Explication, H-D confirmation, and Simplicity

Lukáš Bielik

Department of Logic and Methodology of Sciences

Faculty of Arts, Comenius University

(draft)

Abstract: Explication usually plays the role of the method of language revision. The paper sticks to the Carnapian project of explication and develops some of the formal requirements imposed on the explicatum. However, it departs from Carnap's view when it comes to how to construe the simplicity condition. It is suggested that in many cases the simplicity condition, which in the Carnapian project plays the derived role with respect to the other three conditions – the similarity, exactness, and fruitfulness conditions – is, in fact, substantive for the overall evaluation of explications. Based on a case study of three different explications of the H-D concept of confirmation (provided by Schurz 1991; 1994; Gemes 1993; 1998; and Sprenger 2011), we show that there are cases where competing explicata of a common explicandum satisfy the first three conditions equally well. In those cases, then, the simplicity condition makes the difference. Instead of using Carnap's construal of simplicity, we suggest a Principle of Instrumental Simplicity according to which, ceteris paribus, the simpler the explicatum is, the more likely is its 'survival' in competition with other explicata. Moreover, it is suggested that whereas the similarity, exactness and fruitfulness conditions are, in some sense, formal criteria, the simplicity condition is rather tested empirically.

Keywords: Carnap, Explication, H-D confirmation, Simplicity condition, Tacking paradoxes.

1 Introduction

Explication is usually construed as a method of language revision, both in science and philosophy.[1] To some extent, the present paper sticks to this 'Carnapian' project and develops some of the formal requirements imposed on the explicatum.[2] However, it departs from the Carnapian programme as far

[1] In fact, we can distinguish three different meanings of 'explication': i) the method of explication; ii) the resulting product of that method; and iii) a methodological project of using the method of explication as a systematic means for the construction of semantically precise (formal) languages. In what follows I limit my attention to i) and ii).

[2] Interest in the Carnapian project of explication is currently being revived. However, sceptical voices have also been raised against the adequacy of explication as a proper method for elucidation of philosophical problems in natural language (cf. Strawson's 1963; see also Carnap's reply in Carnap 1963; but currently also by Reck 2012), as well as the critique of the vagueness of criteria imposed on explications (see e.g. Boniolo 2003). For a defence of this method see Maher (2007) and Justus (2012);

as the simplicity condition is concerned. In what follows, I suggest that the simplicity condition, which in the Carnapian project plays the derived role with respect to the other three conditions – the similarity, exactness and fruitfulness conditions – is, in fact, substantive for the overall evaluation of explications in many cases. Based on a case study of different explications of the H-D concept of confirmation (such as provided e.g. by Schurz 1991; 1994; Gemes 1993; 1998; and Sprenger 2011), I suggest the following picture: In philosophical explications it is quite common to find alternative explicata of the same explicandum that fare more-or-less equally well with regard to the similarity, exactness and fruitfulness conditions. In these cases, then, it is obviously the simplicity condition that makes the difference. However, it is far from clear how to evaluate different explicata with respect to the simplicity condition. Indeed, this is a serious problem for the evaluation of competing explications.

I argue that we can come to the simplicity comparison rather indirectly, and in the final part of this paper I propose what I call the *Principle of Instrumental Simplicity*. The principle states that, ceteris paribus, the simpler the explicatum, the more likely its 'survival' in competition with other explicata. Moreover, it is suggested that while the similarity, exactness and fruitfulness conditions are to some extent formal criteria, the simplicity condition is rather tested empirically.

Section 2 presents the essential parts of Carnap's theory of explication and develops the details of his requirements in a systematic way. It also makes clear in what sense we will depart from Carnap's construal of explication. Section 3 presents a case study involving three different explicata (explicates) for a common explicandum, namely the notion of H-D confirmation. In section 4, we compare these explicata applying Carnapian requirements to show that the simplicity condition is an important factor in our evaluation of these explicate.

2 Carnap's explication and beyond

The semantic precision of language is usually a pre-condition for the informativeness of both empirical and non-empirical (i.e. logical or mathematical) sentences (or propositions). The most effective tools for enhancing informativeness are, in our view, definitions and explications. What is the difference between the two? To put it roughly, definitions display or constitute a relation of equivalence or identity between the definiendum (-concept) and definiens (-concepts), while explications play a rather different role. They represent a replacement for or transformation of one concept (or expression) by a non-equivalent concept (expression), provided that the former is not

for a recent analysis of issues related to Carnap's explication project see the collection of papers in Wagner & Beaney (2012).

suitable for fulfilling some previously specified theoretical function and the latter satisfies some set of criteria (discussed below).

Even though Carnap was not the first to employ the method of explication, he certainly did the first systematic exploration of this method.[3] The first introduction of explication as a method of language revision was presented in Carnap (1947), but it was more fully developed in Carnap (1950/1962).

Carnap characterizes explications as conceptual (or linguistic) compounds involving the so-called *explicandum*; that is, the more or less inexact concept (or the term) that is to be replaced by the *explicatum* (or explicate), i.e. the concept satisfying the formal criteria specified below (cf. Carnap 1950/1962, chap. 1). Hence for any (result of the method of) explication there is a relation of replacement of the explicandum with the explicatum.

Carnap's characterization of the explicandum concept is manifold, at least implicitly:

a) First of all, the explicandum is characterized as a kind of *inexact* concept. Note that this is a *semantic qualification*. Basically, the explicandum is a concept that we are usually able to apply to some unproblematic instances (examples). There may, however, be a (fuzzy) class of objects for we cannot unambiguously tell whether we can apply the concept to or not.

b) There is also a *contextual* characterization of the explicandum. It belongs either to a natural language or to a previous stage of the scientific language. It does not mean that the explicandum has to be semantically vague, though Carnap does not pay too much attention to the question of whether the semantic and contextual qualifications are in/dependent. As I will suggest, there may also be different *semantic* kinds of reason to put the explicandum away.

c) Finally, there is a *methodological* aspect of the explicandum, which can be reconstructed from what Carnap says about the explicatum. Since the explicatum should be theoretically fruitful, we can assume that the very motivation for an explication is that the explicandum does not work properly for some specific theoretical aims; the explicandum *is not theoretically fruitful* (given some theoretical desiderata), or it may even generate some serious (conceptual) problems (paradoxes) or impede the development of empirical (or logico-mathematical) hypotheses.

