Judgment and Decision Making, Vol. 6, No. 8, December 2011, pp. 711–721

The empirical content of theories in judgment and decision making: Shortcomings and remedies

Andreas Glöckner* Tilmann Betsch†

Abstract

According to Karl Popper, we can tell good theories from poor ones by assessing their *empirical content* (empirischer Gehalt), which basically reflects how much information they convey concerning the world. "The empirical content of a statement increases with its degree of falsifiability: the more a statement forbids, the more it says about the world of experience." Two criteria to evaluate the empirical content of a theory are their *level of universality* (Allgemeinheit) and their *degree of precision* (Bestimmtheit). The former specifies how many situations it can be applied to. The latter refers to the specificity in prediction, that is, how many subclasses of realizations it allows. We conduct an analysis of the empirical content of theories in Judgment and Decision Making (JDM) and identify the challenges in theory formulation for different classes of models. Elaborating on classic Popperian ideas, we suggest some guidelines for publication of theoretical work.

Keywords: empirical content, theory of science, critical rationalism, methodology, formalization, falsification, critical testing.

1 Introduction

This paper is about the benefits of being wrong. The rise of modern sciences began with the falsification of a theory that has been considered the undisputable truth for centuries: heliocentrism. Copernicus, Kepler, and Galileo provided empirical evidence that falsified the notion that earth is the center of the universe. Their work gave rise to the empirical sciences governed by the notion that a proposition has to stand the ground against reality in order to be accepted.

If we really want to know whether a rule or regularity holds in the real world, we must not be satisfied with instances of verification. We have to seek critical tests to approximate truth.

This may appear to readers as a truism. One would expect scientists to seek strong rules (laws) and put them to a critical test. However, we observe a trend towards the very opposite direction, at least in our field of judgment and decision making. Many theories are weakly formulated. They do not come up with strong rules and are, at least to some extent, immune to critical testing. Therefore, the number of theories that peacefully coexist in the literature is constantly growing. How many theories address judgment and decision making? Which one is superior to the others and makes the best predictions? We do not know, and neither may many readers of this article. We believe that theory accumulation and coexistence of weak theories has its origins in the fact that we lack shared standards for theory formulation and publication. We come up with some suggestion for a remedy. To lay the ground for these, we start with a summary of quite an old contribution made by Karl Popper decades ago.

Popper (1934/2005) suggested that, prior to any empirical testing, scientific theories should be evaluated according to their *empirical content* (empirischer Gehalt), that is, to the amount of information they convey concerning the world. He argues that "[t]he empirical content of a statement increases with its degree of falsifiability: the more a statement forbids, the more it says about the world of experience" (p. 103). Two criteria that determine the empirical content of a theory are their *level of universality* (Allgemeinheit) and their *degree of precision* (Bestimmtheit). The former specifies to how many situations the theory can be applied. The latter refers to the precision in prediction, that is, how many "subclasses" of realizations it allows. As we will describe more formally later, the degree of precision is thereby used as a technical term for how much a theory forbids in the situations to which it can be applied. It does not refer to a lack of ambiguity whether a prediction follows from the theory, because an ambiguous theory would not be considered a theory at all in the strict Popperian sense. To avoid confusion, we adopt the translation of this term from the English edition of Popper's writings, although it might be

*Andreas Glöckner, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10,D-53113 Bonn, Germany. E-mail: gloeckner@coll.mpg.de.

†University Erfurt

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

debated whether it conveys the exact meaning of the concept. An alternative translation could be *specificity*.[1]

The empirical content of a theory is, a priori, independent of empirical testing and therefore also independent of the theory's *degree of corroboration* (Bewährung).[2] Accordingly, falsifiability is a feature of a theory that can be assessed by means of logical analysis. It pertains to the empirical content of a theory which does not require conducting empirical tests. In the current paper, we adopt the Popperian perspective and reflect on the empirical content of models in judgment and decision making (JDM). We presuppose two things: a) theories (in the Popperian sense) are necessary to generalize findings beyond observed data and therefore theory development is an important part of the daily work of scientists; and b) some central aspects of Poppers' philosophical theory of science—particularly concerning the definition of empirical content—are valid.

Note that both assumptions are very basic and do not even necessitate accepting the Popperian approach in general. It should, however, also be noted that *other parts* of Popper's philosophical theory of science have been debated and/or extended in subsequent and contemporary philosophical and methodological writing. Particularly noteworthy are the strong inference approach (Platt, 1964) and Bayesian methods to evaluate the degree of corroboration of hypotheses and theories (e.g., Horwich, 1982; Wagenmakers, 2007). These developments are also reflected in work on JDM (e.g., strong inference using a critical property testing approach: Birnbaum, 1999; Birnbaum, 2008a, 2008b; and for applications of Bayesian approaches to JDM, see Bröder & Schiffer, 2003a; Lee & Newell, 2011; Matthews, 2011; Nilsson, Rieskamp, & Wagenmakers, in press). We will discuss these developments and core points of criticism in the last section of this paper and aim to rule out some misunderstandings.

