Synthese (2016) 193:3293–3309
DOI 10.1007/s11229-015-0929-4

Evaluating competing theories via a common language of qualitative verdicts

Wulf Gaertner[1] · **Nicolas Wüthrich**[2]

Received: 5 March 2015 / Accepted: 22 September 2015 / Published online: 7 October 2015

Abstract Kuhn (The essential tension—Selected studies in scientific tradition and change, 1977) claimed that several algorithms can be defended to select the best theory based on epistemic values such as simplicity, accuracy, and fruitfulness. In a recent paper, Okasha (Mind 129(477):83–115, 2011) argued that no theory choice algorithm exists which satisfies a set of intuitively compelling conditions that Arrow (Social choice and individual values, 1963) had proposed for a consistent aggregation of individual preference orderings. In this paper, we put forward a solution to avoid this impossibility result. Based on previous work by Gaertner and Xu (Mathematical Social Sciences 63:193–196, 2012), we suggest to view the theory choice problem in a cardinal context and to use a general scoring function defined over a set of qualitative verdicts for every epistemic value. This aggregation method yields a complete and transitive ranking and the rule satisfies all Arrovian conditions appropriately reformulated within a cardinal setting. We also propose methods that capture the aggregation across different scientists.

Keywords Theory choice · Social choice theory · Scoring rules · Thomas S. Kuhn · Epistemic values

✉ Nicolas Wüthrich
n.wuethrich@lse.ac.uk

Wulf Gaertner
wulf.gaertner@uni-osnabrueck.de

[1] FB Wirtschaftwissenschaften, VWL/Theoretische Volkswirtschaftslehre, Universität Osnabrück, Rolandstr. 8, Raum 29/B03, 49069 Osnabrück, Germany

[2] Department for Philosophy, Logic, and Scientific Method, London School of Economics and Political Science, Houghton Street, WC2A 2AE London, UK

1 Introduction

Scientists face situations in which they have to choose among competing theories. Kuhn's (1977) paper is still providing the relevant stage-setting for discussing theory choice. Kuhn claimed that a variety of epistemic values, most importantly scope, fruitfulness, accuracy, simplicity, and consistency, influence theory choice. However, he rejected the idea that there is a unique algorithm to amalgamate the information provided by these criteria (Kuhn 1977, pp. 322, 326). Okasha (2011) gave a new twist to this debate. He proposed an analogy between social choice and theory choice. Given the analogy holds, a troubling impossibility result emerges: there exists no aggregation procedure which yields a complete and transitive ranking of the alternatives considered and which satisfies a set of intuitively compelling conditions. Hence, it seems that Kuhn's argumentation has to be reconsidered what primarily motivated this paper.

We suggest a way to avoid Okasha's impossibility result by viewing the theory choice problem in a cardinal context. By doing so, we pick up a suggestion which was made by Okasha in the latter part of his paper, namely to enrich the informational basis of the analysis to allow for inter-criteria comparability. Okasha refers to Sen (1977, 1986) who argued that some degree of interpersonal comparability is needed in order to get out of Arrow's impossibility impasse. This shift in the problem description allows us to use a tool which has been successfully applied in the social choice context, i.e. scoring rules. We argue that a general scoring rule characterised by Gaertner and Xu (2012) is flexible enough to illuminate, and hopefully solve, the problem of theory choice. This general scoring rule uses a set of qualitative verdicts. These take into account ordinal information from the epistemic values as well as information about the epistemic projects pursued in the process of constructing and evaluating theories. Furthermore, these qualitative verdicts permit, so our argumentation goes, a meaningful comparison across different criteria that are used to evaluate alternative theories. This inter-criteria comparability allows us then to avoid Arrow's (1963) and Okasha's impossibility result in a cardinal context. To be more precise, given a finite number of epistemic values and three or more alternative theories, the aggregation method we propose yields a complete and transitive ranking within a cardinal setting and thus satisfies Arrow's rationality postulate together with 'his' four conditions of unrestricted domain, weak Pareto, non-dictatorship, and independence of irrelevant alternatives appropriately reformulated for cardinal values.

The paper is structured as follows. First, we revisit Kuhn's discussion of theory choice (Sect. 2). Secondly, we briefly recall Okasha's reconstruction of Kuhn and refer to some very recent discussion (Sect. 3). Thirdly, we introduce the basic idea of our solution, present a formal characterisation, and show how our proposal satisfies the Arrovian conditions in a cardinal context (Sect. 4). Fourthly, we discuss what our solution reveals about the aggregation across different scientists (Sect. 5).

2 Revisiting Kuhn's discussion of theory choice

We revisit Kuhn's discussion of theory choice for two reasons. First, we want to state precisely what we take the problem of theory choice to be. Secondly, this discussion

allows us to clarify in Sect. 4 what, from our point of view, the status of a solution of the theory choice problem is.

Let us begin with the *problem of theory choice*. We take Kuhn's key claim to be the following: Given a finite set of theories X with more than one element and a finite set of epistemic values N (containing accuracy, consistency, scope, simplicity, and fruitfulness), different scientists can arrive at more than one ranking of the alternatives even if they agree that the evaluation should be done solely with reference to the set N. This is tantamount to saying that there is more than one algorithm to determine the overall ranking of the alternative theories based on the epistemic values (Kuhn 1977, pp. 322, 326).