On the other side of the explication relation there is an explicatum (concept) which is supposed to satisfy these four requirements "to a sufficient degree":

[3] Hempel (1952) and Kemeny & Oppenheim (1952) formulated the method in a similar fashion. However, Boniolo (2003) shows that there was a version of the explication method already in Kant's *Critique of Pure Reason*, which Carnap ignored.

i) The explicatum is to be *similar to the explicandum* in such a way that in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.

ii) The characterization of the explicatum, that is, the rules for its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explicatum into a well-connected system of scientific concepts.

iii) The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept).

iv) The explicatum should be as *simple* as possible; this means as simple as the more important requirements (i), (ii) and (iii) permit. (Carnap 1950/1962, 7)

There have been various attempts to account for these requirements in a more developed way by Carnap and other scholars (see e.g. Hanna 1968; Kuipers 2007; Justus 2012; or Dutilh Novaes & Reck 2015). In general, I will stick to this Carnapian characterization of explication, but in the following I propose certain modifications and amendments.

For the sake of convenience, let the terms 'explicandum' and 'explicatum' stand for the (meaningful) expressions rather than concepts as such.

Here we are in agreement with Carnap's view that the relation between the explicandum and the explicatum is that of replacement (even though, he is primarily concerned with concepts). Moreover, it is evident that any explication relation R^E is *irreflexive*, *asymmetric* and *semi-transitive* (since there may be cases where E2 is an explicatum for E1, E3 is an explicatum for E2, but E3 is not an explicatum for E1).

What kind of expression, then, could the explicandum stand for? I propose to count the explicandum as *the meaning specifier*. That is, the explicandum is *a meta-expression that cites an object-term and describes which meaning (concept) it expresses* in either of the following forms:

1) a complete definitional form;

2) a too broad or a too narrow (i.e. non-definitional) form;

3) a semantically trivial (although complete) form;

4) by providing some uncontroversial examples of its otherwise fuzzy (inexact) meaning.

To give an idea of what this list of different forms of explicandum amounts to, let us have a look at the following examples:

a) 'Fish' applies to carps, tunas, salmons, etc. but not to whales or seals.

b) 'Truth' means *truth.*

c) 'Truth' means *a kind of correspondence.*

d) 'A knows that p' means that *A has a true justified belief that p*.

The first example represents a category of expressions with an imprecise (indirect) specification of meaning. This seems to be a paradigm case for the original Carnapian explication, which corresponds to his semantic characterization of the explicandum as an *inexact* concept. However, there may be other *semantic* reasons for replacing one term (the explicandum) by another term, reasons that are almost neglected in the literature on explication. These are suggested by examples b) through d).

Example b) contains a meta-expression that mentions the semantically trivial meaning of the term 'truth'. As such, the expression may be semantically correct, but it may not suit a contextually-driven methodological purpose (for instance to build a philosophically informative theory of truth). Hence, such an expression may be in need of an explication replacement even though it is semantically correct.

Case c) is an instance of an incomplete meaning specifier, in this case one that is too broad. Of course, we could imagine a too narrow example as well; it would still count as a definitionally incomplete meaning specification.

Finally, d) is a case in which the explicandum could be a whole definition; or in other words, a meaning-complete specifier. Indeed, it should be emphasized that it is this case of the explicandum replacement that is so commonly found in papers from within many areas of analytic philosophy.

At the first stage of a conceptual enterprise, a philosopher will often start with some pre-theoretical (or everyday) term (or concept) that does not play an explicitly specified theoretical function, and which she tries to replace with some explicit definition. However, as the proposed definition comes into use, she may discover some formerly unforeseen problems emerging from its use (e.g., paradoxes). In that case, she may need to replace this definition by another one, which would generate neither the problems of the former definition, nor any new ones. Hence, a definition that is suggested as an explicatum at the first stage, may later become a new explicandum that – even though it is semantically correct and complete – may be in need of replacement by some other, theoretically fruitful definition. In section 3, I will present a case study illustrating this point in detail. But before we arrive there, let us have a closer look at the requirements that Carnap imposes on the explicatum.

First, the *similarity condition*. Carnap had been well aware that it is difficult to spell out the similarity relation between the explicatum and the explicandum. Even though he says that the explicatum should preserve most of the instances to which the explicandum (concept) has been applied so far, he himself mentions an example of the replacement of a pre-scientific concept of *fish* by a zoological one that breaks down such a narrowly construed condition. In fact, there has been disagreement between the first generation of advocates of explication over the question of how close the explicatum (concept) should correspond to the explicandum (concept) (see, e.g. Hanna 1968).[4]

Recently, Kuipers (2007) has suggested a progressive way of capturing the similarity condition. He proposes to work with what he calls "conditions of adequacy" and "conditions of inadequacy" that are to be derived from an explicandum (concept). Moreover, conditions of adequacy go hand in hand with typical examples of the explicandum (concept), while conditions of inadequacy reflect typical non-examples of the explicandum (concept). These conditions (along with typical (non-)examples) are supposed to make our intuitions about the explicandum as explicit as possible. So, for instance, in the case of Hempel's notion of confirmation by instances, Nicod's criterion and the Equivalence Condition could play the role of the conditions of adequacy. In that case, any adequate explicatum of confirmation by instances should fulfil these conditions. Indeed, in many cases fulfilling the conditions of adequacy and violating the conditions of inadequacy seem to be an optimal way of how to construe the similarity condition. However, in general, adopting the requirement of typical examples and non-examples of the explicandum (concept) does not work. Consider, again, Carnap's example of two concepts of *fish*: a pre-zoological one and a zoological one. According to the former, whales and dolphins are typical examples of the explicandum concept. However, they haven't been included as instances of the zoological explicatum.