2 Evaluation of Theories

2.1 Development and formulation of scientific theories

Popper suggests a standardized process of developing theories and selecting among them prior to empirical testing. It starts with putting forward a tentative idea from which conclusions are drawn by means of deductive logic. Then: (I) these conclusions are compared amongst themselves to test internal consistency; (II) the logical form of the system of conclusions is tested whether they have the character of a scientific theory; and (III) they are compared with predictions of existing theories to determine whether establishing them would constitute a scientific advantage.

Only after a theory has successfully passed evaluation on the logical level can it be subjected to critical empirical testing. In the current paper, we will focus on the steps that come *before* empirical testing and start our discussion with Poppers third criterion.

2.2 Scientific advantage (III): Empirical content, universality and precision

Popper argues in favor of replacing inductive with deductive testing of theories (later on elaborated by Hempel & Oppenheim, 1948). The question why a phenomenon happened "is construed as meaning according to what general laws, and by virtue of what antecedent conditions does the phenomenon occur" (Hempel & Oppenheimer, 1948, p. 136). A theory can be translated into a set of premises (law-like statements, definitions, etc.) that, given a set of antecedent conditions, allow for deriving predictions concerning empirical phenomena by logical deduction. According to Popper's notation, theories can therefore be written as a set of general implications of the form $(x)(\varphi(x) \rightarrow f(x))$, which means that all values of x which satisfy the statement function $\varphi(x)$ also satisfy the statement function $f(x)$. Both statement functions can thereby consist of multiple elements or conditions (connected by logical operators such as AND and OR).

Level of Universality. A theory[3] has a high level of universality if the antecedent conditions summarized in $\varphi(x)$ include as many situations as possible. An expected value (EV) model, for example, which predicts "when selecting between gambles people choose the gamble with the highest expected value", has a higher universality than a priority heuristic (PH; Brandstätter, Gigerenzer, & Hertwig, 2006). PH has multiple antecedent conditions that all have to be fulfilled and therefore reduce universality. Specifically, PH predicts "if people choose between pairs of gambles, *and* if the ratio of the expected value of the two gambles is below 1:2, *and* if neither gamble dominates the other then people chose the gamble with the higher minimum payoff".[4] The class "decisions between

[1] We thank Jon Baron for pointing us to this issue.

[2] Although often misunderstood, Popper explicitly stated: "We have introduced falsifiability solely as a criterion of the empirical character of a system of statements. As to falsification, special rules must be introduced which will determine under what conditions a system is to be regarded as falsified" (Popper, 1934/2005; p. 66).

[3] In this article, if not explicitly stated otherwise, we use "theory" in a broad sense, referring to any explicated set of general implications of the form $(x)(\varphi(x) \rightarrow f(x))$, independent of whether they fulfill the criteria of a scientific theory and are consistent or not.

[4] The PH includes further steps and criteria that are discussed in more detail elsewhere (Brandstätter, et al., 2006; for a formalization see also Glöckner & Betsch, 2008a).

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

pairs of gambles with a limited ratio in expected values and without dominance", for which PH is defined, is a subclass of the class "all decisions between gambles", for which EV is defined. Therefore, EV has a higher level of universality than the PH.

Degree of Precision. A theory's degree of precision increases with the specificity of the predicted phenomenon or, more formally, with the classes of elements that the phenomenon specification function $f(x)$ forbids. A theory EV+C that predicts that "when selecting between gambles, people choose the gamble with the higher expected value and their confidence will increase with the difference between the expected values of the gambles" is more specific than the EV model mentioned above. It describes the behavior more specifically and all findings that falsify EV also falsify EV+C, but not vice versa. For one dependent measure, a theory is more precise than another one if it allows fewer different outcomes (e.g., it predicts judgments in a smaller range or allows choices for fewer alternatives).

Empirical Content. The empirical content of a theory (in the Popperian sense) increases with universality and precision. A theory that is superior in both aspects to another one has a higher empirical content. The scientific advantage of a new theory that is dominated on both aspects would be zero and the theory should be disregarded even prior to any empirical testing (i.e., Poppers' criterion III). On the other hand, a theory that dominates another theory on at least one of the two dimensions has "unique" empirical content which could constitute a scientific advantage if it holds in empirical testing. The gist of the concept, however, is that the more a statement prohibits, the more it says about our world.

One possibility ultimately to annihilate empirical content is to formulate the core premises of a theory in terms of *existential statements*. A statement like "Some decision makers may sometimes use strategy X" can never be falsified because the *modus tollens* is not applicable. In this case, it makes no sense to search for violations of the theory because there is nothing to be violated. If one ignores the issue of logic—with modus tollens unavailable—any instance of applying another strategy Y is rendered theoretically irrelevant because it is not forbidden by the propositions. Numerous decision makers relying on strategy Y would not count. Detecting a single user of strategy X would verify the existential statement that constitutes the theory (see Betsch & Pohl, 2002, for an example).[5]

[5]According to Popper, existential statements would not be considered scientific theories at all because they can never be overcome by empirical testing. Furthermore, please note also the more fundamental point that even a high proportion of choices in line with the predictions of a strategy is not a valid indicator for usage of this strategy (e.g., Hilbig, 2010b; Hilbig & Richter, 2011).

