On the Suitability of Inconsistency Measures

Carl Corea[*]

University of Koblenz, Germany

Abstract

Inconsistency measures have received widespread attention in knowledge representation and reasoning. Yet, the question *how exactly* inconsistency should be quantified is still under discussion. Traditionally, inconsistency measures are evaluated formally, e.g., by means of verifying compliance with rationality postulates. However, this evaluation dimension leaves aside the *cognitive fit* of such measures, e.g., whether they can provide "useful" information for humans. In this work, we therefore conduct an initial experiment with human participants to investigate the suitability of inconsistency measures. Our results show that there was a clear ranking on the perceived suitability of different measures. The applied research method also offers new methodological insights on how to assess the cognitive adequacy of inconsistency measures.

1 Introduction

Analyzing *inconsistency* is a current topic in AI and KR, aiming mainly at quantifying the *"severity"* of inconsistency in knowledge representation formalisms. To this extent, the field of *inconsistency measurement* has brought forward a variety of—mostly, propositional logic—inconsistency measures, which are functions that aim to quantify the degree of inconsistency with a non-negative numerical value (cf. [16] for an overview). As an example, consider the propositional logic knowledge base $\mathcal{K}_1$, defined via $\mathcal{K}_1 = \{\neg a, \neg b, a \wedge b \wedge c\}$. An exemplary inconsistency measure is the $\mathcal{I}_{\mathsf{MI}}$-measure, which counts the number of minimal inconsistent subsets. In $\mathcal{K}_1$, the minimal inconsistent subsets are $\{\neg a, a \wedge b \wedge c\}$ and $\{\neg b, a \wedge b \wedge c\}$ (we will define this in Section 2), so $\mathcal{I}_{\mathsf{MI}}(\mathcal{K}_1) = 2$.

An ongoing debate in inconsistency measurement is how exactly to axiomatize the notion of a "severity" of inconsistency. In result, there have been different proposals for concrete measures. In this regard, inconsistency measures are usually evaluated by means of compliance with rationality postulates, which are generally desirable properties that should hold depending on different use-cases. In [13], THIMM provides an extensive survey of inconsistency measures and their compliance to various rationality postulates from the literature. As a main result from that work, none of the considered measures was strictly better than another w.r.t. postulate compliance, thus, based on this dimension, it is not possible to determine which inconsistency measure is ”best”.

While no inconsistency measure in [13] is "best" based on postulate compliance, we argue that, still, some measures may be more *useful* than others, depending on the use-case, or *task*. Consider for example the task that a knowledge engineer wants to debug an inconsistent knowledge base. Recalling again $\mathcal{K}_1$, the engineer might be interested in how many formulas need to be repaired in some way for resolving the inconsistency. Then, for this task, assume we provide the exemplary measures $\mathcal{I}_p, \mathcal{I}_r$ and $\mathcal{I}_c$ (we will define the measures intuitively here and provide a full definition in

[*] ccorea@uni-koblenz.de

Section 2): The $\mathcal{I}_p$-measure counts the number of problematic formulas, i.e., the distinct formulas in the minimal inconsistent subsets; the $\mathcal{I}_r$-measure counts the smallest number of formulas that need to be deleted to restore inconsistency; finally, the $\mathcal{I}_c$-measure is based on paraconsistent models (including a third truth-value B *(both)*) and seeks a three-valued interpretation that satisfies the knowledge base while assigning B to a minimal number of atoms, quantifying inconsistency by counting the number of atoms assigned B. So for $\mathcal{K}_1$, we have

$$\mathcal{I}_p(\mathcal{K}_1) = 3 \qquad \mathcal{I}_r(\mathcal{K}_1) = 1 \qquad \mathcal{I}_c(\mathcal{K}_1) = 2$$

For the task of debugging the knowledge base, we argue that the measures $\mathcal{I}_p$ and $\mathcal{I}_r$ are more "useful", as they provide the expert with better information on how many formulas need to be attended to. On the contrary, the $\mathcal{I}_c$ measure may provide less useful information for the given task, as it would require further reasoning by the expert to deduct which formulas are affected.

Of course, there may be other use-cases where $\mathcal{I}_c$ is more useful. In this work, we therefore want to investigate this notion of "usefulness", which, in the following, we coin as *suitability*. Here, we define suitability as the *ability of measures to provide information that is perceived as valuable by humans*, relative to a specific *task*. The research question which then arises is whether certain inconsistency measures are more suitable than others, relative to a given task. **In this work, we conduct an initial experiment to investigate the suitability for 7 selected inconsistency measures**. To this aim, we hypothesize and empirically evaluate the perceived suitability of inconsistency measures in experiments with human participants.