In fact, Kuipers' appeal to conditions of adequacy is very close to the way I will suggest we should construe the similarity condition. Instead of trying to capture the similarity condition by searching for a maximal set of properties that the entities denoted by both the explicandum and explicatum should exemplify, we could specify that condition stipulatively for every particular case of explication in the following way:

Let $F = \{F_1, \ldots, F_n\}$ be the set of properties or relations exemplified by the objects denoted by the explicandum; then select a non-empty subset F^* of properties (relations) from F that, relative to some formerly specified theoretical aims T, are considered necessary in the following sense: If an object does not have the properties (belonging to set) F^*, then such an entity cannot be denoted by the explicatum. In other words, given theoretical aims T, select just those properties of the explicandum objects that are required to be exemplified by objects in order to be explicatum objects. Although these properties should not jointly be considered sufficient for the identification of the explicatum objects, they should help us look for the features that have to be preserved by the explicatum.

For instance, the notion of H-D confirmation (which we discuss extensively in the following section) involves, beside other things, this essential characterization: "hypotheses (theories) are

[4] As Dutilh Novaes & Reck (2015) argue, there is a tension between the similarity condition and the fruitfulness condition, and since the fruitfulness condition is the most important one priority should be afforded to the latter.

confirmed by their successful, deductively derived predictions." Any explication of the H-D confirmation (concept) should accordingly preserve the relation of deductive entailment between a hypothesis and the prediction sentence (the evidence) derived from it. To be sure, this relation is not sufficient for any concept of H-D confirmation but it represents a minimal, necessary condition without which there is no explicatum of the H-D confirmation.

Second, *the exactness condition*. Carnap and other scholars working on this subject are quite clear about what they mean by *exactness*. On the one hand, there is a requirement of syntactic transparency. Every term of the explicatum should be unambiguous with respect to the linguistic category to which it belongs. Moreover, if there are expressions of a natural (or scientific) language that may be represented in a formal *language* by some predicate, then the arity of the predicate should be linguistically transparent. On the other hand, syntactic transparency is supposed to go hand in hand with semantic exactness or clarity. Any term of the explicatum should have been assigned a precise meaning. Even though there are different ways to do that, Carnap recommends the use of (stipulative) definitions as appropriate.[5]

An important qualification should be added here. Carnap prefers the use of those rules (or definitions) in the explicatum that connect the explicatum with some broader network of theoretically useful concepts. In order to satisfy this desideratum, it may be appropriate in certain cases to use a system of definitions rather than a single definition.

For Carnap, the *theoretical fruitfulness* of the explicatum is a straightforward condition. He believes that the more the explicatum contributes to the formulation of empirical or non-empirical hypotheses, the more fruitful it is. Moreover, the explicatum is fruitful also in those cases when it eliminates the (conceptual) problems or puts away the paradoxes generated by the explicandum (provided that it does not generate new serious problems; cf. Kuipers 2007).

What about the *simplicity condition*? Carnap does not say much about it. He says that the explicatum should be as simple as the other requirements permit. Can we spell out this point in a more satisfactory way?

It appears that the simplicity condition may be considered in different ways. At least two senses of simplicity are often distinguished in the literature. First, there is simplicity in the sense of syntactic minimality or elegance (see Baker 2013): the fewer expressions there are in a theory, the more elegant it is. Therefore, in the case of explications simplicity could also be approached in terms of the

[5] In what follows, I will be using the term 'stipulative' or 'codificatory definition' in roughly this way: "Let expression 'E' mean *such-and-so*". In essence, a stipulative definition selects a simple or complex expression (the definiendum) and assigns a meaning to it by using other meaningful expressions of a given language (the definiens). For more on definitions see, for instance, Gupta (2015).

complexity of the definitions used as the explicata. In addition, it may be relevant to take into account also the interconnectedness of the explicatum terms with other expressions of the underlying language and then assess how natural these links appear to be. The more natural the connection is, the simpler it is.

On the other hand, there is an ontological level of simplicity, also known as 'parsimony'. The fewer kinds (or the less number) of entities the theory postulates, the better (cf. Baker 2013). It is, however, unclear what this principle amounts to in the case of explications. Explications are linguistic-conceptual devices of a language revision. Is there any reason to judge them on the basis of entities their concepts postulate? Maybe so.

However, explications are primarily concerned with meanings of expressions (or concepts), and we should accordingly separate the question concerning the simplicity of entities these concepts presuppose from the question concerning the simplicity of concepts (meanings) as such featuring in these explications. But, again, how should we approach the question of the simplicity of concepts? One way would be to follow the syntactic structure of expressions that express them. But this approach would be adequate only given the supposition that the syntactic structure is isomorphic to the semantic structure. Such an assumption is clearly problematic. On the other hand, why should the simplicity of concepts matter here? There may be simple concepts that are theoretically useless and very complex ones that are indispensable for the formulation of certain principles. In that case, does the complexity of concepts employed in explications bear any relation to the adequacy of an explicatum?

Thus it appears that there are several different layers of simplicity. However, when it comes to the degrees of simplicity that could be applied in the context of explication, we lack a reasonable scale or measure. If we are to evaluate a single explication, we can roughly rely on our qualitative estimates. But what about cases where we have two (or more) different explicata of the same explicandum which are qualitatively on a par with respect to the first three conditions but may differ with respect to the simplicity condition? How could we tell a simpler explicatum from a more complex one?