2.3 Internal consistency (I) and satisfying the character of a scientific theory (II)

Consistency. A theory should be consistent by avoiding inconsistent predictions such as that people chose A and "not A" at the same time. Assume, for example, an imaginary decision theory postulating that people make all decisions concerning which of two cities is bigger by applying one of two strategies: Recognition Heuristic (RH; Goldstein & Gigerenzer, 2002) and Inverse-Recognition Heuristic (I-RH). RH predicts that for all choices between options from which one is known and the other one is not, people select the recognized object without considering further information. Assume that the I-RH predicts for the same comparisons that people select the unrecognized object. Note that the theory does allow the application of both heuristics at the same time. The theory would consequently be inconsistent because it predicts choices for and against recognized objects at the same time. The theory could, however, be made consistent by defining exclusive and non-overlapping subsets of tasks for which the two heuristics are applied.

Satisfying the Character of a Scientific Theory. One temptingly easy solution to solve the above problem would be to define the subsets conditional on the observed choices: the RH is defined only for the tasks in which the recognized object is chosen and the I-RH is defined for tasks in which the unrecognized object is chosen. However, this definition is tautological in that the antecedence conditions defined in $\varphi(x)$ are implied by the phenomenon specification function $f(x)$, and therefore the theory does not satisfy the character of a scientific theory. Another solution would be to define the antecedence conditions in $\varphi(x)$ so that people apply RH vs. I-RH if they have learned that the strategy is more successful in the respective environment. This solves the problem but reduces the universality of the theory in that the theory makes only predictions for cases in which previous learning experiences with the strategies exist and in which the success rates differ.

3 General issues

3.1 Simplicity and parameters of a theory

According to Occams' razor, it has been argued that—everything else being equal—simple theories should be preferred over complex ones (Brandstätter, et al., 2006; Gigerenzer & Todd, 1999). However, "simplicity" is a vague concept which can be understood in many different ways (e.g., esthetic, pragmatic; Popper, 1934/2005). Is one model simpler than another because it assumes

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

a particularly simple decision process, such as non-compensatory heuristics, compared to an expected utility model? The definition of empirical content allows specifying the aspect of "simplicity" that is relevant for theory selection and to avoid misunderstandings; for the logical analysis, simplicity essentially reduces to the empirical content and the falsifiability of a theory: according to Popper, a theory is "simpler" than another if a) it is more precise in its predictions, but can be applied at least as broadly as another theory; or b) it can be applied more broadly, although it is at least as precise as the other theory.[6]

Based on this understanding of the concept "simplicity", some common arguments in JDM might require reconsideration. Comparing non-compensatory heuristics and weighted compensatory models, it is often argued that the former are simpler because they have fixed parameters, while the latter have free parameters and can therefore fit a broader set of behavior (e.g., Brandstätter, et al., 2006, p. 428). According to the Popperian concept of "simplicity", this statement is valid when comparing models that are fitted to the data (Roberts & Pashler, 2000) and the difference in flexibility has to be taken into account (Pitt & Myung, 2002; Pitt, Myung, & Zhang, 2002). The argument is, however, irrelevant if parameters are assumed to be fixed or can be estimated from independent measures. Hence, accepting the Popperian definition of simplicity, models are not per se "simpler" than others, but this evaluation depends on the way the parameters are (and can be) determined. For example, an expected utility model is as "simple" as a non-compensatory heuristic if a) fixed parameters for the utility and the probability transformation function are used; or b) the coefficients are measured independently or using a cross-prediction procedure.

The "simplicity" of the theory decreases if multiple outcomes are allowed. This can result from the fact that different sets of parameters are possible or multiple strategies are simultaneously allowed without specifying how to select between them. The phenomenon specification function $f(x)$ would consequently contain many logical OR statements. Hence, a theory assuming that persons are risk-neutral, risk-seeking, or risk-averse and all apply the same utility function (e.g., $U(x) = x^{\alpha}$) with three different sets of parameters (e.g., $\alpha = 1$ for risk-neutral persons; $\alpha = 1.2$ for risk-seekers; and $\alpha = 0.88$ for risk-averse persons) would be less simple than a theory assuming one set of parameters for all ($\alpha = 0.88$, as in Tversky & Kahneman, 1992). However, both theories would level in terms of Popperian simplicity if the risk aversion type of each person is known because it has been measured using a separate scale (e.g., Holt & Laury, 2002).[7]

[6]The simplicity of a theory is therefore always relative and it can be evaluated only in comparison to another competing theory.