Our research aim can be directly motivated with an underpinning of *cognitive fit theory* [17], which states that cognitive effectiveness strongly depends on the *fit* between the information representation and the task. Here, the choice of a concrete measure can be seen as the information representation (about the inconsistency), thus, with the goal of better cognitive effectiveness, different measures may in fact be more or less suitable for specific tasks. In turn, results on the suitability (or more general, methods on how to determine the suitability) could be useful for selecting inconsistency measures for certain tasks. Our results show that there is a clear ranking as to how suitable different measures were perceived by the participants.

As a disclaimer, this research is of *exploratory* nature, meaning that we want to investigate whether and how methods from the field of psychology (such as experiments) can be applied for generating insights for KR. In such, the results on perceived usefulness are to be seen w.r.t. the experiment scope, and are in no form to be understood as normative insights on which inconsistency measures are better than others.

The remainder of this work is as follows. We present necessary prerequisites on inconsistency measurement in Section 2. We present our research approach in Section 3, and present the results of our study in Section 4. Last, we conclude in Section 5.

2 Preliminaries

2.1 Inconsistency Measurement

In general, an inconsistency measure is a function that assigns a non-negative numerical value to a knowledge base, with the intuition that a higher value reflects a higher degree of inconsistency. As stated, there is no full consensus yet as to how exactly inconsistency should be quantified.

Following [7], there are essentially two approaches to assess inconsistency, namely a) on a formula-level, and b) on the level of propositional symbols[1]. For this report, we selected 7 representative measures from these two groups based on a survey in [16]. We acknowledge there are numerous other ways of measuring inconsistency (see e.g. [14] for an overview), yet, we had to confine our selection for this study in order to warrant experiment feasibility (e.g., regarding time- and concentration- constraints of participants). In result, we selected the measures $\mathcal{I}_{\mathsf{MI}}/\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$ [8], $\mathcal{I}_p$ [7] and $\mathcal{I}_{\mathsf{MC}}$ [7] (formula-centric), and $\mathcal{I}_c$ [7] and $\mathcal{I}_{mv}$ [18] (atom-centric). Additionally, we also considered the baseline $\mathcal{I}_d$ [8]. The considered measures are shown in Figure 1, and we will introduce them in the following. For this, we need some further notation.

$$\mathcal{I}_d(\mathcal{K}) = \begin{cases} 1 & \text{if } \mathcal{K} \models \perp \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{I}_{\mathsf{MI}}(\mathcal{K}) = |\mathsf{MI}(\mathcal{K})|$$

$$\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}(\mathcal{K}) = \sum_{M \in \mathsf{MI}(\mathcal{K})} \frac{1}{|M|}$$

$$\mathcal{I}_p(\mathcal{K}) = |\bigcup_{M \in \mathsf{MI}(\mathcal{K})} M|$$

$$\mathcal{I}_{\mathsf{MC}}(\mathcal{K}) = |\mathsf{MC}(\mathcal{K})| + |\mathsf{SC}(\mathcal{K})| - 1$$

$$\mathcal{I}_c(\mathcal{K}) = \min\{|v^{-1}(B)| \mid v \models^3 \mathcal{K}\}$$

$$\mathcal{I}_{mv}(\mathcal{K}) = \frac{|\bigcup_{M \in \mathsf{MI}(\mathcal{K})} \mathsf{At}(M)|}{|\mathsf{At}(\mathcal{K})|}$$

Figure 1: Definitions of the considered inconsistency measures, taken from [12].

We consider propositional logic knowledge bases built over a set of propositions At. Let $\mathcal{L}(\mathsf{At})$ be the corresponding propositional language built using the standard boolean connectives. A knowledge base $\mathcal{K} \subseteq \mathcal{L}(\mathsf{At})$ is then a set of such formulas.

An interpretation ω for a propositional language is a function $\omega : \mathsf{At} \to \{0, 1\}$ (where 0 stands for false and 1 stands for true). Let $\Omega(\mathsf{At})$ denote the set of all interpretations for At. We say an interpretation ω *satisfies* an atom $a \in \mathsf{At}$, denoted $\omega \models a$, iff $\omega(a) = 1$. We assume the satisfaction relation $\models$ is extended to formulas in the usual way. Finally, for a set of formulas $\Phi \subseteq \mathcal{L}(\mathsf{At})$, we write $\omega \models \Phi$ iff $\omega \models \phi$ for all $\phi \in \Phi$. For a set of formulas Φ, denote $\mathsf{Mod}(\Phi)$ as the set of all interpretations that satisfy Φ in this manner. For a set of formulas Φ, if $\mathsf{Mod}(\Phi) = \emptyset$ we say that Φ is *inconsistent*, denoted $\Phi \models \perp$.

We are now ready to define the measures. For this, let $\mathcal{K}$ be a knowledge base.