This formulation entails two important clarifications of what we take to be the problem of theory choice. First, we are setting aside two arguments in Kuhn's earlier writings. The first argument can be put as follows: Theories, more precisely paradigms, come equipped with standards for assessing theories. These standards can vary across different paradigms. Hence, there is no unique way of choosing between paradigms (Kuhn 1996, pp. 6, 141). The second argument goes like this: Even if there are shared standards for assessing paradigms, paradigms do not solve an identical set of problems. Accordingly, if one chooses between two or more paradigms, one has to weigh the importance of the problems against each other. Hence, there is no unique way of choosing between paradigms (Kuhn 1996, pp. 85, 103; Hoyningen-Huene 1993, p. 242). Therefore, we only deal with the situation in which the set of problems as well as the standards of theory evaluation are shared.[1]

This leaves us with the following two arguments of Kuhn. To start, epistemic values can be interpreted differently by the scientists who are involved in the evaluation process (Kuhn 1977, p. 322). Hence, it is possible that two scientists, who are committed to the same set of epistemic values, come up with different rankings of the alternatives under consideration (Hoyningen-Huene 1993, p. 236). Furthermore, the epistemic values can be weighed differently by scientists who are evaluating theories (Kuhn 1977, p. 322; Hoyningen-Huene 1993, p. 236). The commitment to the same set of values does not entail or presuppose a commitment about their relative weight. Hence, it is again possible that two scientists, who are committed to the same set of epistemic values, come up with different rankings of the alternatives under consideration.

Secondly, the formulation entails a particular view regarding the level at which the problem of theory choice is located. To be more precise, is the theory choice problem an issue on the level of individual scientists or the scientific community? This is a tricky issue, but we think the problem occurs on both levels. Kuhn's own remarks on this give a somewhat mixed picture:

> (…) it is the community of specialists rather than its individual members that makes the effective decision. (Kuhn 1996, p. 200)

[1] We take it to be the case that Okasha (2011) as well as the replies in the literature, which will be discussed below, share this focus. It is worthwhile to note that thereby Kuhn's discussion of methodological incommensurability is left aside.

> (…) shared values can be important determinants of *group behaviour* even though the *members of the group* do not all apply them in the same way. (Kuhn 1996, p. 186, our emphasis)

How do we have to understand these remarks about decision making on the group level and how is it related to decision making of individual scientists?

We follow Hoyingen-Huene's (1993) interpretation here. Accordingly, the most relevant debate occurs on the level of the individual scientist. The choice of an individual scientist is influenced but not determined by the epistemic values. The individual scientist chooses a preferred theory and invests her energy in its development. Which theory comes out on top in the scientific community is determined by a historical process consisting of the choices of individual scientists (Hoyningen-Huene 1993, pp. 153–154). Subsequently, we first focus on the choice situation of an individual scientist (Sects. 3 and 4). Based on our procedure for this situation, we then discuss the aggregation across different scientists (Sect. 5).[2]

3 Okasha's Arrovian reconstruction of Kuhn and beyond

Since Okasha's paper is now fairly widely discussed, we will be brief here, solely remind the reader of the basic set-up and use the opportunity to introduce some notation.

Okasha treats each of Kuhn's epistemic values (simplicity, accuracy, fruitfulness, consistency, and scope) as if it were an individual with a preference ordering over the alternative theories. To be more precise, every epistemic value can be viewed as a decision criterion $n \in N$ (where N is the set of relevant criteria) which can be expressed as a binary relation R_n (e.g. 'is at least as simple as', 'has at least the scope as') defined on the set of alternative theories X. Each binary relation R_n imposes a weak ordering on X, i.e. it is reflexive, transitive, and complete (Okasha 2011, p. 91).

Given this framework, Kuhn's algorithm can be expressed as a theory choice rule (ibid., 92). A theory choice rule is a mapping from the set of all logically possible combinations of weak orderings $(R_1, \ldots, R_n, \ldots, R_t)$ to a single weak ordering R^* which is the aggregate relation defined on the set of alternative theories X and interpreted as "is at least as good as".

According to Okasha, all five requirements that Arrow (1963) postulated in his path-breaking work on the nonexistence of a social welfare function have to be met by a theory choice rule (ibid.). First of all, the aggregate relationship has to be a weak ordering, i.e. the aggregate relationship has to be transitive and complete. This is a requirement of collective rationality. Then the postulate of an unrestricted domain (**U**) means that the theory choice rule should yield an overall ranking R^* for all logically possible combinations of t-tuples of binary relations R_n (ibid.). The requirement of the weak Pareto principle (**P**) states that if theory T_1 does better than another theory T_2 with respect to all considered epistemic values $1 \ldots, n, \ldots, t$, then T_1 should be

[2] Note that Weber (2011) offers an alternative interpretation of Kuhn. He views Kuhn as a social epistemologist (ibid., 3) who treats the scientific community level as the relevant decision making entity (ibid., 7). Accordingly, he would reject our two step approach to the choice problem.

preferred to T_2 overall (ibid.). Arrow's non-dictatorship (**D**) requirement states that there is no epistemic value such that if this value ranks, for all profiles of preference rankings, any T_1 above any other T_2, T_1 is ranked automatically above T_2 in the overall ranking (ibid., 93). Finally, the independence of irrelevant alternatives (**I**) condition requires that the overall ranking of any T_1 and T_2 depends only on how the epistemic values rank T_1 and T_2 and not on how they rank other theories in relation to T_1 and T_2 (ibid.).

For this reconstruction of the theory choice problem, Arrow's famous impossibility result applies: For a finite number of epistemic values and at least three alternative theories, there exists no theory choice function satisfying conditions **U**, **P**, **D**, and **I** (ibid.).