3.2 Process models vs. outcome models

Many models in JDM predict choices or judgments only. These outcome models (also called "as-if models", or paramorphic models) predict people's choices or judgments, but they are silent concerning the cognitive operations that are used to reach them. Expected utility theory (Luce, 2000; Luce & Raiffa, 1957) and weighted-linear theories for judgment (e.g., Brunswik, 1955; Karelaia & Hogarth, 2008) are prominent examples. Both are models with a high degree of universality and they are precise concerning their predictions for choices or judgments, respectively (assuming fixed or measurable parameters as mentioned above).

Process models could, however, have a higher precision by making additional predictions on further dependent variables such as time, confidence, information search, and others. A Weighted Additive Strategy (WADD) could, for instance, assume a stepwise deliberate calculation process consisting of elementary information processes (EIPs; Newell & Simon, 1972; Payne, Bettman, & Johnson, 1988) of information retrieval, multiplying, and summing. It makes predictions concerning multiple parameters of information search (Payne, et al., 1988), decision time, and confidence (Glöckner & Betsch, 2008c). Everything else being equal, the empirical content of a theory increases with the number of (non-equivalent) dependent variables, on which it makes falsifiable predictions. Process theories therefore per se yield potentially higher empirical content than outcome theories.[8]

3.3 Distinct predictions and universality

It seems trivial to assume that a new theory should make predictions that are distinct from previous theories. Following Popper's criteria for evaluating theories according to their empirical content, however, this is not always true. If a new model has a higher universality and is equal in precision, it could constitute a scientific advantage even without distinct predictions - just by replacing a set of theories that each predicted parts of the phenomena.

Furthermore, if the universality of a theory is reduced to a set of situations in which more universal theories

[7]Similarly, an adaptive toolbox model assuming two heuristics without specifying which one is used for each situation is less "simple" than an expected utility model with a fixed set of parameters. And it might additionally be inconsistent as described above.

[8]The advantages of exact process specifications using cognitive models for an improved reasoning in psychology have been repeatedly highlighted (for elaborated discussions see Farrell & Lewandowsky, 2010; Glöckner & Witteman, 2010; Lewandowsky, 1993).

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

make the same predictions, it becomes obsolete. Assume a theory predicting that people use a take the best (TTB) heuristic (Gigerenzer & Goldstein, 1996), which states that people decide according to the prediction of the most valid differentiating cue. Further, assume that the theory additionally defines the antecedence conditions that TTB is only used in situations with high costs for information acquisitions (i.e., monetary cost and cognitive costs for retrieving information from memory), and particularly in situations in which these costs are higher than the gain in expected utility of additional cues after retrieving the first cue. Such a theory would make exactly the same choice predictions as an expected utility theory, since expected utility theory would predict utilization of only the first cue under this specific condition either. The expected utility theory, however, is more universal and hence the TTB-theory would be obsolete.[9]

4 Evaluating the empirical content of models in JDM

From the descriptions above, some evaluation of theories in JDM concerning empirical content might have already become apparent. We will summarize and condense them in the following.

4.1 Universal outcome theories

Universal outcome theories are theories that are defined for a wide range of tasks. They have a high level of universality. Universal outcome theories make predictions for choices or judgments but not for other outcome variables. Some may comprise free parameters. The degree of precision is intermediate to high. For example, expected utility theories, cumulative prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), and the transfer-of-attention-exchange model (Birnbaum, 2008) for risky choices belong to this class.

The theory that claims the maximum of universality in this category is a generalized expected utility theory by Gary Becker ("'the economic approach"; Becker, 1976, p. 14) stating: "I am saying that the economic approach provides a valuable unified framework for understanding *all* human behavior [...] all human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets." Precision is, however, relatively low in that neither the set of preferences nor the transformation function for utility is defined or easily measurable.[10] If, by contrast, one suggested that persons have monetary preferences only and assumed that the utility of money follows a specific concave utility curve, precision would be high and the theory would have high empirical content.

4.2 Single heuristic theories

Single heuristic theories are theories that consist of a single heuristic for which a specific application area is defined. The heuristics PH, RH, TTB, mentioned above, and many more, such as the original formulation of the equal weight strategy (Dawes & Corrigan, 1974), elimination by aspect (Tversky, 1972), and affect heuristic (Slovic, Finucane, Peters, & MacGregor, 2002), fall into this class. The application area defined for these heuristics is sometimes rather limited, and therefore universality of these models can be low. The theories can become tautological if their application area is defined by their empirically observable application or by non-measurable variables as described above. Single heuristic theories often make predictions concerning multiple dependent variables such as choices, decision time, confidence, and information search. Hence, their precision is potentially high. To achieve this degree of precision, however, existential statements and ambiguous quantifiers such as "some people use heuristic X" or "people may use heuristic X" must be avoided because they reduce precision and empirical content to zero. Replacing these statements by "under the defined conditions more than 50% of individuals use heuristic X" (and will show the respective choice behavior) is one way to solve the problem. Given that sufficiently precise measurement of this proportion is possible (and only if this is the case, as argued by Hilbig, 2010a), modus tollens can be applied. If one wishes to be more conservative, a much lower limit (e.g., 1%) could be used, which then conveys important information concerning the universality of the theory (i.e., probabilistic universality).

