}(\mathbf{AML_9})$. The single application of $\mathbf{AML_{14}}$ contributes its full proof-simplicity coefficient, $\mathrm{CS}_{\mathrm{proof}}(\mathbf{AML_{14}}) = 8$.

4.3 Structural parameters for proofs

Given that a mathematical proof is a sequence of steps, in addition to the coefficient of simplicity, it would be useful to have other coefficients: e.g., the total number of steps in the proof; the value of the most difficult step in the proof; the number of different steps of high difficulty in the proof; the number of different types of steps (lemmas) in the proof; a proof script; a numerical description of the proof; and a corresponding line chart or proof trace.

Therefore, to fully characterise a formal synthetic proof produced by a GATP, the following coefficients can be defined:

- $\mathrm{CS}_{\mathrm{proof}}$, the simplicity coefficient (as above); it gives the simplicity coefficient for the overall proof;
- $\mathrm{CT}_{\mathrm{proof}}$, the total number of steps in the proof;
- $\mathrm{CS}_{\mathrm{proofmax}}$, the highest simplicity coefficient of the lemmas/definitions applications; it gives the simplicity coefficient for the most difficult step of the proof;
- $\mathrm{CD}_{\mathrm{typeproof}}$, the number of different types of lemmas used in the proof;
- $\mathrm{CD}_{\mathrm{highproof}}$, the number of different steps of high difficulty in the proof.

and, to complete this information:

- the proof script, as defined above;
- the corresponding line chart or proof trace in *TikZ* format.[9]

To compute $\mathrm{CD}_{\mathrm{highproof}}$, the area-method lemmas implemented in *GCLC* were analysed. Let hp denote the proof-simplicity coefficient assigned to a lemma. Using the k-means clustering function in the *Octave* statistics package, the lemmas were divided into three categories: low difficulty ($hp < 284$), medium difficulty ($284 \leq hp < 1848$), and high difficulty ($hp \geq 1848$).

Using the defined coefficients above, the following values for the proof of $\mathbf{AML_9}$ can be found:

$$\mathbf{AML_9} \begin{cases} \mathrm{CS}_{\mathrm{proof}} = 74 = 22 + 4 + 11 + 8 + (18 + 11) \\ \mathrm{CS}_{\mathrm{gcl}} = 22 \\ \mathrm{CT}_{\mathrm{proof}} = 19 \\ \mathrm{CS}_{\mathrm{proofmax}} = 18 \\ \mathrm{CD}_{\mathrm{typeproof}} = 2 \\ \mathrm{CD}_{\mathrm{highproof}} = 0 \end{cases}$$

Using the *GCLC* implementation of the area method it is possible to produce the desired proof scripts and line chart or proof trace.

As an example, consider `GEO0001` from the *TGTP* repository, which formalises Ceva's theorem.

[9] https://tikz.net

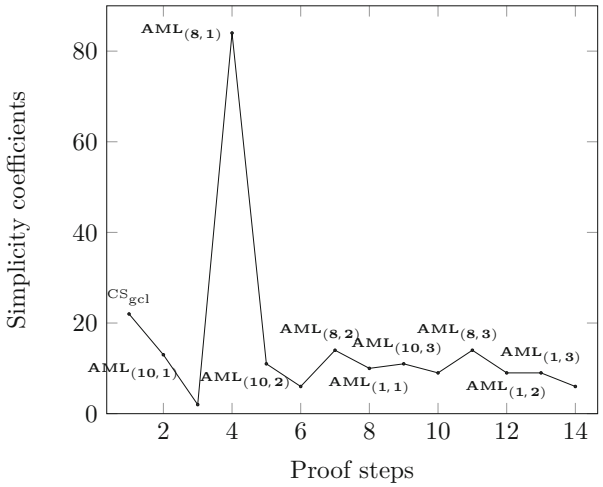
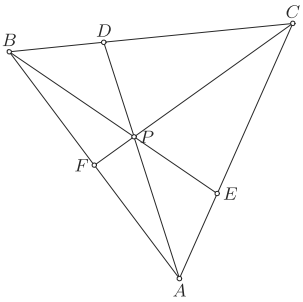


Fig. 1 Geometrographic proof trace of the *GCLC* area-method proof of Ceva's theorem. The labels identify the area-method lemmas applied, while the vertical coordinates give the occurrence-specific weights assigned to the represented proof items