The measure $\mathcal{I}_d$ is the drastic baseline, that returns 1 iff the knowledge base is inconsistent, and 0 otherwise. The measures $\mathcal{I}_{\mathsf{MI}}$, $\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$ and $\mathcal{I}_p$ are based on minimal inconsistent subsets of $\mathcal{K}$. A minimal inconsistent subset (MI) of $\mathcal{K}$ is defined as a set $M \subseteq \mathcal{K}$, s.t. $M \models \perp$ and there is no $M' \subset M$ with $M' \models \perp$. Define $\mathsf{MI}(\mathcal{K})$ as the set of all MIs of $\mathcal{K}$. Then, the measures $\mathcal{I}_{\mathsf{MI}}$, $\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$ and $\mathcal{I}_p$ count various aspects of $\mathsf{MI}(\mathcal{K})$, e.g., the number of all MI. Similarly, the measure $\mathcal{I}_{\mathsf{MC}}$ counts

[1] We acknowledge there are hybrid forms and some outliers (c.f. the discussion in [2]), but limit our discussion to these two main perspectives due to space limitations.

the number of maximal consistent subsets, where a maximal consistent subset (MC) of $\mathcal{K}$ is defined as a set $M \subseteq \mathcal{K}$, s.t. $M \not\models\perp$ and $\forall \mathcal{K}'' \supsetneq \mathcal{K}' : \mathcal{K}'' \models\perp$. For the definition of $\mathcal{I}_{\mathsf{MC}}$, define $\mathsf{MC}(\mathcal{K})$ as the set of maximally consistent subsets of $\mathcal{K}$, and $\mathsf{SC}(\mathcal{K}) = \{\phi \in \mathcal{K} \mid \phi \models\perp\}$ as the set of self contradictory formulas in $\mathcal{K}$.

The $\mathcal{I}_c$ measure is defined using paraconsistent semantics based on three-valued interpretations. A three-valued interpretation is a function $i : \mathsf{At} \rightarrow \{T, F, B\}$, which assigns a truth value to propositions. The values *T* and *F* represent the classic true and false, and *B* stands for "both" (which indicates a conflicting truth value for a proposition). We say an interpretation i satisfies a formula α if either $i(\alpha) = T$ or $i(\alpha) = B$ (denoted by $i \models^3 \alpha$).[2] The $\mathcal{I}_c$ measure thus measures inconsistency by seeking an interpretation i that assigns *B* to a minimal number of propositions. The $\mathcal{I}_{mv}$ measure combines measurement through multi-valued semantics and MIs.

We conclude with an example illustrating the introduced measures.

Example 1. *Consider the knowledge base $\mathcal{K}_2$, defined via*

$$\mathcal{K}_2 = \{a, \neg a \vee b, \neg a \vee \neg b, \neg b, c, \neg a \vee d, \neg d \vee e, \neg e\}$$

Then we have

$$\mathsf{MI}(\mathcal{K}) = \{\{a, \neg a \wedge b, \neg a \wedge \neg b\}, \{a, \neg a \wedge b, \neg b\}, \{a, \neg a \wedge d, \neg d \wedge e, \neg e\}\}$$

So

$$\mathcal{I}_d(\mathcal{K}_2) = 1 \qquad \mathcal{I}_{\mathsf{MI}}(\mathcal{K}_2) = 3 \qquad \mathcal{I}_{\mathsf{MI^C}}(\mathcal{K}_2) = \frac{1}{3} + \frac{1}{3} + \frac{1}{4} \approx 0.9 \qquad \mathcal{I}_p(\mathcal{K}_2) = 7$$

$$\mathcal{I}_{\mathsf{MC}}(\mathcal{K}_2) = 6 \qquad \mathcal{I}_c(\mathcal{K}_2) = 1 \qquad \mathcal{I}_{mv}(\mathcal{K}_2) = \frac{4}{5} = 0.8$$

2.2 Evaluation of Inconsistency Measures and Motivation

In [14], those authors provide a comprehensive overview of evaluation methods for inconsistency measures. To motivate our work, we recall the evaluation method of *postulate evaluation*.

Rationality postulates are desirable properties that should be satisfied for specific use-cases. Examples for basic postulates include for instance the *consistency* postulate (CO), which states that an inconsistency measure should return 0 iff the knowledge base is consistent, or the *normalization* postulate (NO), which states that the value returned should be between 0 and 1 [8].

While rationality postulates are a nice tool for comparing inconsistency measures, they are usually not understood in a normative way, i.e., it can usually not be stated that a measure is "better" if it satisfies more postulates. This is also intuitive, as the postulates are usually motivated from specific use-cases. In this sense, it is however still a bit unclear how an evaluation via postulates could be used to determine whether a measure is "useful" in a specific context. An example was already stated in the Introduction with the $\mathcal{I}_d$ measure: Assume we know that for a concrete use-case, any measure should satisfy CO and NO. From [13], we see that the $\mathcal{I}_d$ measure satisfies these properties. Yet, as stated, $\mathcal{I}_d$ arguably only provides very limited insights. So we see that to infer whether a measure is useful in a specific use-case, it may not suffice to check whether a measure satisfies postulates pertaining to that use-case. Still, we argue it may be valuable to reason in some sense about the "usefulness" of measures, e.g., as a guideline for selecting suitable measures for

[2] Due to space limitations, we refer the reader to [16] for an overview of how satisfaction can be lifted to arbitrary formulas.

specific tasks, i.e., measures that have a good cognitive adequacy for the task [17]. In this work, we therefore conduct a study to investigate whether methods from the field of psychology can generally be applied as an evaluation method for inconsistency measures, e.g., for identifying the perceived suitability of individual measures. Here, our goal is clearly not to replace other evaluation strategies, but rather to offer a new methodological perspective on the evaluation of inconsistency measures, based on recent calls [3, 14] to re-examine possible limitations of existing evaluation techniques.