[9] Similarly, if a theory is put forward stating that the recognition heuristic (RH) is used in probabilistic inference tasks if recognition is a valid cue and whenever no other cues are available, then the theory would become obsolete because a more universal WADD theory (in all probabilistic inferences chose the option with the higher weighted sum of cue predictions) would make the same prediction.

[10] Becker also disclaims in this respect by adding: "although I recognize, of course, that much behavior is not yet understood, and that non-economic variables and the techniques and findings from other fields contribute significantly to the understanding of human behavior." Also note that the theory can easily be defined tautologically, in that all violations are reinterpreted as being due to preferences that have not yet been revealed (i.e., post-hoc rationalization; for a comprehensive discussion see Mantzavinos, 2007): if persons, for instance, look up irrelevant information, this indicates that irrelevant information has a utility for them (i.e., an "irrelevant information preference" is defined). The theory has empirical content only if a specified set of preferences (and utility functions) is assumed.

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

4.3 Multiple heuristic theories

These kinds of theories define multiple strategies or heuristics that are applied adaptively. Models in this class are the contingency model (Beach & Mitchell, 1978), the adaptive decision maker approach (Payne, et al., 1988), and the adaptive toolbox (Gigerenzer & Todd, 1999). Their (potential) level of universality is higher than for single heuristics because the theory's scope is not limited to a certain decision domain. However, frameworks that propose an open set of heuristics without defining them have no empirical content because each deviation can be explained by adding a new heuristic (i.e., precision is zero; for a similar argument see also Fiedler, 2010). To have empirical content, a theory has to be defined as a fixed (or at least somehow limited) set of heuristics. The degree of precision of such a fixed-set theory can still be low if many heuristics are included and no clear selection criteria among the heuristics are defined. Examples for selection criteria are trade-offs between costs and benefits of applying strategies, selection contingent on time constraints, or selection based on previous learning experiences. All of these have to be measurable to increase empirical content. Concerning precision, the issues previously described for single heuristic theories apply.

4.4 Universal process theories

Universal process theories describe cognitive process on a fine-grained level. They often rely on one general mechanism that is applied very broadly to explain many seemingly unrelated phenomena. Considering such basic mechanisms often allows replacing multiple theories that explain phenomena on a more abstract level (e.g., Dougherty, Gettys, & Ogden, 1999; Fiedler, 2000; Fiedler, Freytag, & Meiser, 2009; see also Tomlinson, Marewski, & Dougherty, 2011). Universal process theories have a high level of universality because they are applied not only to decision making, but also to phenomena of perception or memory. They allow us making predictions on many dependent variables (e.g., Glöckner & Herbold, 2011) and therefore can be very precise. One of the problems that has to be solved, however, is that universal process models sometimes have many free parameters which are hard to measure and might therefore decrease precision (Marewski, 2010; Roberts & Pashler, 2000).[11] Examples are sampling approaches (Fiedler, 2000; Fiedler, et al., 2009), evidence accumulation models (Busemeyer & Johnson, 2004; Busemeyer & Townsend, 1993; Pleskac & Busemeyer, 2010; Usher & McClelland, 2001), parallel constraint satisfaction models (Betsch & Glöckner, 2010; Glöckner & Betsch, 2008b; Holyoak & Simon, 1999; Simon, Krawczyk, & Holyoak, 2004), cognitive architectures assuming production rules (Anderson, 1987), and multi-trace memory models (Dougherty, et al., 1999; Thomas, Dougherty, Sprenger, & Harbison, 2008).

[11] This problem obviously applies equally to any kind of mathematical models that necessitate extracting parameters from the data to be explained. As discussed by Popper, the empirical content decreases with an increasing number of parameters (which reduces precision) and a decreasing number of observations that can be accounted for (which reduces universality).

5 Conclusions and remedies

We aimed to show that the Popperian criteria for evaluating the empirical content of theories are useful guides towards achieving scientific progress in JDM research. A rigorous analysis of the level of universality and degree of precision of theories increases sensitivity concerning how theories have to be formalized to maximize their empirical content. It also reveals the criteria for them to have empirical content in the first place, and it makes transparent which general weaknesses are present in current models. Many current theories lack precision. As we have shown, precision could easily be improved by adding simple assumptions or limitations to the theories (i.e., replace vague quantifiers; fix a set of parameters; define a method to measure parameters; limit the set of strategies; define precise selection criteria and application conditions).

Unfortunately, researchers could be caught in a social dilemma:[12] it is desirable for the scientific community to have sufficient theory specification to gain scientific progress, but it is individually rational not to have it because it allows for an easier path to defending one's own theory (assuming that retaining one's own theory has high utility for researchers). Also, it could just be convenient not to specify the theory, thus gaining more time to find out which specifications are most appropriate.