The conjecture I want to propose here, and argue for in the final section, is that other things being equal, the simpler the explicatum is, the more likely it is to survive in competition with the rival explicata. However, as I am going to suggest in section 4, this condition can be tested only indirectly, and, contrary to Carnap's formal (a priori) view of simplicity, rather empirically on a long-term basis; that is, if there are different explicata of the (more or less) same explicandum and those explicata fare equally well relative to the first three criteria, then it is likely that the simpler explicatum (if any) will be adopted in a language of a given theoretical field while the alternative explicata will gradually go out of use. In the final section I return to and argue for this idea.

Let us recap this section briefly: I have suggested that explications are best construed as the relations of replacement that hold between the explicandum, or the meaning specifier, which is unsuitable for some formerly delineated theoretical aims, and the explicatum, or a (system of) stipulative (or codificatory) definition(s) that satisfy the requirements of similarity, exactness, theoretical fruitfulness, and simplicity.

3 The Notion of H-D Confirmation: the Different Explicata

Now let us take a closer look at a notion that has been vigorously discussed over the last few decades, the concept (or rather concepts) of H-D confirmation. H-D confirmation has been defined in slightly different ways in the literature. In this section, I first introduce one simple (pre-theoretic) concept of H-D confirmation and add two more elaborate ones that have often been accused of generating the *tacking paradoxes*: tacking by conjunction and tacking by disjunction. Subsequently, I present three different theoretical explications of the theoretical notion of H-D confirmation based on more or less different assumptions.

Let us begin with the simplest form of the H-D confirmation concept, which is usually described this way:

(H-D1) Hypothesis (theory) T is HD-confirmed by its successful prediction E.

The two slightly more complex concepts are fleshed out in these definitions:

(H-D2) Sentence E HD-confirms T if i) E is contentful ($\nvdash$E); ii) T is consistent; iii) E is true; and iv) T$\vdash$E.
(cf. Hesse 1970; Schurz 1991)

(H-D3) Hypothesis T is HD-confirmed by E relative to B if and only if i) E is true; ii) T$\wedge$B is consistent; iii) T$\wedge$B$\vdash$E; and iv) B$\nvdash$E.
(cf. Hempel 1945/1965; Glymour 1980; or Sprenger 2011 – omitting condition i))

The first two notions depict what is sometimes called 'unconditional H-D confirmation', while the third one describes H-D confirmation relative to some background knowledge B. All three concepts suffer from the so-called *tacking paradoxes*.

The first paradox arises due to the monotonicity of deduction: If E HD-confirms T relative to B (that is, [T$\wedge$B]$\vdash$E), then E confirms T and X relative to B (since, again [(T$\wedge$X)$\wedge$B]$\vdash$E), where X may be any sentence (proposition) unrelated to the content of T (cf., for instance, Hempel 1945/1965; or Glymour 1980, 322). Thus, for example, the evidence of an object *a* being *black* confirms not only the hypothesis "All ravens are black" with respect to the background information of *a* being *raven*, but also the compound hypothesis that "All ravens are black and there's a planet made completely out of cheese in our galaxy." This result clashes with the intuition of many scientists and philosophers, who accordingly think that something's gone wrong with the concept of H-D confirmation.[6]

A similar problem arises on the evidence side. The tacking by disjunction (addition) concerns cases where we want to admit as confirmatory not only the logically stronger evidence but also the weaker evidence. For instance, if our evidence consists of various observations such as *a is black*, *b is black, c is black*, then we take that evidence as HD-confirming the hypothesis that "All ravens are black" with respect to the condition that *a is a raven, b is a raven* and *c is a raven*. If that's the case, then we'd like to consider also a single piece of evidence, such as *a is black*, as HD-confirming the hypothesis "All ravens are black" relative to "a is a raven". So, not only does the logically stronger piece of evidence E HD-confirm some T with respect to B, but so does a logically weaker one, such as $E \vee E^*$ where E^* may be any sentence (proposition) whatsoever. In other words, if E HD-confirms T relative to B (since $[T \wedge B] \vdash E$), then $E \vee E^*$ HD-confirms T relative to B ($[T \wedge B] \vdash E \vee E^*$). Hence, this case seems to allow for a too inclusive concept of hypothetico-deductive confirmation (cf. Hesse 1970; Schurz 1991; or Sprenger 2011).

In the remaining part of this section I will present three different solutions to these paradoxes. All of them basically consist in replacing one of the problematic definitions – (H-D2) or (H-D3) – by a different definition or a system of definitions. Since the authors involved in this fervent discussion themselves describe their own activity as searching for the explicatum of the H-D notion of confirmation, I will interpret their proposed solutions as explications, too.

First, I will look at Gerhard Schurz's relevant deduction approach (cf. Schurz 1991; 1994). Schurz develops his system of relevant deduction to solve not only the problem of the tacking paradoxes, but also the Ross paradox, the Tichý-Miller paradox of verisimilitude and the Prior Paradox of Is-Ought inferences. If his proposal is successful, then it works well also for other problematic concepts. However, I will here restrict my attention to his treatment of the tacking paradoxes.

[6] However, Kuipers (2000, chap. 2) is an interesting exception. He suggests that his classificatory-cum-comparative view of deductive (H-D) confirmation, based on the so-called comparative principles, can accommodate both the irrelevant conjunction objection and the irrelevant disjunction objection. He accepts that if E HD-confirms H, then both: i) E HD-confirms $H \wedge H'$ for any H'; and ii) $E \vee E'$ HD-confirms H for any sentence E'. Nevertheless he argues that the HD-confirmation (or 'd-confirmation') is still localizable, and so the resulting paradoxes are not as absurd as they originally seemed to be.

Instead of refusing formal tools of logic for a philosophical analysis (the strategy preferred by e.g. Strawson 1963) or building a new logical system, Schurz proposes restricting classical deductive inference by certain relevance criteria. He does so by distinguishing the formal *validity* of arguments on the one hand and the *appropriateness* of applied arguments on the other hand (cf. Schurz 1991, 399). Accordingly, he defines relevant deductive arguments via the definitions of premise-relevant and conclusion-relevant deduction. Since Schurz (1991) suggests both more informal (but still precise) as well as completely formal definitions of those notions, we will focus on the more informal ones, which have also been discussed by Gemes in his reaction to Schurz's project (see e.g. Gemes 1993; 1994a; and 1998).