Theorem 1 (Ceva's Theorem) *Let ΔABC be a triangle and P be any point in the plane. Let $D = AP \cap CB$, $E = BP \cap AC$, and $F = CP \cap AB$. Show that: $\frac{\overline{AF}}{\overline{FB}} \times \frac{\overline{BD}}{\overline{DC}} \times \frac{\overline{CE}}{\overline{EA}} = 1$. The point P must not lie on any of the lines parallel to AC, AB, and BC and passing through B, C, and A, respectively [49].*



When *GCLC* is set to display the proof at the fullest level of detail, the proof script of Ceva's theorem occupies four pages, because the individual algebraic steps are reported rather than condensed; see Appendix A. Figure 1 visualises the geometrographic proof trace of the *GCLC* area-method proof of Ceva's theorem. The horizontal axis records the ordered sequence of the proof items represented in the trace, while the vertical axis reports the geometrographic weight assigned to each item. The initial point represents the constructional coefficient $\mathrm{CS}_{\mathrm{gcl}}$. Consecutive algebraic or geometric simplifications are grouped into single points in order to provide a more compact representation of the proof. The labels $\mathbf{AML}_{(\mathbf{1},\mathbf{r})}$, $\mathbf{AML}_{(\mathbf{8},\mathbf{r})}$, and $\mathbf{AML}_{(\mathbf{10},\mathbf{r})}$ identify the r-th application of the corresponding area-method lemma. Peaks in the trace therefore identify proof items with comparatively higher assigned geometrographic weights; they should not, by themselves, be interpreted as measures of cognitive difficulty.

Therefore, the geometrography of the proof of Ceva's theorem is the following: $4\mathbf{D} + 18\mathbf{C} + 23\mathbf{AS} + 3\mathbf{AML}_{(\mathbf{1},\mathbf{r})} + 3\mathbf{AML}_{(\mathbf{8},\mathbf{r})} + 3\mathbf{AML}_{(\mathbf{10},\mathbf{r})}$.

$$\textbf{Ceva's Theorem} \begin{cases} \mathrm{CS}_{\mathrm{proof}} = 220 \\ \mathrm{CS}_{\mathrm{gcl}} = 22 \\ \mathrm{CT}_{\mathrm{proof}} = 33 \\ \mathrm{CS}_{\mathrm{proofmax}} = 84 \\ \mathrm{CD}_{\mathrm{typeproof}} = 3 \\ \mathrm{CD}_{\mathrm{highproof}} = 0 \end{cases}$$

4.4 From complexity to a structural proxy for readability

Building on these parameters, Quaresma and Graziani [30] introduced the *Geometrographic Readability Coefficient of Proofs* (GRCP) as a formally defined coefficient intended to estimate structural aspects of the readability of area-method proofs. The coefficient combines formally computable features of proof structure, including the weighted complexity of non-trivial steps, the occurrence of high-difficulty lemma applications, and the diversity of the lemmas employed. It is defined as follows:

$$\mathrm{GRCP} = (\mathrm{CS}_{\mathrm{proof}} - \mathrm{CT}_{\mathrm{proof}})(\mathrm{CD}_{\mathrm{highproof}} + \mathrm{CD}_{\mathrm{typeproof}}).$$

The first factor, $(\mathrm{CS}_{\mathrm{proof}} - \mathrm{CT}_{\mathrm{proof}})$, approximates the contribution of non-trivial steps in the proof by removing the unit-weighted steps counted in $\mathrm{CT}_{\mathrm{proof}}$ from the weighted simplicity measure $\mathrm{CS}_{\mathrm{proof}}$. This factor subtracts a unit baseline for each proof item from the total weighted coefficient. It therefore represents the aggregate excess weight of the proof items above that baseline. The second factor, $(\mathrm{CD}_{\mathrm{highproof}} + \mathrm{CD}_{\mathrm{typeproof}})$, captures the impact of structurally difficult steps by combining the number of high-difficulty lemma applications with the number of distinct lemmas used.

In the original study, the GRCP was used to classify a corpus of area-method proofs into high-, medium-, and low-readability classes. However, as explicitly acknowledged in that study, the coefficient has not been empirically validated against human readability judgements. Accordingly, in the present paper we treat it as a formally defined structural proxy for proof readability, rather than as a validated measure of the actual ease with which human readers understand a proof.

Considering 71 theorems from the *TGTP* repository and their *GCLC* area-method proofs, and using, again, the k-means clustering function from *Octave*, the proofs can be classified into the following classes of geometrographic readability:[10] readable (high-readability), $GRCP \leq 48000$; medium-readability $48000 < GRCP \leq 135000$; low-readability, $GRCP > 135000$.