Okasha's result has stimulated a variety of replies (Morreau 2014, 2015; Rizza 2014; Stegenga 2015; Weber 2011). All of them focus on the conditions which need to be fulfilled by a satisfactory aggregation mechanism according to the analogy to social choice theory. Weber (2011) suggests that the non-dictatorship assumption cannot be defended in the case of theory choice. Specifically, fruitfulness (defined as the capability to provide problem solutions) has to be viewed as a dictatorial criterion amongst the epistemic values. Morreau (2014, 2015) argues that the assumption of unrestricted domain does not hold for theory choice. Rizza (2014) argues that Okasha's result disappears if one uses the correct information encoded in the ordinal information of the epistemic criteria. In the situation of three alternative theories, using sequenced triples instead of triples of pairs of alternatives as input for the aggregation process avoids the potential for an intransitive overall ranking. Finally, Stegenga (2015) claims that Okasha's own solution, namely to enrich the information which can be fed into the aggregation mechanism, fails. The complexity and diversity of the epistemic values does not generally allow expressing more than ordinal information. Our solution strategy is in contrast to Stegenga's claim. We take up Okasha's idea to enrich the informational basis and propose a solution that allows for inter-criteria comparability within a cardinal set-up.

4 A new approach: using scoring functions over qualitative verdicts to establish comparability of theory choice criteria

Now that the stage has been set, we propose our solution to avoid the Arrovian impossibility result in the case of theory choice. First, we introduce the basic idea with the help of a decision situation in a committee (Sect. 4.1). Secondly, we explain how this basic idea can be made fruitful for the problem of theory choice. In particular, we motivate a reformulation of the theory choice problem in a cardinal context (Sect. 4.2). Thirdly, we capture the basic idea formally and highlight intuitively attractive features of this rule. Here, we also show how this rule is able to fulfil all of the appropriately reformulated Arrovian conditions in a cardinal context (Sect. 4.3). Finally, we further clarify our proposal by spelling out the notion of context-dependency (Sect. 4.4).

Before going on, let us specify what, from our point of view, the *status of a solution* to the theory choice problem is.

As we saw at the end of the last section, a common attempt to circumvent Arrow's impossibility result is to relax one of the requirements for the aggregation mechanism (**U**, **P**, **D**, and **I**). Stegenga (2015) nicely highlights that such a move requires a stance on what the standards are for evaluating these requirements (ibid., 265). He introduces three possible standards (ibid.):

(a) Normative adequacy (theory choice ought to satisfy the requirement)
(b) Complete descriptive accuracy (theory choice always satisfies the requirement)
(c) Partial descriptive accuracy (theory choice sometimes satisfies the requirement)

Stegenga's reflections also apply to our proposal. Is our solution meant to be a descriptively (fully or partially) accurate picture of theory choices in the sciences? Or is it a normative proposal stating how theory choices should be done in the sciences?

We do not aim at providing a descriptively accurate account of theory choice processes in the sciences. By this we mean that we are not claiming that scientists necessarily form their beliefs and come to their judgments as specified in our aggregation procedure. However, we do understand our procedure as a rational reconstruction of theory choice processes. To be more precise, our procedure belongs in the normative domain to the extent that it either can be viewed as one of multiple possible prescriptive procedures to arrive at an overall judgment in theory choice processes, or as a standard to judge outcomes of theory choice processes.

4.1 The basic idea

Imagine you are one of the members of a committee that has to decide among a certain number of research proposals for funding. Let us suppose that k proposals were submitted. Let us further assume that the chairperson of your committee comes forward with the following procedure. She declares that there are m categories (from excellent to fail, let us say, with $m-2$ categories in between), with rank scores from m to 1 attached to these categories. The chairperson asks all members of your committee to allocate the k proposals to the m available categories. It is explicitly not required that every member comes up with a strict ordering and that all categories have to be filled by each and every committee member. Furthermore, the chairperson announces that, as soon as each member has assigned the k proposals to the m categories, she would count the rank numbers assigned to each proposal and then construct a ranking over the k proposals from the highest rank sum to the lowest, the proposal with the highest aggregate sum being the winner, though more than one proposal may be selected depending on the available budget. We claim that this aggregation procedure can be made fruitful for the theory choice case. Let us flesh out this analogy in more detail.

4.2 Transferring the basic idea to the problem of theory choice

In the above scenario, replace the research proposals with alternative theories, the members of the committee with Kuhn's epistemic values, and the chairperson with an *individual* scientist. Furthermore, consider a set of discrete verdicts corresponding to

the categories: 'very high', 'high', 'satisfactory', 'just sufficient', and 'insufficient'.[3] This yields the following set-up for theory choice: A scientist considers the alternative theories in light of the epistemic values. For each epistemic value, she independently assigns to the alternative theories a qualitative verdict (e.g. 'T_1 is sufficiently accurate', 'T_2's accuracy is very high'). The five qualitative verdicts constitute a discrete scale with rank scores. The overall ranking of the theories is determined by the sum of rank scores of each alternative theory.

Drawing this parallel between the case of the committee and the problem of theory choice amounts to switching to a cardinal description of the theory choice problem. Note that the epistemic values, such as simplicity or accuracy, are assumed to only provide ordinal information in our proposal. The introduction of qualitative verdicts, and their respective expression in rank scores, is, hence, an additional step, which needs to be justified. By taking this step, we do not propose to solve Okasha's problem in an ordinal context. Furthermore, note that the general scoring rule which was introduced in the committee case is not identical to the Borda rule, since it does not presuppose that the epistemic values always provide a strict ordering of the alternative theories.

The presence of a cardinal scale is, in our point of view, an accurate description of theory choice situations for the following reasons. Theory choice does not take place in a vacuum. Rather, scientists are evaluating competing theories with respect to *broad epistemic projects* they are conducting in a discipline or a sub-discipline. For example, in engineering such a broad epistemic project could consist in building more efficient combustion engines. In molecular biology, a broad epistemic project could consist in synthesizing new functional germs. Finally, in astrophysics a broad epistemic project could be the understanding of the distribution of clusters of galaxies across the universe. These epistemic projects each define a particular *assessment context* with respect to which the evaluation of competing theories, models, or hypotheses takes place. Looking more closely at these assessment contexts reveals that such a context provides additional informational structure beyond the ordinal information given by the epistemic values. To be more precise, evaluating a theory in light of the epistemic values in an assessment context allows introducing statements about whether a theory is sufficiently simple, accurate, consistent, fruitful or broad in scope to contribute to the realisation of a particular epistemic project. We assume that the assessment context provides enough informational structure beyond the ordinal information encoded in the epistemic values to assign a sufficiently fine-grained set of qualitative verdicts.