First, the definition of *conclusion-relevant* deduction:

> Assume $\Gamma \vdash A$. Then A is a *relevant conclusion* of Γ if, and only if (henceforth 'iff'), no predicate in A is replaceable on some of its occurrences by any other predicate of the same arity, salva validitate of $\Gamma \vdash A$. Otherwise, A is an *irrelevant* conclusion of Γ. (Schurz 1991, 409)

Here, predicates of 0-arity are simply sentence (or propositional) variables. Now, consider the case in which evidence E confirms hypothesis T relative to B, and hence $T \wedge B \vdash E$. Then the deduction $T \wedge B \vdash E \vee E^*$ has an irrelevant conclusion due to the fact that one of the propositional variables – in this case E* – may be replaced by any other propositional variable (here, on its single occurrence) without affecting the original validity of the argument. The tacking-by-disjunction objection is thus taken care of.

The second step is to offer a definition of *premise-relevant* deduction. Since Gemes's (1998) wording of Schurz's original definition is easier to follow (but equivalent) we will use Gemes's definition (with a slight notational change):

> Assume $\Gamma \vdash A$. Then $\Gamma \vdash A$ is a *premise-relevant* deduction iff (i) there is no single occurrence of a predicate in Γ such that its replacement in Γ by any other predicate of the same arity results in a Γ^* such that $\Gamma^* \vdash A$; and (ii) there are no predicate occurrences in Γ such that they are replaceable by other predicates of the same arity resulting in a Γ^* such that $\Gamma^* \dashv\vdash \Gamma$. (cf. Schurz 1991, 421-422; and Gemes 1998, 4)[7]

[7] The conditions (i) and (ii) are, in fact, independent.

In order to show how this definition solves the problem of irrelevant conjunctions, let us assume that there is a deduction $T \wedge B \vdash E$ such that E HD-confirms T relative to B. Provided that both $T \nvdash E$ and $B \nvdash E$ hold, then $T \wedge B \vdash E$ is the premise-relevant deduction while $[T \wedge B] \wedge X \vdash E$ is not.

Hence, Schurz's (1991) proposal to define the concept of H-D confirmation based on these two preliminary definitions is given in a clear way (here B being a tautology):

(H-D4) Sentence E HD-confirms T iff i) E is contentful ($\nvdash E$); ii) T is consistent; iii) E is true; iv) $T \vdash E$; and v) $T \vdash E$ is a premise-relevant and conclusion-relevant deduction. (cf. Schurz 1991, 422)

This, in fact, is the explicatum for the initial definition (H-D2) that resolves the tacking paradoxes. Even though Schurz (1994) introduces a further modification of this definition due to some problems indicated by Gemes (1994a), the fundamental idea remains the same: only relevant deductions play a role in H-D confirmation.

A different attempt at solving the tacking paradoxes has been made by Ken Gemes (cf. Gemes 1993; 1994a; 1998). His strategy combines two elements: i) denying that every contingent consequence of a theory is part of its content; and ii) that there are so-called natural axiomatizations of theories with respect to which it is possible to define (the notion of) H-D confirmation.

In order to give a positive motivation for i), Gemes develops an account of the content (parts) of theories. His (1994b) gives a syntactic version of the account, and later (in Gemes 1997) he formulates the idea of content-parts within a model-theoretic framework.

To arrive at the first syntactic version of the content-part definition, let α be a variable for well-formed formulas (wffs) of some language L and β a variable for wffs or sets of wffs of L. For any wff σ let us say that σ is stronger than α iff $\sigma \vdash \alpha$ but $\alpha \nvdash \sigma$. Abbreviating 'α is a content part of β' as '$\alpha < \beta$', we can then define the notion of content part as follows:

$\alpha < \beta$ iff α and β are contingent, $\beta \vdash \alpha$, and there is no σ such that $\beta \vdash \sigma$, σ is stronger than α, and every atomic wff that occurs in σ occurs in α. (Gemes 1993)

Thus, for instance, if $\beta = \{(\forall x)[F(x) \rightarrow G(x)], F(a)\}$ and $\alpha = G(a)$, then it holds that α is a content part of β; but for any $\alpha^* = [G(a) \vee H(a)]$, it is not the case that α^* is a content part of β; since $G(a)$ is a consequence of β that is stronger than $[G(a) \vee H(a)]$ and every atomic wff that occurs in $G(a)$ (that is, $G(a)$ itself) occurs also in $[G(a) \vee H(a)]$. Hence, tacking by irrelevant disjunction (or addition) doesn't arise.

However, to satisfactorily solve the problem of irrelevant conjuncts, Gemes needs to employ an additional concept, that of a natural axiomatization of a theory. Put more informally, Gemes requires that "evidence E only confirms those parts of theory T whose content is needed in order to derive E from T" (Gemes 1998, 8). If a theory amounts to a set of wffs closed under the consequence relation, then the definition of a natural axiomatization of theory T goes as follows:

> T′ is a natural axiomatization of T iff (i) T′ is a finite set of wffs such that $T' \equiv T$, (ii) every member of T′ is a content part of T′, and (iii) no content part of any member of T′ is entailed by the set of the remaining members of T′. (Gemes 1993, 483; Gemes 1998, 9)

Hence if theory T consists of a set $\{(\forall x)[F(x)\rightarrow G(x)], F(a)\}$ closed under the consequence relation, then $T' = \{`(\forall x)[F(x)\rightarrow G(x)]', `F(a)'\}$ is a natural axiomatization of T, but $T^* = \{`(\forall x)[F(x)\rightarrow G(x)]', `F(a)', `H(a)'\}$ is not. Now, having the notions of a content part and a natural axiomatization of theories at hand, Gemes provides a different explicatum for the notion of H-D confirmation; in particular, he replaces definition (H-D3) by the following one:

(H-D5) Where N(T) is a natural axiomatization of theory T and A is an axiom of N(T), evidence E HD-confirms axiom A of theory T relative to background evidence B iff E and (non-tautologous) B are content part of $(T\wedge B)$, and there is no natural axiomatization N(T)′ of T such that for some subset S of the axioms of N(T)′, E is a content part of $(S\wedge B)$ and A is not a content part of $(S\wedge B)$. (Gemes 1993, 486; cf. also Gemes 1998, 10)

In other words, only those parts (i.e. axioms) of theory T are confirmed by evidence E that are necessary for the derivation of E relative to some background B, and E and B are content parts of the conjunction of T and B. However, there is one condition that has been left out and which could be added to this definition to complete it, namely the condition that E be a true sentence (or accepted sentence).