For the *GCLC* area-method proof of `GEO0001` (Ceva's theorem), the GRCP value is:

$$\mathrm{GRCP}_{\mathrm{GEO0001}} = (220 - 33)(0 + 3) = 561.$$

Since $561 \leq 48000$, the proof falls within the geometrographic class labelled *high-readability*.

[10] The actual values were rounded for better (textual) readability.

5 An exploratory geometrographic interpretation of survey-based interestingness

Having introduced the geometrographic framework, we now examine whether selected dimensions identified in Survey 1 can be related to formally computable features of the available area-method proofs. The survey categories and the geometrographic parameters operate at different levels: the former express reasons for judging theorems interesting, whereas the latter describe the structure of particular proofs. The aim is therefore to formulate theorem-level hypotheses that can guide Survey 2 and subsequent comparative analyses. Some dimensions may admit plausible structural counterparts; others — notably historical, pedagogical, aesthetic, and emotional dimensions — may not.

Proof readability and theorem interestingness are conceptually distinct. Readability may contribute to interestingness in some contexts, but it is neither necessary nor sufficient: routine theorems may have readable proofs, whereas conceptually or historically important theorems may have difficult ones. Accordingly, GRCP is treated here as a formally defined structural indicator rather than as an empirically validated measure of human readability or theorem interestingness.

The examples included in the geometrographic comparison were not selected solely on the basis of their frequency in the survey. The comparison is constrained by the intersection between the theorems mentioned by respondents and those for which a complete *GCLC* area-method proof and a computed geometrographic profile are currently available. For transparency, Table 2 also reports frequently cited survey cases for which no corresponding profile is available. Missing values are explicitly marked rather than replaced by estimates. This makes it possible to distinguish the prominence of a theorem in the survey from its availability for the present structural analysis.

Because Survey 1 yielded no usable entries in the section asking participants to identify uninteresting theorems, Table 2 contains no contrast group. This is a limitation of the present analysis rather than evidence that the selected theorems lack meaningful negative counterparts. Survey 2 will therefore need to elicit contrastive cases explicitly.

Against this methodological background, we now consider each of the six survey-based dimensions in turn. For each dimension, we indicate its frequency in Survey 1, identify any plausible geometrographic counterpart, and formulate either a theorem-level hypothesis for Survey 2 or a limitation of the structural approach.

Simplicity and clarity Descriptors grouped under simplicity and clarity occurred 11 times among the 88 coded descriptors (12.5%). The most direct geometrographic candidates for comparison with this dimension are CS_{proof}, $CS_{proofmax}$, and GRCP: together they describe overall weighted complexity, the weight of the most demanding lemma application, and the combined structural profile of the proof. This suggests the hypothesis that proofs with lower values on these parameters are more likely to be judged simple or easy to follow.

Ceva's theorem provides an illustrative case. It was mentioned twice in the survey, including once with a simplicity-and-clarity descriptor, and its area-method proof has a GRCP of 561, placing it in the low-GRCP class. Survey 2 will test whether this pattern extends across theorem-level judgements.

Capacity to surprise Descriptors grouped under surprise or counter-intuitiveness occurred 7 times among the 88 coded descriptors (7.95%). Proof traces make visible local discontinuities in structural weight, such as isolated peaks corresponding to unusually costly lemma applications. This motivates a theorem-level question for Survey 2: do steps located at such

Table 2 Survey information and available geometrographic profiles for selected theorems

Theorem	Mentions	Dominant descriptors	CS_{proof}	CT_{proof}	$CS_{proofmax}$	$CD_{typeproof}$	$CD_{highproof}$	GRCP	Class
Pythagorean theorem	13	A(×1); D(×7); F(×2); G(×5)	–	–	–	–	–	–	–
Thales' theorem	7	D(×1)	854	54	287	6	0	4,800	High-readability
Euler's line	3	—	14711	2491	2807	15	4	232,180	Low-readability
Ptolemy's theorem	3	B(×2); D(×2)	–	–	–	–	–	–	–
Ceva's theorem	2	A(×1); B(×1); C(×1); D(×2)	220	32	84	3	0	561	High-readability

The Pythagorean theorem and Ptolemy's theorem cannot be proved by GCLC using the area method. Dominant descriptors: simplicity and clarity (A); capacity to surprise (B); conceptual depth (C); usefulness and applicability (D); aesthetic value - beautiful (E); historical, pedagogical, or emotional relevance (F); many proofs (G)

discontinuities coincide with steps that readers identify as unexpected or conceptually decisive?