To further motivate the plausibility of the move from ordinal information encoded in the criteria of theory evaluation to a cardinal scale, reconsider the case of the committee. The members of the committee also transfer their ordinal assessment of proposals into a cardinal scale. How are they doing this? Following the previous line of reasoning, they are capable of doing this mentally challenging task, which involves a certain degree of imprecision or granulation, by implicitly making assumptions about the research setting at the university or in the discipline more broadly. Accordingly, they are judging whether a research proposal is sufficiently well-structured or highly

[3] We will say more about the details of this scale. For the moment, the reader should not be irritated by the facts that we specify an exact number of grades and choose a particular formulation of the qualitative verdicts.

original in view of the expectations of the profession, such as the prospect of producing work which could be published in a peer-reviewed journal.[4]

Notice the work that the qualitative verdicts are doing. The qualitative verdicts stated in the rank scoring system impose a cardinal representation on the preference orderings over the alternative theories.[5] It is this *constructed* cardinal representation which allows inter-criteria comparison and, hence, enables us to avoid the Arrovian impossibility result. In order to make the cardinal scores for each of the theories comparable across the set of criteria and thereby to achieve inter-criteria comparability, the process of construction of the scale is of utmost importance. This is analogous to establishing a common grammar or language that creates some unifying basis for comparison.

In order to clarify more concretely our proposed approach, let us go back to the committee and its members again. Each individual has to transform his or her ordinal preference relation over the alternative proposals into a cardinal ranking with the requirement that if proposal x, let us say, is at least as good as proposal y, the cardinal rank or score attached to x is at least as high as the rank assigned to y so that for all $x,\ y \in X$, the following relationship holds: $xRy \leftrightarrow s(x) \geq s(y)$, where $s(x)$ stands for the cardinal value or score attached to x, and likewise for y. This is a very basic requirement in the sense that one must neither lose nor distort ordinal information when one makes a transition from the ordinal to the cardinal world. Furthermore, we postulate that score differences among the different alternatives are meaningful and comparable, so that for four alternatives $x,\ y,\ z,\ w$, let us say, one may come to the conclusion that $s_n(x) - s_n(y) > s_n(z) - s_n(w)$, where $s_n(x)$, for example, is the score assigned to alternative x by committee member n. Note that any affine transformation of these scores with a common positive scale factor over all n does not destroy this comparison of score differences.

Coming back to our problem of theory comparison, each scientist is assumed to examine the given theories in the light of the set of criteria that are relevant for the problem at stake. More precisely, each scientist starts for each single criterion with an ordinal ranking over the theories to be evaluated and then transforms this ranking into a sequence of cardinal scores according to the specified relationship from above. We assume that the scientist can translate the ordinal into the cardinal information for every epistemic value in isolation. Accordingly, assigning an alternative x a rank score with respect to simplicity, let us say, is independent from assigning x a rank score with respect to accuracy, for example. This involves the assumption that the epistemic values are independent of each other. We think that this is, first, in line with Kuhn's discussion

[4] Let us add two further clarifications to our motivation for a move from the ordinal to the cardinal world. First, note that the informational content of the assessment context (i.e. the broad epistemic projects and its features) need to be distinct from the epistemic values. If this informational content were identical to the Kuhnian epistemic values or could be rephrased in terms of additional epistemic values, then, given our problem set-up, it would only provide ordinal information and, hence, we would not be able to justify the move to a cardinal scale in this way. Secondly, let us emphasize that we do not presuppose that the epistemic values directly provide cardinal information. Rather, we assume an additional and mentally challenging step by the scientist which involves the careful consideration of the assessment context of a theory choice problem.

[5] See Pivato (2014, p. 50) for a similar discussion of the possibility to impose cardinality on a ranking of alternatives.

and, secondly, even if there are (conceptual or empirical) dependency relations between the values, these relations might not hold under all possible interpretations of these values which should be considered when one discusses theory choice on this level of abstraction.

The result we get is attractive. Given a finite number of epistemic values and three or more alternative theories, the proposed aggregation method yields a complete and transitive ranking and the rule satisfies the four Arrovian conditions of unrestricted domain (**U**), weak Pareto (**P**), non-dictatorship (**D**), and independence of irrelevant alternatives (**I**) within a cardinal setting. We will show carefully how these results are achieved once we have formally specified our proposal (see Sect. 4.3).

Since the qualitative verdicts are establishing the inter-criteria comparability, let us motivate them further. First, the success of our solution does not depend on the particular formulation of qualitative verdicts we have introduced above. We can allow for a more fine-grained and less fine-grained set of qualitative verdicts. In addition, different formulations of the qualitative verdicts could be chosen for different areas in science. However, we take it to be the case that the following requirements have to be fulfilled by a plausible set of qualitative verdicts: (a) the qualitative verdicts need to be framed in evaluative terms. The evaluative terms transport a substantial meaning which can be made sense of in the context of theory choice (e.g. a *highly* fruitful theory). Furthermore, using evaluative terms instead of going directly for the rank scores implies the commitment to justify the ascription of a particular evaluative term; (b) the evaluative terms need to suggest a natural ranking amongst them; (c) the evaluative terms need to make sense with respect to every epistemic value under consideration.

We should mention that the idea to introduce qualitative verdicts and thereby establish a common language or "grammar" was made before by Balinski and Laraki (2007, 2010). Their proposal of preference aggregation, called majority judgment, however, completely remains within the framework of ordinal information and therefore systematically differs from our own approach.