Here, for the sake of simplicity, I leave open the question whether the notion of natural axiomatization is clear enough and suitable for the application to any theory. However, Gemes's explication of the H-D confirmation (notion) at least solves what it aims to solve, namely the tacking paradoxes.

Finally, let us focus on a third attempt to replace the problematic notion of H-D confirmation suggested by Jan Sprenger (see Sprenger 2011). Sprenger makes use of the notion of *content part*, as defined by Gemes (1994b; 1997), and combines it with the transposition of $T\vdash E$ – that is, $\neg E\vdash\neg T$.

Moreover, he distinguishes between a theory T and a hypothesis H that may be a content part of theory T, and uses Hempel's notion of the restriction of H to the domain of E in the usual manner (cf. Hempel 1945/1965). If E consists of, say, $G(a)\wedge G(b)$, then the domain of E is simply the set {a, b}. More precisely, if α is any wff, then the domain of a wff α (designated as 'dom(α)') is the set of singular terms that occur in the atomic wffs of a given language and are relevant for α. So, '$H_{|\mathrm{dom}(E)}$' abbreviates 'the restriction of H to the domain of E' (see Sprenger 2011, 504-505).

Now we can introduce the explicatum of the H-D confirmation (notion) proposed by Sprenger:

(H-D6) Evidence E HD-confirms theory T relative to background knowledge B iff

i) E is a content part of $T\wedge B$ (that is $E < [T\wedge B]$ or, in Sprenger's notation: $[T\wedge B]\vdash_{CP} E$);

ii) There are wffs $H_1,\ldots,H_n$ such that $H_1, \ldots, H_n \vdash T$ and for all $i\leq n$, H_i is a content part of T and there is a wff E_i such that: a) E_i is a content part of E; and b) $\neg(H_{i|\mathrm{dom}(E)})\wedge B$ is a content part of $\neg E_i\wedge B$ (that is: $\neg E_i\wedge B \vdash_{CP} \neg(H_{i|\mathrm{dom}(E)})\wedge B$). (cf. Sprenger 2011, 505)

This definition essentially expresses the idea that for any theory T consisting of particular hypotheses H_i (potentially a single one) that all (individually) represent the content part of T, it holds that T is HD-confirmed by evidence E (relative to B) just in case E is a content part of that theory (and background B), and there is a content part E_i of evidence E such that the negation of a domain-restricted hypothesis H_i conjoined with B is a content part of the negation of E_i conjoined with B. Again, as a result we get rid of the tacking paradoxes since conditions i) and ii).a) eliminate tacking by disjunction and condition ii).b) prevents the case of irrelevant conjunctions. However, Sprenger's definition (as well as Gemes's) lacks an important condition for E H-D confirming T: We should include E's being true as an additional condition.

Now we have three different explicata (explicates) for two initial theoretical definitions of H-D confirmation – (H-D2) and (H-D3). However, if the background B is set to be a tautology, then definition (H-D3) reduces to (H-D2). Accordingly, if (H-D2) is supplemented by additional elements of background knowledge B, then it becomes an equivalent of (H-D3). Indeed, from now on I will treat (H-D2) and (H-D3) as equivalent and interpret all three suggestions as explications of the equivalent explicandum.

First, note that what is common to all three explications is that they *replace* one *clear definition* generating some problems by another definition or definitions that avoid these problems and, at the same time, are theoretically fruitful. So in our case of the H-D notion of confirmation we are not replacing a semantically vague concept by a sharper one. On the contrary, the explicandum is a formally defined notion that, unfortunately, generates some problems (viz. tacking paradoxes).

And how do the definitions (H-D2)–(H-D6) fare with respect to the account of explications we provided in section 2? Well, even though they are not explicitly expressed as meaning specifiers (in the case of H-D2 and H-D3) or as stipulative definitions (in the case of H-D4, H-D5, and H-D6), they can be reconstructed as having that form. For instance, the definition (H-D2) may be equivalently transformed to the meaning specifier:

(H-D2)* The expression 'sentence E HD-confirms T' means that *i) E is contentful* ($\nvdash E$)*; ii) T is consistent; iii) E is true; and iv)* $T\vdash E$.

And analogously, the definition (H-D4) suggested as the explicatum for (H-D2)* could be transformed into an explicit stipulative definition:

(H-D4)* Let the expression 'sentence E HD-confirms T' mean that *i) E is contentful* ($\nvdash E$)*; ii) T is consistent; iii) E is true; iv)* $T\vdash E$*; and v)* $T\vdash E$ *is a premise-relevant and conclusion-relevant deduction*.

Hence, on our view, Schurz's explication may be construed as the replacement of the meaning specifier (H-D2)* by a stipulative definition (H-D4)*. And similarly for the explications proposed by Gemes and Sprenger.

Now, if we assume (with the qualifications made above) that all three authors were trying to find an adequate explicatum for an equivalent explicandum, then whose explicatum (if anyone's) is the best one? Before drawing any general conclusions, we should first analyse these three explicata with respect to the criteria discussed in section 2.