Conceptual depth Descriptors grouped under conceptual depth occurred 10 times among the 88 coded descriptors (11.36%). The parameters $CD_{typeproof}$ and $CD_{highproof}$ describe, respectively, lemma diversity and the presence of high-difficulty steps. They therefore support a testable question: are proofs with greater lemma diversity, or with distinctive concentrations of high-difficulty steps, more frequently associated with judgements of conceptual depth?

Usefulness and applicability Descriptors grouped under usefulness and applicability occurred 21 times among the 88 coded descriptors (23.86%). Constructional simplicity offers a candidate structural variable because CS_{gcl} measures the economy of the geometric configuration used in the available formalisation. A theorem-level comparison can test whether lower constructional complexity is associated with perceived reusability in problem solving or teaching.

Aesthetic value Aesthetic descriptors, including beauty, elegance, symmetry, and harmony, occurred 5 times among the 88 coded descriptors (5.68%). No geometrographic coefficient directly measures aesthetic value, making this category a boundary case for the structural approach. Survey 2 may nevertheless test whether aesthetic judgements co-vary with simplicity, economy, or regularity in proof profiles.

Historical, pedagogical, or emotional context Historical, pedagogical, or emotional considerations occurred 8 times among the 88 coded descriptors (9.09%). These dimensions have no direct geometrographic counterpart because they concern the theorem's history, use, reception, or personal significance rather than the structure of its proof. They therefore identify a clear boundary of geometrographic analysis.

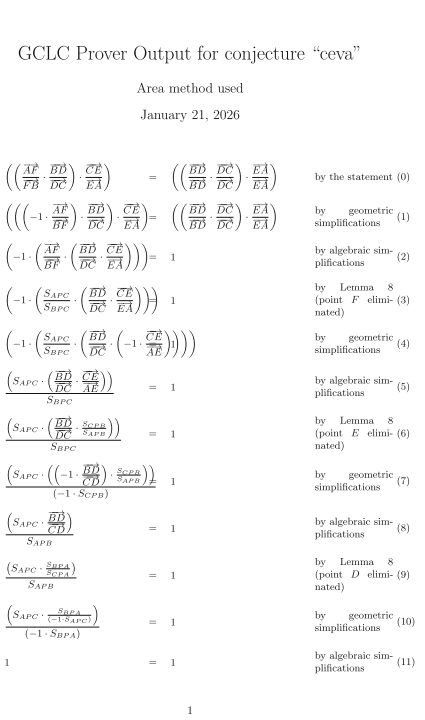
6 Conclusions and future work

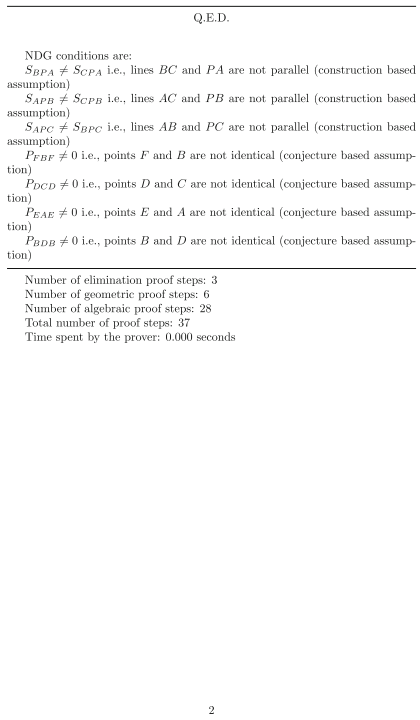
In this paper, we have addressed Wos' Problem 31 from a combined theoretical, empirical, and structural perspective. Starting from the observation that the notion of an *interesting* theorem cannot be captured by purely algorithmic means, we have argued that any adequate analysis of interestingness in automated theorem generation and theorem finding must explicitly take into account the role of human judgement. From a theoretical standpoint, the undecidability result establishes a clear limitation: there is no general algorithmic procedure capable of deciding whether a theorem-generating system can produce interesting theorems. This result does not undermine the relevance of formal evaluation. Rather, it motivates the development of complementary approaches capable of clarifying how interestingness is assessed in mathematical practice. To this end, we presented a descriptive analysis of an exploratory survey on interesting geometric theorems. The analysis identified six recurring, non-exclusive, and non-exhaustive dimensions in the participants' justifications: simplicity and clarity; surprise; conceptual depth; usefulness and applicability; aesthetic value; and historical, pedagogical, or emotional relevance. The central contribution of the paper is the formulation of a structural research programme connecting selected survey-based dimensions with formally computable properties of proofs. Geometrographic parameters such as overall proof complexity, maximum step complexity, lemma diversity, the occurrence of high-difficulty steps, and the