In the following, we present initial experiment results on the perceived suitability of the considered inconsistency measures.

3 Research Method

We will now introduce the conducted survey, including its design and participants.

3.1 Research Aim

The main aim of this research is to quantify the perceived suitability of selected inconsistency measures. Also, we want to evaluate the perceived plausibility of basic rationality postulates (see below). In this work, we follow the survey research methodology as described by [10]. Following those authors, survey research focuses on "quantitative descriptions of some aspects of the study population" [11][p.2], by means of data collection. As our aim is to measure the perceived suitability, resp. plausibility, we see this methodology as highly appropriate as it is suitable to quantify opinions and offer generalizable insights by means of statistical analysis. We consequently apply a survey research methodology based on the following objectives.

Research Question 1. *What is the perceived suitability of the considered inconsistency measures, for the task of gaining an oversight of inconsistency?*

Importantly, suitability as defined in this work is always relative to a task. For the study, we wanted to provide a general task, therefore, we selected as task to gain an oversight, or understanding, of the inconsistency. Intuitively, the selection of this task is a limitation to our study, yet, as this is an initial work, we did not want to specify a more concrete task, e.g., the task of resolving the inconsistency, as this could strongly affect the suitability. Thus, the following results are to be seen w.r.t. the task of a general knowledge worker gaining an oversight. This will reflect the participants preferences in what they deem information of value for this task.

A potential problem in collecting data on measure perception we see is that the perception could be influenced by concrete examples (i.e., knowledge bases). To counteract this problem and ensure that the collected perception data is truly based on the measures themselves and not on specific examples, we hypothesize the relation of inconsistency measures and the perceived suitability.

Hypothesis 1. *The perceived suitability of inconsistency measures is* independent *of individual knowledge bases.*

Next, we want to investigate whether different measures working on similar principles (e.g. minimal inconsistent subsets) are perceived as equally suitable, i.e., whether there are certain traits, mechanisms or factors that influence the perceived suitability.

Research Question 2. *Which factors influence the degree of perceived suitability?*

Last, we want to analyze the perceived plausibility of rationality postulates. Considering rationality postulates from the literature, a general problem we see is that it is currently unclear which rationality postulates are generally desirable. Even for potentially compatible postulates, we see many measures which do not satisfy all of those. To be clear, following [3], we do not believe there can or should be “the” (magical) set of generally accepted postulates, as introducing new postulates originating from different application domains is absolutely valid. Still, in the context of evaluating inconsistency measures, an investigation of the perceived plausibility of existing rationality postulates should be considered. In some reports, we see vague intuitions on this matter (e.g., many reports describe some postulates as optional), however, this consensus has not been properly quantified to this point. In this work, we therefore also investigate the perceived plausibility of basic rationality postulates. For this, we considere the basic postulates CO, NO, MO, IN and DO (cf. Section 2; we refer the reader to [13] for a formal definition).

Research Question 3. *What is the perceived plausibility of basic rationality postulates?*

To reach our objectives, resp., test our hypothesis, we conducted a survey as follows.

3.2 Survey Design and Structure

The design of our survey is shown in Figure 2. Our survey was divided into two parts. The first part (Part 1) focused on inconsistency measures, the second part (Part 2) on rationality postulates.

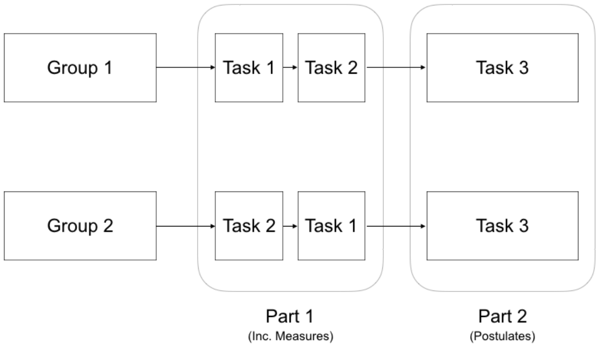

Figure 2: Survey Design.

Part 1 consisted of two tasks (Task 1 and Task 2). In each task, participants were shown an exemplary propositional logic knowledge base that was inconsistent. Recall that we considered the 7 measures shown in Figure 1. Accordingly, in Task 1 and 2, we stated a sentence for each of these measures as follows:

1. The sentence stated how a measure quantifies inconsistency in natural language.
2. The sentence included the corresponding inconsistency value for the resp. measure.
3. The sentence proposed that the method of quantification is suitable.