4 Evaluation of Explications and the Simplicity condition

All the explicata discussed so far preserve the *similarity* condition with the explicandum (definition) in the minimal sense that was specified in section 2. In particular, any explicatum suggested above preserves the relation of entailment between the set of sentences (or propositions) representing a hypothesis or a theory (and some background), and a sentence expressing the evidence (for it). In other words, the entailment $(T\wedge B)\vdash E$ is a necessary condition for each of the suggested explicates (B being a tautology in Schurz's case). Certainly, there may be other properties (relations) of similarity exemplified by both the explicandum and the explicatum, such as *T being consistent* (in the case of H-D2 and H-D4) or *T and B being consistent* (in the case of H-D3 and H-D5). However, as Carnap himself argued, "close similarity is not required, and considerable differences are permitted"

(Carnap 1950/1962, 7). Hence, the preservation of the deductive relation seems to provide enough information for seeking an adequate explicatum of the H-D confirmation (notion).

As far as the *exactness* condition is considered, Schurz, Gemes and Sprenger use the apparatus of explicit definitions and put them into a coherent system with other notions adopted from logic and related fields. It is hard to tell whether it even makes sense to ask who among them uses the syntactically and semantically sharpest tools. Their suggested explicata are as exact as possible (notwithstanding a minor but repairable caveat, namely that Gemes' and Sprenger's definitions need the truth condition of E to be added). Hence I take them to satisfy this condition (roughly) equally well.

The three explicates are also *theoretically fruitful* in the sense that they get rid of the tacking paradoxes, the elimination of which was the main incentive behind the proposals. Of course, these solutions may generate other problems, but these additional complications seem to be resolvable (see e.g. Gemes 1994a on some problems for Schurz's 1991 account; then Schurz's 1994 reply to Gemes, and Gemes's 1998 follow-up to Schurz). Thus, to evaluate all three explicata with respect to the tacking paradoxes, they have all successfully done their job. Anyway, the question is whether Schurz's approach couldn't be construed as being a bit more general because it aims at solving not only the problems of the H-D confirmation concept but also the other interesting issues associated with the problem of irrelevant disjunctions. That may suggest that Schurz's approach is theoretically more fruitful than the alternative views. Still, it is an open question whether Gemes's concept of content parts wouldn't put his approach on a par with Schurz's definition of conclusion-relevant deduction and thus also solve the other problems associated with irrelevant disjunctions. Again, as far as the tacking paradoxes have been concerned, I take it that all three explicata fare equally well, and are accordingly equally fruitful, at least with respect to the originally considered problems.

Finally, the *simplicity* condition. What linguistic-semantic elements do the three explicata make use of?

We've seen that Schurz's refined definition of H-D confirmation is made possible by adopting two preliminary definitions of premise-relevant and conclusion-relevant deductions. Hence, his strategy amounts to using the language of first-order logic supplied by definitions of relevance; nothing more and nothing less.

The idea behind Gemes's attempt is similar. However, he employs two definitions of different kinds: he first defines a content part of (a set of) well-formed formula(s) (or hypotheses) and then provides the definition of a natural axiomatization of theory. Again, the common framework is the language of first-order logic enriched with these two additional concepts; nothing more and nothing less.

Sprenger's account does not involve the concept of natural axiomatization of a theory. However, it is again embedded in the framework of the language of first-order logic, and makes use of Gemes's definition of a content-part as well as the (Hempel's) concept of the domain restriction. Finally, he uses the transposition of implication (which, of course, is a standard property of implication in first-order logic) and puts all these elements together in one definition.

If we are to compare these three explicata on the basis of simplicity, it is quite difficult to tell which of the proposed explications is the simplest one, or simpler than the other two. There does not seem to be a clear formal a priori criterion available on the basis of which we would be able to tell that the concepts employed by one explication are ontologically more parsimonious than those employed by the other ones. Nor does there seem to be a clear formal criterion for how to judge the syntactic/semantic complexity of these explications.

Anyway, even though we seem to lack a clear-cut a priori criterion for the evaluation of simplicity with respect to competing explications, there is, I suggest, another way to think about simplicity. To put things straightforwardly, it appears that we, in general, *tend* to prefer *simpler* solutions to more complex ones, conditional on *other things being equal*, whether they be theories, explanations or, as in our case here, explications. (Though the-other-things-being-equal condition is crucial here.) This is an empirical assumption which I do not find to be too demanding to accept. Even though I do not provide any independent evidence for this thesis, I believe that besides our folk-psychological intuitions this assumption can also be backed by evolutionary or psychological evidence. What I explicitly claim here is that the connection between simplicity and our behaviour aimed at problem-solving is quite close, in the following sense: When we search for fruitful and effective hypotheses, theories, or explications, and then find some competing candidates that fare equally well according to different criteria except for simplicity, then sooner or later we end up choosing that member of a pool of competing candidates that has continuously been proven to be instrumentally or methodologically simpler than the other candidates. As I've already emphasized, this empirical hypothesis about the connection of simplicity with our problem-solving behaviour substantially depends on the *ceteris paribus* condition.

In order to express this idea about our tendency to prefer simpler solutions to more complex ones in more precise terms I think the following (empirical) statement, which I call 'Principle of instrumental simplicity', will do the job:

Principle of instrumental simplicity

Assume that x and y are two distinct theoretical solutions to some problem z. Then *other things being equal*, $p(\text{Survives}(x, y) \mid \text{Simpler}(x, y)) > p(\text{Survives}(x, y) \mid \text{Simpler}(y, x))$.

The principle says that, in general, if there are, for instance, two distinct theories or hypotheses aimed at solving (explaining, predicting) some common problem (data), then *other things being equal* it is more likely that the simpler solution survives than that the more complex one survives. Hence, in the case of explications, if x and y are any two explicata of a common explicandum z, then – ceteris paribus – the probability that the first survives (over) the second given that the first is simpler than the second is greater than the probability that the first survives (over) the second given that the second is simpler than the first. Hence, given this principle, and the premise that E_1 and E_2 are the alternative explicata of a common explicandum E_d, it follows that the probability of (theoretical) survival of the simpler one is greater than the probability of the (theoretical) survival of the more complex one.