Second, let us provide some plausibility for the claim that our five qualitative verdicts 'very high', 'high', 'satisfactory', 'just sufficient', and 'insufficient' are indeed applicable to the five Kuhnian epistemic values. From our point of view, it is fairly obvious that accuracy, scope, fruitfulness and simplicity can be fulfilled to a greater or lesser extent. It is less obvious for consistency. Under the heading of this epistemic value, Kuhn discusses internal consistency (i.e. the theory is free of any contradictions) and external consistency (i.e. the theory does not contradict already accepted theories) (Kuhn 1977, p. 321). We accommodate internal and external consistency in our set of qualitative verdicts as follows: If the theory contains no contradictions and does not entail contradictions with already existing theories, it receives the verdict 'very high', otherwise it receives the verdict 'insufficient'. Accordingly, we treat 'consistency' as a matter of binary choice.[6]

[6] Notice that we do not treat the qualitative verdict 'insufficient' as an eliminative verdict in the sense that whenever with respect to one epistemic value an alternative gets an 'insufficient' verdict, this alternative is eliminated from the choice set. The reason for this is that we, in line with Kuhn, do not share the intuition that one of the epistemic values should be treated as a killer criterion. Accordingly, even if a theory is

Thirdly, the talk about theory choice problems in an assessment context introduces rather straightforwardly a notion of context-dependency. Let us be very clear what dependency relations we have in mind here.

In our framework, the graininess of the partition can vary between assessment contexts. Furthermore, the prerequisites in order to assign the qualitative verdict of 'sufficient', let us say, to a particular theory can vary among the epistemic values and may depend on the assessment context. What is important, however, is that if the scientist has come to the conclusion that two criteria are sufficiently fulfilled with respect to any particular theory or across two or more theories, then, simply said, 'sufficient means sufficient'. Otherwise, the inter-criterion comparability would not be given. Admittedly, it is a rather complicated mental exercise that we expect the scientist to perform. For example, the scientist may have reasons to say that the requirements to assert that sufficiency is given, may be quite high in relation to one criterion but less rigorous with respect to another criterion. But once sufficiency (of fulfilment) is registered or stated by two or several criteria, it has the same meaning everywhere and enters, via its constructed cardinal representation, directly as input into the aggregation function.

We acknowledge that these claims are fairly abstract. We will illustrate them in Sect. 4.4 with the help of hypothetical examples. But first, we provide a precise formal discussion of the aggregation rule we just have introduced.

4.3 A formal characterisation of our solution

The following formal presentation is based on Gaertner and Xu's (2012) characterisation. Let X be the universal set of scientific theories containing a finite number of elements. Let N be the set of criteria deemed relevant with $t > 1$. Let $E = \{1, \ldots, E\}$, with the cardinality of this set being larger than one, be a set of given positive integers from 1 to E. These integers will in most cases be assumed to be equally distanced and are thought to represent qualitative statements thus constituting a common language of evaluation, as outlined above.[7]

A scoring function $s_i : X \rightarrow E$ is chosen for each criterion $i \in N$, such that, for all $x \in X$, $s_i(x)$ indicates the score that criterion i assigns to x. Let S_i be the set of all possible scoring functions for criterion i. As explained in the last section, the statement how well or how badly a theory fares given in light of a criterion has to be inserted in the commonly given scale constituted by set E.

Let P be the set of all orderings over X. A profile $s = (s_1, \ldots, s_t)$ is a list of scoring functions, one for each criterion. An aggregation rule f is defined as a mapping: $S_1 \times \ldots \times S_t \rightarrow P$. Let $S = S_1 \times \ldots \times S_t$.

Footnote 6 continued
insufficiently simple, let us say, but receives the best qualitative verdicts with respect to all other criteria such that the aggregate rank score is the highest of all alternatives, the low grade in terms of simplicity has to be seen in relation to the high grades obtained from the other criteria.

[7] In general, the required minimal level of graininess depends on the particular theory choice problem at hand. However, if one criterion (i) ranks all alternative theories in a strict order, then due to $x P_i y \leftrightarrow s_i(x) > s_i(y)$, the minimal level of graininess is the number of alternative theories.

f is said to be an E-based scoring rule, to be denoted by f_E, iff, for any $s \in S$, and any $x, y \in X$, it is the case that

$$x \succsim y \Leftrightarrow \sum_{i \in N} s_i(x) \geq \sum_{i \in N} s_i(y)$$

where $\succsim = f(s)$. The asymmetric and symmetric parts of $\succsim$ will be denoted by $\succ$ and $\sim$, respectively.[8]

We are now in the position to articulate precisely how the Arrovian conditions **U**, **D**, **I**, and **P** are fulfilled.

Our aggregation procedure does not restrict the set of possible binary relations R_n over the alternative theories. Hence, unrestricted domain (**U**) is satisfied. The Pareto condition **(P)** is satisfied quite naturally. If all epistemic values find T_1, let us say, better than T_2, then according to our earlier relationship that $xRy \leftrightarrow s(x) \geq s(y)$ and $xPy \leftrightarrow s(x) > s(y)$, respectively, the Pareto condition requires that T_1 is better than T_2 in the "world" of cardinal information. Furthermore, the non-dictatorship (**D**) requirement states that there is no epistemic value such that if it ranks T_1 above T_2, T_1 is ranked automatically above T_2 in the overall ranking. This is satisfied by our general scoring rule because the general scoring rule weights the rank scores of each criterion equally.[9] Finally, the independence of irrelevant alternatives (**I**) requirement means that the overall ranking of T_1 and T_2 depends only on how the epistemic values rank T_1 and T_2 and not on how they rank other alternatives. Our proposed aggregation procedure satisfies condition **I** reformulated within the cardinal context.[10] Verbally, it requires that if two theories T_1 and T_2 receive precisely the same rank scores from the different epistemic values in the case of two separate evaluations, then the aggregate judgment over T_1 and T_2 is identical between the two evaluations. Given this set-up, it is irrelevant for the aggregate ranking of T_1 and T_2 how other theories are evaluated in the two evaluations. Let us illustrate the cardinal version of the independence condition with the help of a simple example. Consider the following two evaluations of three theories T_1, T_2 and T_3 based on three epistemic values where the ranks or scores in the left column are embedded in a simple integer-valued, equally-spaced interval scale:

[8] This E-based scoring function can be characterised in a fairly simple way. Please see the Appendix for a formal characterisation.