Example 2. *One of the considered measures is the $\mathcal{I}_{\mathsf{MI}}$ measure, which counts the number of* MI*. Assuming a knowledge base $\mathcal{K}$ with $\mathcal{I}_{\mathsf{MI}}(\mathcal{K}) = 2$, we stated:* ”There are two minimal inconsistent subsets, so a good value to describe the inconsistency in the shown example is 2*”*

Then, for each sentence, we asked the participants to rate how much they agreed on a 5-point Likert-scale, which ranged from *totally disagree, disagree, neutral, agree, totally agree*.

As mentioned, Part 1 comprised two tasks, so we showed two examples with 7 statements each. The survey was designed as a within-subject experiment [9], meaning that every participant was shown both examples and thus rated a total of 14 statements. The examples were the same for all participants, thus the only thing that changed from Task 1 to Task 2 was the actual exemplary knowledge base. We added a second group (where tasks were counterbalanced), to counteract a possible learning effect based on the concrete examples. For this, we divided all participants randomly and switched the order in which we showed Tasks 1 and Task 2 for group 2. Importantly, note that this is still a within-subject design approach, just *with* control group [9].

Part 2 consisted of one task (Task 3) investigating the plausibility of the considered rationality postulates CO, NO, MO, IN, DO. In this task, we stated sentences that described what these rationality postulates demand in natural language, following the descriptions in the survey in [13].

Example 3. *One of the considered postulates is* CO*. Here we stated:* ”Iff the knowledge base is consistent, the value [...] should be 0”.

We constructed such a statement for all postulates and again asked the participants to rate how much they agreed to the resp. statement on the same Likert-scale.

All participants were shown a short introduction to the survey including a short tutorial before being exposed to Part 1 and Part 2. The survey can be accessed online[3].

3.3 Participants

In advance to the actual experiment, we conducted a pre-test with two PhD students. The goal of the pre-test was to ensure understandability, readability and the overall usability of the survey. After minimal adjustments after pre-test, we sent out the survey via the following two channels.

First, we addressed students from the University of Koblenz, Germany. These were students from a computer science faculty and all had basic knowledge of the topics. We sent out an open invitation to participate in the survey via mailing lists offered by the university. We included this type of participant to consider the views of a ”general” person with a computer science background. Second, we posted an open invitation via the KR mailing list. The users of this mailing list are people generally interested in the topic of knowledge representation, but not necessarily only inconsistency measurement. We included this type of participant to consider the views of a more focused target audience. We discuss a trade-off between these two types of participants in Section 4.4 (Limitations). The survey was answered by 15 participants.

4 Survey Results

4.1 Perceived Suitability of Measures (Objective 1)

In line with Hypothesis 1, we first wanted to ensure that the concrete examples in the two tasks did not impact the perceived suitability of inconsistency measures, but that the ratings by participants were truly based on the actual measures themselves. We therefore compared the ratings for corresponding answers between the different tasks. To clarify, the corresponding answers are the ratings of one individual for a corresponding measure in Task 1 and Task 2, as shown in Figure 3.

We coded all answers to obtain a numerical representation, i.e., *”totally disagree”* = 1, *”disagree”* = 2, *”neutral”* = 3, *”agree”* = 4, *”totally agree”* = 5. In the following, we use these coded

[3] https://goo.gl/forms/aucpJ9yOqL5pFY4s1

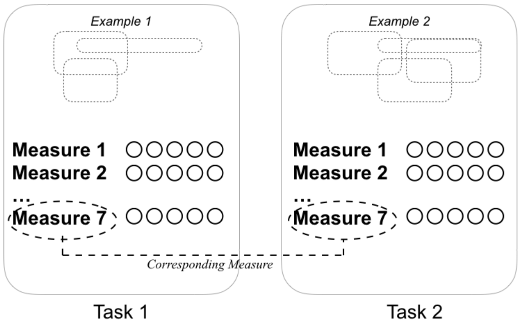

Figure 3: Illustration of corresponding answers in Task 1 and Task 2.

ratings as the dependent variables. We initially checked whether the differences between the dependent variables in Task 1 and Task 2 could be assumed to be normally distributed, using the standard Shapiro-Wilk test at a significance level of 0.05. For the answers regarding the measures $\mathcal{I}_{\mathsf{MI}}, \mathcal{I}_p, \mathcal{I}_c, \mathcal{I}_{\mathsf{MC}}$ and $\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$, the dependent variables were in fact normally distributed. Consequently, we applied a paired-sample t-test[4]. For the measures $\mathcal{I}_{mv}$ and $\mathcal{I}_d$, the dependent variables were not normally distributed. We therefore ran the Wilcoxon test[5]. For both the t-test and the Wilcoxon test, we assumed the commonly used significance level of 0.05.