Geometrographic Readability Coefficient of Proofs (GRCP) provide a basis for formulating theorem-level hypotheses concerning simplicity, surprise, conceptual depth, and usefulness. At the same time, the analysis shows that some dimensions of interestingness, especially aesthetic, historical, pedagogical, and emotional relevance, cannot be inferred from proof structure alone. The proposed correspondences should therefore be understood as hypotheses rather than established empirical relations. The present study does not demonstrate statistical associations between geometrographic parameters and perceived interestingness, nor does it validate GRCP as a measure of human readability. Geometrography supplies the formal and structural component of a broader empirical framework, while surveys and experimental studies are required to determine whether, and under which conditions, particular proof profiles correspond to human judgements. Several directions for future research follow from this framework. The first and most immediate consists in administering Survey 2 and systematically analysing the resulting data. Survey 2 is designed to evaluate quantitatively the candidate dimensions elicited from Survey 1 by asking participants to assess concrete theorems against explicitly stated criteria. The resulting data will make it possible to examine the stability and relative importance of the identified dimensions and to compare theorem-level judgements with the corresponding geometrographic profiles. A complementary empirical task concerns the validation of GRCP against human judgements of proof readability and comprehension. Studies involving mathematicians, computer scientists, educators, and students could determine whether GRCP, or any of its individual components, consistently reflects how readers follow and understand area-method proofs. Such studies would also help distinguish proof readability from theorem interestingness, two notions that may overlap in some cases but should not be treated as equivalent. Further research could include a systematic comparison between the geometrographic profiles of frequently cited theorems, such as Ceva's, Thales', and Pythagoras' theorems, and the judgements collected through the surveys. Another possible development is the construction of a *perceived-interest index* based on survey frequencies and justificatory categories, to be compared with GRCP-based structural scores. These comparisons could clarify whether particular formal features are associated with specific dimensions of interestingness rather than with a single, undifferentiated notion of mathematical interest. On the computational side, geometrographic parameters could be integrated into automated theorem-generation and theorem-finding systems as candidate features for heuristic filters or ranking criteria. Their effectiveness would then have to be evaluated empirically, rather than assumed in advance. Finally, repositories of automatically generated theorems could be enriched with survey-derived metadata and human evaluations, enabling larger-scale investigations of the relationships among mathematical interestingness, proof readability, and formal complexity. Taken together, these directions place geometrography within a broader empirical programme for the study of mathematical interestingness. It provides formal descriptions of proof structure, while surveys and experiments assess their relationship with human judgement. This integrated approach offers a promising framework for advancing the study of interestingness in automated reasoning while respecting both the limits of algorithmic evaluation and the central role of human mathematical understanding.

Appendix A: Ceva's Theorem

GCLC's Ceva's Theorem, with proof level 1, i.e. with multiple algebraic steps condensed in one step.





If needed, proof level 7 displays all the steps. At that level, the proof is four pages long.

Appendix B: First survey — interesting theorems

With this survey[11] the goal will be to find the characteristics that make a theorem interesting, or not. A list of questions about geometric theorems found to be interesting, or uninteresting.

B.1 Interesting and why?

A list of situations/explanations about interesting theorems.

[11] The survey is a *Google Form*, https://forms.gle/KNaHGMou9jNczhY48. If you did not answer it yet, please take some time to do it.

Can you describe in detail a situation (during classes or lectures) in which you have used
the adjective interesting applied to a theorem in geometry?

nth Situation _____

Can you explain in detail the reasons why you used the adjective interesting in the first situation?

nth Explanation _____

B.2 Five interesting theorems in geometry

A list of 5 questions, each about an interesting theorem.

Can you list at least five theorems in geometry that you consider interesting?

Theorem n _____

Can you explain in detail the reason for your choice by listing at least five adjectives that describe characteristics of the previous theorem making it interesting?

B.3 Five uninteresting theorems in geometry

A list of 5 questions, each about an uninteresting theorem.

Can you list at least five theorems in geometry that you consider NOT interesting?

Theorem n _____

Can you explain in detail the reason for your choice by listing at least five adjectives that describe characteristics of the previous theorem making it NOT interesting?

Appendix C: Second survey — characteristics of interesting theorems

This survey will only be designed after studying the results of the first survey. The second survey will propose theorems (taken from the first survey) and will provide characteristics (taken from the first survey) for each of them. The survey will ask the participants to express their opinion on characteristics that (presumably) make the theorems interesting or uninteresting.