This dilemma could be solved on the level of scientific policy. The rigorous standards for empirical research provide the foundation of our science's success and reputation. These standards are documented in catalogues such as the Publication Manual of the American Psychological Association, handed down in academic courses, and serve as guidelines for publication. We suggest that our publication standards *should be extended by principles for the formulation of theories*. Specifically, authors should be required to *maximize empirical content* in their formulation of theories. We suggest three criteria or obligations that may serve as standard evaluation of theoretical papers in a review process.

[12] Social dilemmas are defined by two properties: "(a) each individual receives a higher payoff for a socially defecting choice ... than for a socially cooperative choice, no matter what the other individuals in society do, but (b) all individuals are better off if all cooperate than if all defect." (Dawes, 1980, p. 169)

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

(1) Specification. Authors must explicitly state a finite set of definitions and propositions that together constitute their theory. The gold standard to achieve this aim is formalization. Importantly, this set of premises must be separated from antecedent conditions and auxiliary assumptions (Lakatos, 1970). Moreover, authors should be obliged to state more precisely how and to what extent their theory advances prior attempts of theorizing.

(2) Empirical content. Authors are obliged to formulate the propositions in such a way that the theory as a whole clearly predicts particular states of the world to occur and others not to occur. Wording and formulations that are defective to empirical content (e.g., existential statements) *must not* be included in those premises that constitute the nomological part of the theory.

(3) Critical properties. The authors should be obliged to explain which kind of empirical observations they would consider a fundamental violation of their theory. In other words, the authors themselves should make deductions from their postulated set of premises that could be potentially refuted on empirical grounds.

6 Epilog: Developments and some popular objections

According to our own experience, psychologists sometimes feel uncomfortable with Popper's rigid approach to theory construction and evaluation. As mentioned above, his corroboration approach was criticized. Note, however, that our paper focuses on his ideas of theory construction and evaluation *prior* to empirical testing. These aspects have remained to be the basis for further methodological work, which we will outline next.

6.1 Developments in theory testing

In his influential plea for using strong inference in scientific investigation, Platt (1964) suggests a stepwise procedure for scientific progress. It starts with devising alternative hypotheses, followed by developing and conducting crucial experiment(s) to exclude (branches of) hypotheses. This procedure should be iterated to gradually approach the truth. Scientific advance is achieved by conducting critical experiments that exclude branches of hypotheses. Such a procedure requires as a fundamental prerequisite that theories are formulated in a falsifiable fashion. To put it in Popper's terminology, Platt's approach necessitates that theories have empirical content. Platt suggested that researchers should use the following test questions to evaluate whether a hypothesis and experiment is a step forward: "But Sir, what experiment could disprove your hypothesis?'; or, on hearing a scientific experiment described, 'But Sir, what hypothesis does your experiment disprove?"' (p. 352).

In a similar vein, the issue of falsifiability is central in debates on the persuasiveness of a good model fits and concerning ways to take into account model flexibility. Roberts and Pashler (2000) have argued that model fit per se is no persuasive evidence for a model since one also has to take into account whether "plausible alternative outcomes would have been inconsistent with the theory" (p. 358). This idea is mathematically formalized in the Normalized Maximum Likelihood (NML) approach (e.g., Davis-Stober & Brown, 2011; Myung, Navarro, & Pitt, 2006). In NML the likelihood for observing the data vector given the theory is normalized (i.e., divided) by the average likelihood for observing any possible data vector given the theory.

Obviously, a singular finding in conflict with a theory according to classic hypothesis testing with an alpha level of .05 does not suffice to falsify a theory. Bayesian approaches have been suggested to compare multiple hypotheses on equal footing and allow to easily integrating evidence over multiple studies (e.g., Matthews, 2011; Wagenmakers, 2007). Resulting posterior probabilities provide excellent quantitative measures for the degree of corroboration of a set of theories considering the available data in total.

6.2 Objections

A frequently raised knockout argument is that *Popper's ideas are outdated*. It is true, indeed, that falsificationism as a *methodology* has faced many critiques. The most striking one refers to its difficulties of dealing with probabilistic evidence. Presumably, the Bayesian approach mentioned in the previous section is more viable as a pragmatic methodology of testing theories than Popper's underspecified program of corroboration (e.g., Howson & Urbach, 2006). Note, however, that this critique concerns the pragmatic methodological level only. Pragmatics deal with the problem how to treat a theory vis-à-vis empirical evidence (e.g., whether to keep, alter, or dismiss a theory). On the logical level, we deal with the formulation of a theory prior to testing. The question at this level is: does the theory allow us to derive precise predictions that could be critically tested? The pragmatic level concerns the issue of falsification (e.g., which methods are used when applying modus tollens), whereas the logical level concerns logical falsifiability of a theory. Therefore, the critique raised on the pragmatic level does not apply to the logical level. The point we want to make here is that a theory lacking empirical content is a bad theory because it is immune to reality, regardless of what kind of testing methodology will be applied. This follows purely from

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

logical rules. These rules have not become outdated. If one is to criticize the notion of empirical content one has to dismiss logic: "Of course you are free to dismiss the principles of logic, but then you have the obligation to provide new ones." (Hans Albert, in a personal communication to the second author in 1987)