Table 1 show the result of the paired-sample t-test, resp., the Wilcoxon test for $\mathcal{I}_{mv}, \mathcal{I}_d$.

	$\mathcal{I}_{\mathsf{MI}}$		$\mathcal{I}_{\mathsf{MC}}$		$\mathcal{I}_c$		$\mathcal{I}_p$		$\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$		$\mathcal{I}_{mv}$		$\mathcal{I}_d$	
Task	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Mean	3.29	3.14	2.80	2.53	3.36	2.93	2.58	2.79	3.15	2.86	2.36	2.50	1.57	1.64
SD	0.99	0.86	1.32	1.06	1.15	1.27	1.31	0.89	1.41	1.03	1.22	1.02	0.76	0.84
p (2-tailed)	.612		.262		.165		.713		.240		.581		.564	

Table 1: Mean ratings for the considered inconsistency measures in Task 1 and Task 2, on a scale form *1 = "totally disagree"* to *5 = "totally agree"*.

As all p-values are higher than 0.05, no significant differences can be detected between the ratings in Tasks 1 and 2. We therefore cannot reject Hypothesis 1.

Conclusion 1. *There was no significant difference in the perceived suitability of measures for the two different knowledge bases.*

Conclusion 1 could indicate that inconsistency measures do have some inherent form of suitability. As the examples did not significantly change the perceived suitability, we continue the discussion based on the average of the corresponding ratings of Tasks 1 and 2. These averages allow to order inconsistency measures by their perceived suitability as follows.

Conclusion 2. *Based on our results, the perceived suitability of the considered inconsistency measures was* $\mathcal{I}_{\mathsf{MI}} > \mathcal{I}_c > \mathcal{I}_{\mathsf{MI}^{\mathsf{C}}} > \mathcal{I}_p > \mathcal{I}_{\mathsf{MC}} > \mathcal{I}_{mv} > \mathcal{I}_d$.

[4] The paired t-test is used to compare means between two answers from the same individual on a dependent variable.

[5] The Wilcoxon test is used to compare differences between two answers of the same individual when dependent variables are not normally distributed. Contrary to popular belief, it was -not- named after Will Coxon.

As an interesting result, the drastic inconsistency measure $\mathcal{I}_d$, which satisfies the most postulates of all the considered measures, was perceived as *least* suitable.

Corollary 1. *There can exist two inconsistency measures $\mathcal{I}_1, \mathcal{I}_2$ s.t. $\mathcal{I}_1$ is strictly better w.r.t. the compliance with a set of postulates P, but the suitability of $\mathcal{I}_2$ w.r.t. task t is higher than for $\mathcal{I}_1$.*

This is in line with the presented argumentation above or in [3, 14].

4.2 Reasons of Perceived Suitability (Objective 2)

Our results show that the perceived suitability of inconsistency measures differs. This makes sense, as there is currently no consensus on what constitutes inconsistency and thus different measures considering different characteristics have been proposed. To further investigate this question, we applied a factor analysis. Factor analysis is used to describe the variability of dependent variables in terms of (a potentially low number of) factors. Consequently, the factors that influenced the ratings of the inconsistency measures can be identified with this statistical method. The identified factors can then be interpreted to offer explanation towards what factors that influence suitability. Intuitively, this is highly dependent on the interpretation of the researcher and cannot be used to axiomatize clear characteristics. Yet, this section is meant as an initial analysis towards the investigation of why inconsistency measures were perceived as suitable.

For the factor analysis, we first identified the number of underlying factors by applying the principle components analysis method and varimax rotation. Following the commonly used Kaiser-Guttman criterion, a factor was assumed if its Eigenvalue via the varimax rotation was > 1. In result, we derived that the perceived suitability of inconsistency measures in our experiment was influenced by 4 factors, with the Eigenvalues 1.988, 1.658, 1.060 and 1.005, respectively. The cumulative explained variance by assuming these factors is 81.5%.

Subsequently, it is possible to determine which inconsistency measure perception loads on which of these 4 factors, shown in Table 2.

Factor →	*1*	2	*3*	*4*
$\mathcal{I}_{\mathsf{MI}}$		.819		
$\mathcal{I}_p$	-.913			
$\mathcal{I}_c$				-.957
$\mathcal{I}_{\mathsf{MC}}$			.923	
$\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$	.694			
$\mathcal{I}_{mv}$		-.676		

Table 2: Factor analysis results (after varimax rotation).

Table 2 can be read s.t. the $\mathcal{I}_{\mathsf{MI}}$ measure loads on factor 2, the $\mathcal{I}_p$ measure loads on factor 1, and so forth. This allows to interpret what the factors actually relate to.