[9] Our approach can allow different weights of the epistemic values. A natural way to account for this difference is to divide a criterion (e.g. 'fruitfulness') into two sub-criteria (e.g. 'fruitfulness with respect to the discipline' and 'fruitfulness with respect to neighbouring disciplines'). In this way, the initial criterion gets, quite naturally, a higher weight in our summation procedure.

[10] A discussion of the reformulated version of the independence requirement in the context of the utilitarian rule can be found in Gaertner (2013, pp. 125–126).

Evaluation 1				Evaluation 2			
4				4	$\boldsymbol{T_3}$	$\boldsymbol{T_3}$	
3	T_1	T_3	T_3	3	T_1		
2	T_2	T_1	T_2	2	T_2	T_1	T_2
1	T_3	T_2	T_1	1		T_2	T_1
0				0			$\boldsymbol{T_3}$

According to our general scoring function, in the first evaluation T_1 (with an associated total rank score 6) is strictly preferred to T_2 (having an associated total rank score 5). The exact scores with respect to T_1 and T_2 are precisely retained in the second evaluation. As this is the case, the aggregate relations between the two theories are exactly the same in the two evaluations despite the fact that an irrelevant alternative (here T_3) is positioned differently in the two situations.

Furthermore, the suggested aggregation procedure is sensitive towards the degree of criteria fulfilment. This is, in social choice theory, denoted as a form of positive responsiveness or positive association in the sense that a unilateral change in the fulfilment of some criterion in favour of x, let us say, should be reflected on the aggregate level in the same and not in the opposite direction.

The model also satisfies a property that is sometimes called consistency, at other times reinforcement (Young 1974), demanding that if the set of criteria is split up into two parts and a certain theory wins in both subsets, then this theory must also win in relation to the complete set of criteria.

Finally, it seems to us that the method proposed is superior to the Borda rule. While the latter rule requires that each and every criterion rank the alternative theories in a linear order, such a high degree of uniformity is not demanded by the method proposed here. Different criteria can rank or rather assign scores to the given alternatives in completely different ways as explained previously. We consider this as an advantage since the single criterion has more flexibility 'to articulate its preferences', i.e., it has more flexibility to express to what degree or extent it finds itself represented among the various theories under consideration.

4.4 Spelling out context-dependency in our proposal

In this section, we provide further motivation for our proposal of shared qualitative verdicts in the context of theory evaluation. So far we have argued that the theory choice problem should be viewed in a cardinal context and that the qualitative verdicts can be imposed on all of the Kuhnian values (see Sect. 4.2). Now we show that the set of qualitative verdicts allows for considerable flexibility by spelling out the element of context-dependency. We do this in two steps. First, we motivate the claim that the *level of graininess* of the qualitative verdicts can vary among assessment contexts. Secondly, we argue that the *prerequisites in order to reach a particular qualitative verdict* can change across epistemic values and assessment contexts.

To start, think about the following two hypothetical examples of assessment contexts. **Martina**, a particle physicist at the CERN, the European Organisation for

Nuclear Research, evaluates two theories about the structure of the decay of Higg's Bosons. **Tom**, a sociologist, evaluates two theories about the causes of the recent increase in immigration to the United Kingdom. Martina and Tom are each using the five Kuhnian values to reach an overall ranking of theories. They impose a set of qualitative verdicts on their ordinal preferences for each criterion. To do this, Martina and Tom might be using different sets of qualitative verdicts. Martina might work, in light of the small differences in the content of the theories and the necessary precision of the predictive tasks, with a five-item scale. In contrast, Tom might be using a three-item scale which is appropriate to deal with the recent aggregate data on immigration flows.

Let us turn to the prerequisites to reach a particular qualitative verdict. Assume that Martina is using the following verdicts: 'very high', 'high', 'sufficient', 'just sufficient', and 'insufficient'. When she assigns the competing theories to the qualitative verdicts, she reviews the ordinal information provided by every epistemic value. Now, to do this pairing of theories and verdicts it is, as we asserted earlier in Sect. 4.2, absolutely necessary that a qualitative verdict (e.g. 'insufficient') means the same for every epistemic value. The prerequisites to reach a particular qualitative verdict can, however, be quite different. With respect to accuracy, for example, 'high' could refer to a specific number of decimals at which the prediction of a theory matches the data. With respect to simplicity, 'high' could denote the fact that a theory allows stating the key differential equation for the system under study in closed form. The same could be argued for the other qualitative verdicts.

What about the prerequisites to reach a particular verdict in different assessment contexts? Assume that Tom and Martina are using the same set of verdicts, as specified in the previous paragraph. Furthermore, assume that both of them are attaching the verdict 'high' to one of their theories with respect to accuracy. Since their application contexts (particle physics vs. sociology) differ substantially, the reasoning behind the respective verdicts can differ. Martina could interpret it as a specific number of decimals at which the prediction of a theory matches the data. In contrast, Tom could refer to the fact that one of his theories is able to reflect qualitatively what people have reported in narrative interviews.