The set of theorems will deliberately include contrastive cases, such as theorems with comparatively readable proofs that may nevertheless be judged routine or uninteresting, as well as theorems regarded as conceptually significant despite proofs of lower geometrographic readability. Participants will therefore assess both positive and negative cases.

Please express whether you consider the following theorems interesting or not, and why?

Is Theorem *n* interesting?
☐ YES ☐ NO

Why? Because it has the characteristic A.
☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Why? Because it has the characteristic B.
☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Why? Because it has the characteristic C.
☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Why? Because it has the characteristic D.
☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Acknowledgements The first author was partially supported by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. The second author was partially supported by the Italian Ministry of Education, University and Research through the PRIN 2022 project "Developing Kleene Logics and their Applications" (DeKLA), project code: 2022SM4XC8.

Author Contributions P.Q. and P.G. jointly conceived the study, developed the arguments, analysed the material, and wrote the manuscript. Both authors reviewed and approved the final version of the manuscript.

Funding Open access funding provided by Università degli Studi di Urbino Carlo Bo within the CRUI-CARE Agreement.

Data Availability The data analysed during the current study consist of expert survey responses collected through a Google Form. A summary of the responses is available on the project website (https://geotiles.mat.uc.pt/GeoTiles/Projects/SimpleProofs/simpleProofs.php). Further information may be obtained from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Wos, L.: Automated Reasoning: 33 Basic Research Problems, Prentice-Hall (1988)
2. Wos, L.: The problem of automated theorem finding. J. Autom. Reason. **10**(1), 137–138 (1993). https://doi.org/10.1007/BF00881868
3. Cheng, J.: Entailment calculus as the logical basis of automated theorem finding in scientific discovery. In: Systematic Methods of Scientific Discovery: Papers from the 1995 Spring Symposium, pp. 105–110. AAAI Press (1995)
4. Cheng, J.: A strong relevant logic model of epistemic processes in scientific discovery. In: Kawaguchi, E., Kangassalo, H., Jaakkola, H., Hamid, I.A. (eds.) Information Modelling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press (2000). https://doi.org/10.1007/3-540-49292-5_42
5. Colton, S., Bundy, A., Walsh, T.: On the notion of interestingness in automated mathematical discovery. Int. J. Hum Comput Stud. **53**(3), 351–375 (2000). https://doi.org/10.1006/ijhc.2000.0394
6. Gao, H., Goto, Y., Cheng, J.: A systematic methodology for automated theorem finding. Theoret. Comput. Sci. **554**, 2–21 (2014). https://doi.org/10.1016/j.tcs.2014.06.028
7. Gao, H., Cheng, J.: An epistemic programming approach for automated theorem finding. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCIxCC), pp. 49–58. IEEE Computer Society, Los Alamitos, CA, USA (2015). https://doi.org/10.1109/ICCI-CC.2015.7259365
8. Gao, H., Li, J., Cheng, J.: Measuring interestingness of theorems in automated theorem finding by forward reasoning based on strong relevant logic. In: 2019 IEEE International Conference on Energy Internet (ICEI), pp. 356–361. IEEE Computer Society, Los Alamitos, CA, USA (2019). https://doi.org/10.1109/ICEI.2019.00069
9. Puzis, Y., Gao, Y., Sutcliffe, G.: Automated generation of interesting theorems. In: FLAIRS Conference, (2006)
10. Abánades, M., Botana, F., Montes, A., Recio, T.: An algebraic taxonomy for locus computation in dynamic geometry. Comput. Aided Des. **56**, 22–33 (2014). https://doi.org/10.1016/j.cad.2014.06.008
11. Botana, F., Abánades, M.A., Escribano, J.: Computing locus equations for standard dynamic geometry environments. In: Shi, Y., Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) International Conference on Computational Science. Lecture Notes in Computer Science, vol. 4488, pp. 227–234. Springer, (2007). https://doi.org/10.1007/978-3-540-72586-2_32
12. Todd, P.: A method for the automated discovery of angle theorems. In: Proceedings of ADG 2021, vol. 352, pp. 148–155. Open Publishing Association, (2021). https://doi.org/10.4204/EPTCS.352.17
13. Peng, X., Chen, Q., Zhang, J., Chen, M.: Automated discovery of geometric theorems based on vector equations. J. Autom. Reason. **65**(6), 711–726 (2021). https://doi.org/10.1007/S10817-021-09591-2
14. Peng, X., Zhang, J., Chen, M., Liu, S.: Heuristic strategies for geometry theorem extension based on complex number identity*. Ann. Math. Artif. Intell. (2025). https://doi.org/10.1007/s10472-025-09988-4
15. Kovács, Z., Peng, X.: Automated proving in planar geometry based on the complex number identity method and elimination (2025). https://doi.org/10.48550/arXiv.2511.14728
16. Chen, X., Song, D., Wang, D.: Automated generation of geometric theorems from images of diagrams. Ann. Math. Artif. Intell. **74**(3–4), 1–26 (2014). https://doi.org/10.1007/s10472-014-9433-7
17. Dalzotto, G., Recio, T.: On protocols for the automated discovery of theorems in elementary geometry. J. Autom. Reason. **43**(2), 203–236 (2009). https://doi.org/10.1007/s10817-009-9133-x
18. Kovács, Z., Recio, T., Vélez, M.P.: Automated reasoning tools in GeoGebra discovery. ACM Commun. Comput. Algebra **55**(2), 39–43 (2021). https://doi.org/10.1145/3493492.3493495
19. Kovács, Z., Yu, J.H.: Automated discovery of geometrical theorems in geogebra. In: Marcos, J., Neuper, W., Quaresma, P. (eds.) Proceedings 10th International Workshop on Theorem Proving Components for Educational Software. Electronic Proceedings in Theoretical Computer Science, vol. 354, pp. 1–12 (2022). https://doi.org/10.4204/EPTCS.354.1
20. Recio, T., Vélez, M.P.: Automatic discovery of theorems in elementary geometry. J. Autom. Reason. **23**, 63–82 (1999). https://doi.org/10.1023/A:1006135322108
21. Ariño-Morera, B., Kovács, Z., Recio, T., Tolmos, P.: Solving with geogebra discovery an austrian mathematics olympiad problem: Lessons learned. Electron. Proceed. Theoret. Comput. Sci. **398**, 101–109 (2024). https://doi.org/10.4204/eptcs.398.13
22. Bartha, S., Kerekes, A., Kovács, Z., Recio, T.: Automated analysis of the difficulty of secondary school geometry theorems. Ann. Math. Artif. Intell. (2025). https://doi.org/10.1007/s10472-025-09983-9