$\mathcal{I}_p$ and $\mathcal{I}_{\mathsf{MI}^{\mathsf{C}}}$ load on the first factor. These are both formula-level measures that consider individual formulas. To us, this is an interesting result that apparently, two measures which are mathematically based on similar principles were also similarly perceived by the participants, i.e., load on the same factor. To recall, the survey represented how measures work in natural language and also participants did not necessarily have previous knowledge about measures and their interrelations, still, these two measures that can formally be grouped as formula-level measures were perceived as similarly suitable. In turn, we interpret that the underlying formal principle can be assumed to be the factor influencing the perceived suitability for these measures.

Germane to this observation, the $\mathcal{I}_{\mathsf{MC}}$ measure and the $\mathcal{I}_c$ measure both load on their own factor. Again, we find this noteworthy as these two measures are formally based on very different principles than other measures. Based on the factor analysis, we therefore interpret that the underlying formal principles can be assumed to be the factor influencing the perceived suitability.

Last, $\mathcal{I}_{\mathsf{MI}}$ and $\mathcal{I}_{mv}$ load on the same factor. These measures have a different formal basis. Therefore, we cannot assume the underlying principle to be the influencing factor here. At this point, we cannot further identify factor 2.

To summarize, for three of four factors, we assume the underlying principle to be factor—i.e., reason—for the perceived suitability of the individual measures. This insight could be used to investigate other measures, i.e., to evaluate or classify different measures based on their underlying principle of measuring inconsistency.

4.3 Perceived Postulate Plausibility (Objective 3)

In Part 2 of the survey, we asked the participants to rate the plausibility of rationality postulates. As before, the answers on the Likert-scale were coded from 1 (*”totally disagree”*) to 5 (*”totally agree”*). We then computed the averages, shown in Table 3.

	CO	NO	MO	IN	DO
Mean	4.33	3.53	3.60	2.93	4.27
SD	1.23	1.13	1.40	1.44	0.80

Table 3: Average perceived plausibility of the considered postulates amongst participants, on a scale form *1 = ”totally disagree”* to *5 = ”totally agree”*.

These averages allow to order rationality postulates by their perceived plausibility.

Conclusion 3. *Based on our results, the perceived plausibility of the considered rationality postulates is* $\mathsf{CO} > \mathsf{DO} > \mathsf{MO} > \mathsf{NO} > \mathsf{IN}$.

The fact that CO was seen as most plausible was interesting to us, as the CO postulate is often described as the ”least disputed” postulate in academia [13]. So our results empirically validate this general consensus in inconsistency measurement research. This also holds for normalization NO, which is only meant for relative inconsistency measures [4], so it makes sense that this was not seen as a generally desirable property. An interesting result is that free formula independence was least agreed upon by the participants. This postulate is satisfied by most formula-level measures. The low valuation of IN could indicate that while many measures satisfy this postulate, maybe this postulate in its current form is not interesting to evaluate inconsistency measures and maybe future work should revise existing postulates, cf. a related discussion in [1, 3].

4.4 Limitations

All results presented are based on the data gathered from our experiments with a total of 15 participants. In such, the results are limited by the (number of) participants and further studies should be conducted to confirm our results. Also, as the participation was anonymous, we did not have means to classify participants by their degree of knowledge on the topic or similar factors, thus, the ratings of all assessors were treated equal. In general, it is difficult for us to judge whether such studies should be conducted with specialized scholars, or with a more general audience such as students. For specialized scholars, it could be problematic that they might have too strong of a pre-exposure to the measures and might not be able to give unbiased answers (cf. the demand characteristics effect [9]). On the contrary, generalized audiences might have comprehension problems in logic [6] and might give unreasonable answers. In future work, it would be interesting to conduct similar studies with different groups. For this study, we considered both groups. The survey was designed in a manner accessible for general scholars, which was tested in a face-to-face pretest.

A further limitation is that the ratings of the assessors could depend on the phrasings or the order of the individual questions. Especially the transformation of formal measure or postulate definitions into natural language can introduce ambiguity. While we cannot rule this out entirely, we counteracted this problem by using the descriptions of measures and postulates as provided in published surveys, e.g. [12, 13]. In this context, a different approach would be to show participants a knowledge base and ask them to quantify the severity of inconsistency with a numerical number. Then, one could compare the numbers as provided by participants with actual measures. However, a severe problem with that approach is that many measures return the same numerical value for the same knowledge base (c.f. the related discussion on expressivity [12], or a recent experiment in [15], where those authors found that there were high correlations between the numerical assessments of some measures regarding identical knowledge bases). In result, such an approach would make it very difficult to draw connections between the numerical values and the reasons that were considered towards the severity of inconsistency. In our approach, this was possible based on the factor analysis.

An important aspect to bare in mind is that we asked for the *perceived* suitability (and nothing more). This is intuitively highly subjective and in no way allows to discard any measures.

5 Discussion

In this work, we conducted an empirical survey to measure the perceived suitability of inconsistency measures. Based on our results, an influence of concrete knowledge bases on the perceived suitability of specific measures could not be shown, thus we argue that inconsistency measures have some inherent form of perceived suitability.