6.3 Existential statements in theory formulation and theory testing

Findings that show the existence of certain phenomena are crucial for the development of science. According to Platt (1964), this is particularly the case if a theory (or a whole class of theories) exists that explicitly prohibits the occurrence of this phenomenon. Consider, for example, the case of coherence effects, this is, the finding that information is modified during the decision process to achieve a coherent pattern (e.g., DeKay, Patino-Echeverri, & Fischbeck, 2009; Glöckner, Betsch, & Schindler, 2010; Russo, Carlson, Meloy, & Yong, 2008; Simon, Krawczyk, & Holyoak, 2004). Coherence effects violate normative and many descriptive models of decision making which assume informational stability. More generally, much valuable work in JDM has been devoted to demonstrating effects of systematic deviations from rational behavior (i.e., biases, fallacies). Without any doubt, it is extremely important to know that these phenomena exist. They give fresh impetus to theorizing (e.g., Busemeyer, Pothos, Franco, & Trueblood, 2011) and modification of existing models. Furthermore, knowledge of these deviations can also be important for other reasons such as proving important insight for public policy (e.g., Baron & Ritov, 2004; McCaffery & Baron, 2006).

The principle of empirical content is violated, however, if one treats existential statements as theories or if existential statements are used to constitute the core of a theory (Betsch & Pohl, 2002). One might think that this is sometimes the best we can do: to claim and to show that some people show some effects sometimes. Without any doubt, identifying new phenomena is an important first step in any science. However, framing an existential statement concerning an effect as a theory is not the best, but the worst thing we can do. Due to the principles of logic, an existential statement can never be falsified but only be verified, as Popper brilliantly illuminated in his Logic of Scientific Discovery. A single instance of proof suffices to verify an existential statement. Thereafter, a verified existential premise cannot be overcome by means of empirical research. Existential statements are among the reasons responsible for theory accumulation in psychology.

It is, however, important to mention that much work in JDM is concerned with investigating moderator effects for the prevalence of certain types of behavior. Following this approach, showing existence of the behavior can be a first necessary step. The existence statement "there are people that use lexicographic strategies" (which has zero empirical content) can, for example, be a first step towards developing and testing a theory stating that certain factors (e.g., memory retrieval costs; see Bröder & Gaissmaier, 2007; Bröder & Schiffer, 2003b; Glöckner & Hodges, 2011) increase the usage of lexicographic strategies.

6.4 Implications and outlook

Why should we bother about excessive theory accumulation? The community is liberal when it comes to theory (in contrast to empirical methodology!). It allows new researchers to enter the arena quickly by providing demonstrations of phenomena. Obviously, if our main goal is to promote the academic careers of our graduate students, we definitely should not bother with such aloof concepts like empirical content (for an insightful discussion of the social dilemma structure in scientific publication, see also Dawes, 1980, p. 177). We should advise our students to perform like hunters and gatherers: seek a demonstration of an effect; postulate a corresponding "theory" and make the prediction not too strong in order to get published. If the goal, however, is to improve the quality of scientific predictions, we cannot be interesting in maintaining the status quo. We accumulate effects and "theories", but in many cases we have no guidelines how to select a "theory" for making predictions. "Theory" accumulation is not a proof for progress, but rather an indicator for the lack of a shared methodology for theory construction and testing in our science.

"Time will tell", as optimists might conjecture; eventually good theories will survive and the others will be forgotten. According to Poppers' corroboration method, theories should compete against each other. The criterion for completion is to withstand critical tests. The prerequisite for making such tests, however, is that theories have empirical content. Otherwise they cannot take part in the game of competition. They will survive or be forgotten upon arbitrary criteria which are unscientific in nature. More generally, the underlying social dilemma structure has to be broken by introducing and enforcing standards such as the ones described above. To avoid the tragedy of the commons, the standard method for solving dilemma structures has to be applied also for theory construction in JDM: "mutual coercion, mutually agreed upon by the majority of the people affected" (Hardin, 1968).

We would like to close with a positive example. Over half a century ago, Ward Edwards (1954) introduced a powerful theory to psychologists. He imported utility theory together with its set of axioms implying strong pro-

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

hibitions. The influence of this theory was exceptional because it was so strong and because it turned out to be wrong. We learned much about the processes of human choice from critical tests of this theory. The biases, the fallacies, and the violations of principles soon filled our textbooks. They stipulated researchers thinking about how humans *really* make decisions. Our plea is simple. Strive for maximizing empirical content in theorizing. Do not be afraid of failure. Failure breeds scientific advance. This is the gist of Popper's logic of scientific discovery.

References

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 192–210.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74–85.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review, 3*, 439–449.

Becker, G. S. (1976). *The economic approach to human behavior*. Chicago: University of Chicago Press.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry, 21*, 279–294.

Betsch, T., & Pohl, D. (2002). Tversky and Kahneman's availability approach to frequency judgement: A critical analysis. In P. Sedlmeier & T. Betsch (Eds.), *Etc. - Frequency processing and cognition* (pp. 109–119). Oxford: Oxford University Press.