5 Aggregation across scientists

What we have described in the last section essentially is an aggregation procedure to which an individual scientist may resort in order to establish an ordering over a number of competing theories that satisfy a given set of epistemic values in different ways or to a different degree. Different scientists will normally come up with different orderings over the theories to be evaluated. Kuhn (1977, p. 325) writes that "every individual choice between competing theories depends on a mixture of objective and subjective factors, or of shared and individual criteria" so that the construction of a unique algorithm for theory choice is very difficult, if not impossible. He went on saying that

> (…) I have conceded that each individual has an algorithm and that all their algorithms have much in common. Nevertheless, I continue to hold that the

> algorithms of individuals are all ultimately different by virtue of the *subjective* considerations with which each must complete the objective criteria before any computations can be done. (ibid., 329, our italics)

The interesting aspect of our formal approach is that each and every scientist can have his or her own scoring function in order to generate an ordering over alternative theories. But not only this. Different scientists could choose different degrees of graininess with respect to qualitative verdicts. One person could have three verdicts, let us say 'high', 'sufficient' and 'insufficient', another person could decide just to pick two verdicts, namely 'high' and 'insufficient'. If this is the case, the two scientists can come up with different rankings of the theories, as in the following example.

Consider **Anna** and **Peter**, two scientists who evaluate two theories T_1 and T_2 with two different sets of qualitative verdicts:

Qualitative verdicts	Assigned Rank scores	Criterion 1	Criterion 2	Criterion 3
Anna				
High	3	T_1		
Sufficient	2		T_2	T_2
Insufficient	1	T_2	T_1	T_1
Peter				
High	2	T_1	T_2	T_2
Insufficient	1	T_2	T_1	T_1

Notice that in terms of purely ordinal information, Anna and Peter reveal the same preference ordering. Using our scoring procedure, Peter strictly prefers T_2 over T_1 whereas Anna is indifferent between T_1 and T_2.

As can be seen, Anna and Peter assign T_1 and T_2 to different qualitative verdicts. Reasons for their disagreement in the assignment of qualitative verdicts might be that Anna and Peter interpret the qualitative verdicts differently and/or that they interpret the epistemic values in a different way, explanations which come close to the point of subjectivity that Kuhn was making in the quotations at the beginning of this section.

All this can happen. A person may never assign the grade 'very good'—for this person 'good' is the best ever. Another person may be easily satisfied and, therefore, assign the grade 'very good' quite often. As we stated before, once the grades have been assigned, they have to be taken at face value. One should not come up asserting that the grade 'good' of one person is equivalent to the grade 'very good' of another person. We just simply lack this information. Such statements could become acceptable only under special circumstances where one has detailed information about the personality and psychology of different persons, which normally is not the case.

Nevertheless, once an overall verdict among a group of different scientists is found necessary, there is need for a mechanism that aggregates across individual evaluations.

The procedure put forward in this paper establishes, for each evaluating person, an ordinal ranking over the set of alternative theories at stake. Various methods are available to aggregate these orderings, all of which violate at least one of Arrow's requirements. The Borda rule is one candidate, approval voting (Brams and Fishburn

1983) would be a second, plurality voting a third, Condorcet's pairwise majority voting rule a fourth. Subsequently, we want to offer some considerations for why the Borda rule might be an appropriate procedure for the aggregation across different scientists.

To start, notice that it does not make sense to use our general scoring rule at this stage. To do so, would be asking the scientists to use a set of qualitative verdicts to transform their ordinal rankings into cardinal information. However, this would amount to saying that the scientists need to go back to their individual assessment of the alternative theories instead of taking the outcome of this analysis at face value. Furthermore, the Borda method has an advantage in relation to plurality and approval voting, namely, its aggregation procedure uses a lot of positional information that both plurality and approval voting ignore. While the plurality rule restricts itself to using information on the top element within each person's evaluation only so that the ranking of all other options is ignored, approval voting implicitly constructs two indifference classes, the set of acceptable options and the set of unacceptable alternatives, with no further differentiation in either set. Finally, the Borda method might be preferred to Condorcet's pairwise majority voting since the former guarantees that we never get a cyclical ranking of the alternative theories on the level of the scientific community.

6 Concluding remarks

In this paper, we proposed to avoid Arrow's impossibility result in the realm of multi-criteria theory choice by making use of a general scoring function. We have shown that a cardinal description of the problem of theory choice can be motivated and that based on this a set of qualitative verdicts can be brought into play which allow arriving at a complete and transitive relationship over the alternative theories without violating unrestricted domain, non-dictatorship, weak Pareto, and independence of irrelevant alternatives in a cardinal context. In a final step of the paper we argued that our solution can capture Kuhn's statements about the role of subjective factors in the theory choice process. We claimed that if the scientists who are involved in the evaluation procedure agree on a common set of qualitative verdicts, an aggregate ranking of the alternative theories can be achieved. We briefly discussed the Borda rule as a fruitful method of aggregating the rankings across different scientists.

We stated clearly that our proposal should not be read as a descriptively accurate account of how theory choice is done in the sciences. We take it to be an interesting and open question how the proposed solution can be linked to actual scientific practice. Answers to this question will also allow to flesh out the implications of our proposal for the rationality of science; an issue which we have deliberately set aside here. In the context of the Arrovian impossibility, we believe that the proposed solution is a promising step forward.

Acknowledgments We would like to thank an anonymous reviewer for helpful comments regarding the presentation of our overall argument. Furthermore, Claus Beisbart, Georg Brun, Kamilla Buchter, Gregory Fried, Stephan Güttinger, Paul Hoyingen-Huene, Jurgis Karpus, Simon Lohse, Alex Marcoci, James Nguyen, and Mantas Radzvilas provided fruitful feedback on earlier versions of this paper.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

To formally characterise the E-based scoring function, we need to introduce three concepts: (i, j)-variance, a monotonicity condition, and the property of cancellation independence.[11]

For any $s, s' \in S$, any $i, j \in N$, and any $x, y \in X$, we say that s and s' are (i, j)-*variant* with respect to (x, y) if $s_k(x) = s'_k(x)$ and $s_k(y) = s'_k(y)$ for all $k \in N \setminus \{i, j\}$.