Now, what exactly does this principle tell us with respect to a more general problem of the evaluation of explications? Well, it (only?) says that if one explicatum is simpler than another then we will – within an appropriate period of time – probably stick to the simpler one. If true, is this principle of any help?

It depends on what we expect from such a solution. If we preferred to have a direct and a priori decidable criterion then this principle wouldn't suffice. However, if what is instrumentally simple is somehow indirectly displayed in the choices we make over the course of time, then the simplicity of different explicates (and theories) may be indirectly demonstrated by their survival. That is, if the principle is plausible, then the instrumental simplicity may be tested indirectly and empirically on a long-term basis. That does not mean that the criteria of ontological parsimony and syntactic or semantic complexity play no role in our choices of simpler solutions, however. Our proposal only suggests that we do not need to have a clear-cut formal criterion to select the simpler theories; we can indeed choose simpler theories even if we do not have a crystal-clear theory of our preferences based on simplicity.

If this is correct, which one of the three explicata is the simplest (if any)? The answer I suggest here is straightforward: Let's work with them all and see which one survives our theoretical practices.

Acknowledgement:

I'd like to thank to Fredrik Haraldsen, Miloš Kosterec, Theo Kuipers, Marta Sznajder, Jan-Willem Romeijn, and Marián Zouhar for very helpful comments on an earlier draft of this paper. However, I'm solely responsible for any errors hidden from my view at the moment.

The work on this paper was supported by project APVV No. 0149-12.

References

Baker, A. (2013): Simplicity. In: *The Stanford Encyclopedia of Philosophy*, (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/simplicity/>.

Boniolo, G. (2003): Kant's Explication and Carnap's Explication: The Redde Rationem. *International Philosophical Quarterly* 43 (3), 289-298.

Carnap, R. (1947): *Meaning and Necessity.* University of Chicago Press.

Carnap, R. (1950/1962): *Logical Foundations of Probability.* The University of Chicago Press.

Carnap, R. (1963): Replies and Systematic Expositions. In: Schilpp, P. A. (ed.): *The Philosophy of Rudolf Carnap.* La Salle: Open Court, pp. 859-1013.

Dutilh Novaes, C. & Reck, E. (2015): Carnapian Explications, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization. *Synthese*, Open Access, DOI: 10.1007/s11229-015-0816-z.

Gemes, K. (1993): Hypothetico-Deductivism, Content, and the Natural Axiomatization of Theories. *Philosophy of Science* 60 (3), pp. 477-487.

Gemes, K. (1994a): Schurz on Hypothetico-Deductivism. *Erkenntnis* 41, pp. 171-181.

Gemes, K. (1994b): A New Theory of Content I: Basic Content. *Journal of Philosophical Logic* 23, pp. 595-620.

Gemes, K. (1997): A New Theory of Content II: Model Theory and Some Alternatives. *Journal of Philosophical Logic* 26, pp. 449-476.

Gemes, K. (1998): Hypothetico-Deductivism: The Current State of Play; The Criterion of Significance: Endgame. *Erkenntnis* 49 (1), pp. 1-20.

Glymour, C. (1980): Hypothetico-Deductivism is Hopeless. *Philosophy of Science* 47 (2), pp. 322-325.

Gupta, A. (2015): Definitions. In: *The Stanford Encyclopedia of Philosophy*, (Summer 2015 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/sum2015/entries/definitions/.

Hanna, J. F. (1968): An Explication of Explication. *Philosophy of Science* 35 (1), 28-44.

Hempel, C. G. (1945/1965): Studies in the Logic of Confirmation. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: The Free Press, pp. 3-46.

Hempel, C. G. (1952): *Fundamentals of Concept Formation in Empirical Science. International Encyclopedia of Unified Science*. The University of Chicago Press.

Hesse, M. (1970): Theories and the Transitivity of Confirmation. *Philosophy of Science* 37 (1), pp. 50-63.

Justus, J. (2012): Carnap on Concept Determination: Methodology for Philosophy of Science. *European Journal for Philosophy of Science*, No. 2, pp. 161-179.

Kemeny, J. G. – Oppenheim, P. (1952): Degree of Factual Support. *Philosophy of Science* 19 (4), 307-324.

Kuipers, T. (2007): Introduction. Explication in Philosophy of Science. In: *General Philosophy of Science: Focal Issues.* Theo A. F. Kuipers (ed.), Elsevier, pp. vii-xxiii.

Kuipers, T. (2000): *From Instrumentalism to Constructive Realism*. Springer-Science+Business Media, BV.

Maher, P. (2007): Explication Defended. *Studia Logica* 86 (2), pp. 331-341.

Reck, E. (2012): Carnapian Explication: A Case Study and Critique. In: Wagner, P. – Beaney, M. (eds.): *Carnap's Ideal of Explication and Naturalism.* Palgrave Macmillan, pp. 96-116.

Schurz, G. (1991): Relevant Deduction. *Erkenntnis* 35, 1/3 Special Vol., pp. 391-437.

Schurz, G. (1994): Relevant Deduction and Hypothetico-Deductivism: A Reply to Gemes. *Erkenntnis* 41, pp. 183-188.

Sprenger, J. (2011): Hypothetico-Deductive Confirmation. *Philosophy Compass* 6 (7), pp. 497-508.

Strawson, P. F. (1963): Carnap's Views on Constructed Systems versus Natural Languages in Analytic Philosophy. In: Schilpp, P. A. (ed.): *The Philosophy of Rudolf Carnap.* La Salle: Open Court, pp. 503-518.

Wagner, P. – Beaney, M. (eds.): *Carnap's Ideal of Explication and Naturalism.* Palgrave Macmillan.