In many cases, although specific measures satisfied the same basic postulates, they had a different degree of perceived suitability (and they were perceived this way for different reasons). Also, the drastic inconsistency measure $\mathcal{I}_d$, which satisfies the most postulates of the considered measures, had the lowest perceived suitability amongst participants. This shows that there is something more to evaluating inconsistency measures than rationality postulates. Our work could be used to extrapolate an evaluation method for future works, i.e., our experiments can be repeated in the context of specific application domains, however, much more work is needed towards the classification and evaluation aspect in inconsistency measurement (cf. e.g. [5]).

We also investigated the perceived plausibility of rationality postulates. The ranking of perceived plausibility regarding postulates presented in this work is intended to foster the discussion on which postulates can be seen as desirable and thus guide the development of new measures. For example, an interesting result is that free formula independence was least agreed upon by the participants, however, this postulate is satisfied by most formula-level measures. This could give way to investigate some form of "interestingness" of postulates for certain use-cases, cf. e.g. the discussion of postulates for relative inconsistency measures in [4]. In future work, it would be interesting to revise also other postulates, and to understand more clearly when postulates are desirable, cf. also [1].

The final question to now address is: what makes a suitable inconsistency measure? Our results would support to make propositions such as *"It should probably satisfy CO and DO"* (cf. Conclusion 3), *"It should probably be based on paraconsistent semantics and maybe not on maximal consistency"* (cf. Conclusion 2), or *"It should maybe not implement NO"* (cf. Conclusion 2, the two lowest ranked measures are the only ones to satisfy NO). However, we will not make such a requirements catalogue in this work. Rather, we want to advocate the consideration of specific application domains and new evaluation dimensions. Here, the research method introduced in this work yields novel (methodological) insights towards how results from the field of social sciences can be applied to study the cognitive adequacy of approaches in AI.

References

[1] Philippe Besnard. Revisiting postulates for inconsistency measures. In *European Workshop on Logics in Artificial Intelligence*, pages 383–396. Springer, 2014.

[2] Philippe Besnard. Forgetting-based inconsistency measure. In *10th International Conference on Scalable Uncertainty Management (SUM'16)*, pages 331–337, 2016.

[3] Philippe Besnard. Inconsistency measuring over multisets of formulas. In John Grant and Maria Vanina Martinez, editors, *Measuring Inconsistency in Information*. College Pub., 2018.

[4] Philippe Besnard and John Grant. Relative inconsistency measures. *Artificial Intelligence*, 280:103231, 2020.

[5] Glauber De Bona, John Grant, Anthony Hunter, and Sebastien Konieczny. Towards a unified framework for syntactic inconsistency measures. In *32nd AAAI Conference on Artificial Intelligence*, 2018.

[6] Geoffrey L Herman, Michael C Loui, Lisa Kaczmarczyk, and Craig Zilles. Describing the what and why of students' difficulties in boolean logic. *ACM Transactions on Computing Education (TOCE)*, 12(1):1–28, 2012.

[7] Anthony Hunter and Sébastien Konieczny. On the measure of conflicts: Shapley inconsistency values. *Artificial Intelligence*, 174(14):1007–1026, 2010.

[8] Anthony Hunter, Sébastien Konieczny, et al. Measuring inconsistency through minimal inconsistent sets. *KR*, 8:358–366, 2008.

[9] W Lawrence Neuman. Social research methods: Qualitative and quantitative approaches w. lawrence neuman, 2014.

[10] Katja Lozar Manfreda and Vasja Vehovar. Internet surveys. *International handbook of survey methodology*, pages 264–284, 2008.

[11] Alain Pinsonneault and Kenneth Kraemer. Survey research methodology in management information systems: an assessment. *Journal of management IS*, 10(2):75–105, 1993.

[12] Matthias Thimm. On the expressivity of inconsistency measures. *Artificial Intelligence*, 234:120–151, 2016.

[13] Matthias Thimm. On the compliance of rationality postulates for inconsistency measures: A more or less complete picture. *KI-Künstliche Intelligenz*, 31(1):31–39, 2017.

[14] Matthias Thimm. On the evaluation of inconsistency measures. In John Grant and Maria Vanina Martinez, editors, *Measuring Inconsistency in Information*. College Pub., February 2018.

[15] Matthias Thimm. An experimental study on the behaviour of inconsistency measures. In *13th International Conference on Scalable Uncertainty Management (SUM'19)*, December 2019.

[16] Matthias Thimm. Inconsistency measurement. In *13th International Conference on Scalable Uncertainty Management (SUM'19)*, December 2019.

[17] Iris Vessey and Dennis Galletta. Cognitive fit: An empirical study of information acquisition. *Information systems research*, 2(1):63–84, 1991.

[18] Guohui Xiao and Yue Ma. Inconsistency measurement based on variables in min. unsatisfiable subsets. In *European Conference on Artificial Intelligence 2012 (ECAI'12)*, page no, 2012.