Let us now introduce two properties of an aggregation rule f.

Monotonicity (M). For all $s \in S$ and all $x, y \in X$, if $s_i(x) \geq s_i(y)$ for all $i \in N$, then $x \succsim y$ and if $s_i(x) \geq s_i(y)$ for all $i \in N$ and $s_j(x) > s_j(y)$ for some $j \in N$, then $x \succ y$.

Condition M is a simple vector dominance condition. It requires that, in ranking two theories x and y, if the score assigned to x by each criterion i ∈ N is at least as great as the score assigned to y by the same criterion i, then x must be ranked at least as high as y by the evaluating scientist, and if in addition, some criterion assigns a higher score to x than to y, then x must be ranked higher than y.

Cancellation Independence (CI). For all $s, s' \in S$, all $x, y \in X$ and all $i, j \in N$, if s and s' are (i, j)-variant with respect to (x, y), $s_i(x) - s_i(y) = a$, $s_j(y) - s_j(x) = b$, $s'_i(x) = s_i(x)$, $s'_j(y) = s_j(y)$, $s'_i(y) = s_i(y) + \gamma$ and $s'_j(x) = s_j(x) + \gamma$ where $\gamma = min(a, b)$ when $a \geq 0$ and $b \geq 0$ and $\gamma = max(a, b)$ when $a < 0$ and $b < 0$, then $x \succsim y \leftrightarrow x \succsim' y$, where $\succsim = f(s)$ and $\succsim' = f(s')$.

Condition CI makes use of the fact that for any pair of alternatives, rank differences of opposite sign can be reduced without changing the aggregate outcome of the ranking procedure. This reduction procedure is performed in a stepwise fashion, starting with any two theories x and y, let us say, and picking any two criteria whose rank differences for x and y are of opposite sign. The "net" rank difference between x and y for this pair of criteria is determined. Then another criterion is picked whose rank difference for x and y is opposite in sign to the net rank difference of the first two criteria. The new net rank difference for x and y is calculated and the next criterion is picked whose rank difference again is opposite in sign to the just determined net rank difference with respect to x and y, if there is still one such criterion, and so on.

In Condition CI, vectors s and s' define scoring profiles that are aggregate-rank equivalent with respect to any pair of scientific theories. We call s' an s-reduced scoring profile. Condition CI therefore requires that $f(s)$ and $f(s')$ order any x and y in exactly the same way. Note that Condition CI makes an implicit assumption about an inter-criterion comparison of scores for which a common language is required.

Theorem. $f = f_E$ if and only if f satisfies the properties of Monotonicity and Cancellation Independence.

A proof of this result can be found in Gaertner and Xu (2012).

The theorem above establishes, for each evaluating person, an ordering over the set of alternative theories if that person has consented to a common language.

[11] See Pivato (2014) for an alternative axiomatic characterisation of this scoring function.

References

Arrow, K. J. 1963 [1951]. *Social choice and individual values*, 2nd edn. New York: John Wiley.

Balinski, M., & Laraki, R. (2007). A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences*, *104*, 8720–8725.

Balinski, M., & Laraki, R. (2010). *Majority Judgment: measuring, ranking and electing*. Cambridge, MA: The MIT Press.

Brams, S. J., & Fishburn, P. C. (1983). *Approval voting*. Boston: Birkhäuser.

Gaertner, W. (2013). *A primer in social choice theory*. Oxford: Oxford University Press.

Gaertner, W., & Xu, Y. (2012). A general scoring rule. *Mathematical Social Sciences, 63*, 193–196.

Hoyningen-Huene, P. (1993). *Reconstructing scientific revolutions: Thomas S. Kuhn's philosophy of science* (Transl. by A. T. Levine). With a foreword by Thomas S. Kuhn. Chicago: University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension—Selected studies in scientific tradition and change* (pp. 320–339). Chicago: The University of Chicago Press.

Kuhn, T. S. 1996 [1962]. *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Morreau, M. (2014). Mr. Fit, Mr. simplicity and Mr. scope: From social choice to theory choice. *Erkenntnis*, *79*, 1253–1268.

Morreau, M. (2015). Theory choice and social choice: Kuhn vindicated. *Mind*, *124*(493), 239–262.

Okasha, S. (2011). Theory choice and social choice: Kuhn vs. Arrow. *Mind*, *129*(477), 83–115.

Pivato, M. (2014). Formal utilitarianism and range voting. *Mathematical Social Sciences*, *67*, 50–56.

Rizza, D. (2014). Arrow's theorem and theory choice. *Synthese*, *191*, 1847–1856.

Sen, A. (1977). On weights and measures: Informational constraints in social welfare analysis. *Econometrica*, *45*, 1539–1572.

Sen, A. (1986). Social choice theory. In K. J. Arrow & M. D. Intriligator (Eds.), *Handbook of mathematical economics* (Vol. 3, pp. 1073–1181). Amsterdam: North-Holland.

Stegenga, J. (2015). Theory choice and social choice: Okasha versus Sen. *Mind*, *124*(493), 263–277.

Weber, M. (2011). Experimentation versus theory choice: A social-epistemological approach. In H. B. Schmidt, D. Sirtes, & M. Weber (Eds.), *Collective epistemology* (pp. 1–26). Frankfurt: Ontos Verlag.

Young, P. H. (1974). An axiomatization of Borda's rule. *Journal of Economic Theory*, *9*, 43–52.