23. Kovács, Z., Parisse, B., Recio, T., Vélez, M.P., Yu, J.H.: The showproof command in geogebra discovery: Towards the automated ranking of elementary geometry theorems. ACM Commun. Comput. Algebra **58**(2), 27–30 (2024). https://doi.org/10.1145/3712023.3712026
24. Kovács, Z., Recio, T., Vélez, M.P.: Showing proofs, assessing difficulty with geogebra discovery. In: Electronic Proceedings in Theoretical Computer Science, vol. 398, pp. 43–52 (2024). https://doi.org/10.4204/eptcs.398.8
25. Lemoine, É.: Géométrographie Ou Art des Constructions Géométriques. Scientia, Série Physico-Mathématique, vol. 18. C. Naud, Éditeur, Paris (1902). http://catalogue.bnf.fr/ark:/12148/cb36049032t
26. Mackay, J.S.: The geometrography of Euclid's problems. Proc. Edinb. Math. Soc. **12**, 2–16 (1893). https://doi.org/10.1017/S0013091500001565
27. Quaresma, P., Santos, V., Graziani, P., Baeta, N.: Taxonomy of geometric problems. J. Symb. Comput. **97**, 31–55 (2020). https://doi.org/10.1016/j.jsc.2018.12.004
28. Quaresma, P., Graziani, P.: The geometrography's simplicity coefficient for the axioms and lemmas of the area method. Technical Report TR 2021-001, Center for Informatics and Systems of the University of Coimbra (2021). https://www.cisuc.uc.pt/download-file/16035/p5jMyRCjncUcRlfoKnNU
29. Quaresma, P., Graziani, P., Nicoletti, S.M.: Considerations on approaches and metrics in automated theorem generation/finding in geometry. Electron. Proceed. Theoret. Comput. Sci. **398**, 85–100 (2024). https://doi.org/10.4204/eptcs.398.12
30. Quaresma, P., Graziani, P.: Measuring the readability of geometric proofs: The area method case. J. Autom. Reason. **67**(5), (2023). https://doi.org/10.1007/s10817-022-09652-0
31. Abar, C.A.P., Almeida, M.V., Recio, T.: Estimating the Complexity Levels of a Geometric Statement. Boletín de la Sociedad Puig Adam **121**, 76–97 (2026)
32. De Millo, R.A., Lipton, R.J., Perlis, A.J.: Social processes and proofs of theorems and programs. Commun. ACM **22**(5), 271–280 (1979). https://doi.org/10.1145/359104.359106
33. MacKenzie, D.: The automation of proof: A historical and sociological exploration. IEEE Ann. Hist. Comput. **17**(3), 7–29 (1995). https://doi.org/10.1109/85.397057
34. MacKenzie, D.: Slaying the kraken: The sociohistory of a mathematical proof. Soc. Stud. Sci. **29**(1), 7–60 (1999). https://doi.org/10.1177/030631299029001002
35. Aldon, G., Cahuet, P.-Y., Durand-Guerrier, V., Front, M., Krieger, D., Mizony, M., Tardy, C.: Expérimenter des Problèmes Innovants en Mathématiques à l'école. INRP, (2010). https://hal.archives-ouvertes.fr/hal-00989132
36. Bundy, A., Jamnik, M., Fugard, A.: What is a proof? Philosoph. Trans. R. Soc. A: Math. Phys. Eng. Sci. **363**(1835), 2377–2391 (2005). https://doi.org/10.1098/rsta.2005.1651
37. D'Amore, B., Fandiño Pinilla, M.: Una formula per la misurazione oggettiva della difficoltà di comprensione di un testo di matematica da parte degli studenti. Uso valutativo e uso didattico. La matematica e la sua didattica, 1. **2**, 59–78 (2016)
38. Hemmi, K., Julin, E., Pörn, R.: Misconceptions and developmental proof. In: CERME 10 (2017)