Birnbaum, M. H. (1999). Testing critical properties of decision making on the Internet. *Psychological Science, 10*, 399–407.

Birnbaum, M. H. (2008a). New paradoxes of risky decision making. *Psychological Review, 115*, 463–501.

Birnbaum, M. H. (2008b). New tests of cumulative prospect theory and the priority heuristic: Probability-outcome tradeoff with branch splitting. *Judgment and Decision Making, 3*, 304–316.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without tradeoffs. *Psychological Review, 113*, 409–432.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14*, 895–900.

Bröder, A., & Schiffer, S. (2003a). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making, 16*, 193–213.

Bröder, A., & Schiffer, S. (2003b). Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General, 132*, 277–293.

Brunswik, E. (1955). Representative design and the probability theory in a functional psychology. *Psychological Review, 62*, 193–217.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review, 118*, 193–218.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 133–154). Malden, MA: Blackwell Publishing.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432–459.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Davis-Stober, C. P., & Brown, N. (2011). A shift in strategy or "error"? Strategy classification over multiple stochastic specifications. *Judgment and Decision Making*.

DeKay, M. L., Patino-Echeverri, D., & Fischbeck, P. S. (2009). Distortion of probability and outcome information in risky decisions. *Organizational Behavior and Human Decision Processes, 109*, 79–92.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science, 19*, 329–335.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*, 659–676.

Fiedler, K. (2010). How to study cognitive decision algorithms: The case of the priority heuristic. *Judgment and Decision Making, 5*, 21–32.

Fiedler, K., Freytag, P., & Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review, 116*, 187–206.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart. Evolution and cognition*. New York, NY: Oxford University Press.

https://doi.org/10.1017/S1930297500004149 Published online by Cambridge University Press

Glöckner, A., & Betsch, T. (2008a). Do people make decisions under risk based on ignorance? An empirical test of the Priority Heuristic against Cumulative Prospect Theory. *Organizational Behavior and Human Decision Processes, 107*, 75–95.

Glöckner, A., & Betsch, T. (2008b). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making, 3*, 215–228.

Glöckner, A., & Betsch, T. (2008c). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1055–1075.

Glöckner, A., & Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making, 24*, 71–98.

Glöckner, A., & Witteman, C. L. M. (2010). Beyond dual-process models: A categorization of processes underlying intuitive judgment and decision making. *Thinking & Reasoning, 16*, 1–25.

Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making, 23*, 439–462.

Glöckner, A., & Hodges, S. D. (2011). Parallel constraint satisfaction in memory-based decisions. *Experimental Psychology, 58*, 180–195.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75–90.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135–175.

Hilbig, B. E. (2010a). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making, 5*, 272–284.

Hilbig, B. E. (2010b). Reconsidering "evidence" for fast and frugal heuristics. *Psychonomic Bulletin & Review, 17*, 923–930.

Hilbig, B. E., & Richter, T. (2011). Homo heuristicus outnumbered: Comment on Gigerenzer and Brighton (2009). *Topics in Cognitive Science, 3*, 187–196.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*, 1644–1655.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3–31.

Horwich, P. (1982). *Probability and evidence*. New York: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–292.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge: Cambridge University Press.

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making, 6*.

Lewandowsky, S. (1993). The rewards and hazards of computer-simulations. *Psychological Science, 4*, 236–243.

Luce, R. D. (2000). *Utility of gains and losses: measurement-theoretical and experimental approaches.* Mahwah, NJ: Erlbaum.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.

Mantzavinos, C. (2007). Interpreting the Rules of the Game. In C. Engel & F. Strack (Eds.), *The Impact of Court Procedure on the Psychology of Judicial Decision Making* (pp. 13–30). Baden-Baden: Nomos.

Marewski, J. N. (2010). On the theoretical precision and strategy selection problem of a single-strategy approach: A comment on Glöckner, Betsch, and Schindler (2010). *Journal of Behavioral Decision Making, 23*, 463–465.

Matthews, W. J. (2011). What would judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? *Judgment and Decision Making, 6*.

McCaffery, E. J., & Baron, J. (2006). Thinking about tax. *Psychology, Public Policy, and Law, 12*, 106–135.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167–179.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Oxford, England: Prentice-Hall Print.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (in press). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. B. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347–353.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence. *Psychological Review, 117*, 864–901.

Popper, K. R. (1934/2005). *Logik der Forschung* (11th ed.). Tübingen: Mohr Siebeck.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Russo, J. E., Carlson, K. A., Meloy, M. G., & Yong, K. (2008). The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General, 137*, 456–470.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science, 15*, 331–336.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York, NY: Cambridge University Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*, 155–185.

Tomlinson, T., Marewski, J. N., & Dougherty, M. R. (2011). Four challenges for cognitive research on the recognition heuristic and a call for a research strategy shift. *Judgment and Decision Making, 6*, 89–99.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*, 281–299.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779–804.