39. Polya, G.: How to Solve It: A New Aspect of Mathematical Method vol. 246. Princeton University Press, (2004)
40. Stylianides, A.J., Harel, G.: Advances in Mathematics Education Research on Proof and Proving: An International Perspective. Springer (2018). https://doi.org/10.1007/978-3-319-70996-3
41. Arzarello, F., Sabena, C.: Semiotic and theoretic control in argumentation and proof activities. Educ. Stud. Math. **77**(2–3), 189–206 (2010). https://doi.org/10.1007/s10649-010-9280-3
42. Aschbacher, M.: Highly complex proofs and implications of such proofs. Philosoph. Trans. R. Soc. A: Math. Phys. Eng. Sci. **363**(1835), 2401–2406 (2005). https://doi.org/10.1098/rsta.2005.1655
43. Avigad, J.: Mathematical method and proof. Synthese **153**(1), 105–159 (2006). https://doi.org/10.1007/s11229-005-4064-5
44. Barendregt, H., Wiedijk, F.: The challenge of computer mathematics. Philosoph. Trans. R. Soc. A: Math. Phys. Eng. Sci. **363**(1835), 2351–2375 (2005). https://doi.org/10.1098/rsta.2005.1650
45. Burge, T.: Computer proof, apriori knowledge, and other minds: The sixth philosophical perspectives lecture. Philos. Perspect. **12**, 1–37 (1998). https://doi.org/10.1093/acprof:oso/9780199672028.003.0013
46. Gao, H., Goto, Y., Cheng, J.: A set of metrics for measuring interestingness of theorems in automated theorem finding by forward reasoning: A case study in NBG set theory. In: Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques, pp. 508–517. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-319-23862-3_50
47. Gao, H., Cheng, J.: Measuring interestingness of theorems in automated theorem finding by forward reasoning: A case study in Peano's arithmetic. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawinski, B. (eds.) Intelligent Information and Database Systems. Lecture Notes in Computer Science, vol. 10192, pp. 115–124. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-54430-4_12

48. Gao, H., Li, J., Cheng, J.: Measuring interestingness of theorems in automated theorem finding by forward reasoning: A case study in Tarski's geometry. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE Computer Society, Los Alamitos, CA, USA (2018). https://doi.org/10.1109/SmartWorld.2018.00064
49. Janičić, P., Narboux, J., Quaresma, P.: The Area Method: A Recapitulation. J. Autom. Reason. **48**(4), 489–532 (2012). https://doi.org/10.1007/s10817-010-9209-7
50. Santos, V., Baeta, N., Quaresma, P.: Geometrography in dynamic geometry. Int. J. Technol. Math. Educ. **26**(2), 89–96 (2019). https://doi.org/10.1564/tme_v26.2.06
51. Hohenwarter, M.: GeoGebra — a software system for dynamic geometry and algebra in the plane. Master's thesis, University of Salzburg, Austria (2002)
52. Quaresma, P., Janičić, P.: The Area Method, rigorous proofs of lemmas in Hilbert's style axiom system. Technical Report 2009/006, Centre for Informatics and Systems of the University of Coimbra (2009). https://www.cisuc.uc.pt/download-file/13539/8FRDv6OvNKK7d095kCtB

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.