Sprachtheorie und germanistische Linguistik (sugl.eu), 28.1 (2018), 3-35
© 2018 by Nodus Publikationen (Münster), ISSN (Online) 2365-8584

Csilla Rákosi

Remarks on the margins of a debate on the role of metaphors on thinking

Abstract

The replications of the experiments in Thibodeau & Boroditsky (2011) by the authors, and Steen and his colleagues brought a spectacular proliferation and development of experimental designs. Despite this, the two series of replications repeatedly lead to conflicting results. This paper intends to resolve this contradictory situation with the help of a novel theoretical framework called the 'experimental complex'. This framework makes it possible to reconstruct the relationship existing among the original experiment as well as its replications, and to evaluate the problem solving process and propose future developments. The high quality of the experiments analysed in this paper does not allow for easy and clear-cut decisions, but it offers a rich source of inspiration, as well as guidance on further prospects.

Keywords: psycholinguistic experiments on metaphor; replication of experiments; problem solving, diverging evidence

1 Introduction

Thibodeau and Boroditsky presented the results of two series of experiments (2011, 2013) in favour of the hypothesis that "exposure to even a single metaphor can induce substantial differences in opinion about how to solve social problems" (Thibodeau & Boroditsky 2011: 1). Steen and his fellow researchers, however, "consistently found no effects of metaphorical frames on policy preference" (Steen et al. 2014: 21) when they conducted follow-up experiments in order to replicate Thibodeau and Boroditsky's results. Thibodeau and Boroditsky, in response, re-analysed their own, as well as Steen et al.'s earlier experiments and conducted new ones, too. They reported results that reinforced their earlier findings (Thibodeau & Boroditsky 2015). As they put it, this controversy has led to a highly productive proliferation of replication attempts, experiment versions and control experiments:

> […] this example highlights the importance of thinking about replication not in terms of individual studies, but in terms of lines of investigation. Often the interpretation of the results of any one experiment depends on many other ancillary pieces of data, norming results, and control conditions reported elsewhere in the same paper or in the same line [of – Cs. R.] work more broadly. […] Arriving at a meaningful culture of replication will require going beyond a focus on direct replication of disconnected single studies, and instead shifting to a theoretically-informed consideration of the broader set of dependencies needed for interpreting any given finding (Thibodeau & Boroditsky 2015: 20f.).

Nevertheless, the debate continued in a further publication (Reijnierse et al. 2015). Both the evaluation of the earlier results and the conclusions drawn from the newer series of experiments diverge to an even greater extent. Thus, it is not clear whether the more elaborated experiments reinforce the results of the earlier set of experiments or must be regarded as overruling them. Therefore, these series of replication attempts seem to be typical examples of the *paradox of replications* (cf. Rákosi 2017a). That is, on the one hand, each replication is a more refined version of the original experiment and their predecessors, *providing more plausible experimental data.* On the other hand, however, instead of leading to converging results, they *trigger cumulative contradictions* among different replications of the original experiment.

Rákosi (2017a, b) put forward a metatheoretical framework that might make it possible to grasp the relationship between original experiments and their non-exact replications, to overcome the above mentioned paradox and evaluate the effectiveness of the problem solving process. The central concept of this framework is the notion of ‘experimental complex’. Experimental complexes consist of chains of closely related experiments which are modified (refined, improved) versions of an original experiment. The aim of these modifications is the elaboration of an experiment that is, at least temporarily, stable (reliable) and generally accepted by the members of the given research field (i.e., it can be regarded as valid, at least temporarily, on the basis of the information available and the criteria considered to be in force). Such experiments are called the *limit of the experimental complex.* Against this background, the following question can be raised:

(Q) How can the cumulative contradictions between Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014) and Reijnierse (2015) be resolved?

The structure of the paper is as follows. Section 2 will present the metatheoretical framework which will be applied to the experimental complex evolving from Thibodeau & Boroditsky (2011). In Section 3, the relationship between the experiments belonging to this experimental complex will be reconstructed. Section 4 will evaluate the problem solving process. Section 5 will offer an answer to (Q) and a brief summary will be presented.

2 Metatheoretical background

2.1 Experimental complexes

In order to grasp the relationship between (non)-exact replications and original experiments, one has to transgress the boundaries of single experiments and identify more complex structures. This motivates the elaboration of the concept of 'experimental complex':[1]

(D1) An *experimental complex* consists of chains of closely related experiments which re-evaluate some part of the original experiment, such as its reliability, experimental design, research hypothesis, applied methods, etc.

Each member of the experimental complex also re-evaluates the plausibility (acceptability) of the results obtained in the original experiment, and makes them more plausible, less plausible, or shows them to be implausible.[2] Such experimental complexes are considerably more complex than single experiments, because they may involve, among other things,

- *modified (improved) versions* of the original experiment,
- *exact replications* of the original experiment or one of its non-exact replications,

[1] For a more detailed elaboration of this concept, see Rákosi (2017a).

[2] The notion of 'plausibility' is the central concept of the p-model of linguistic theorising and argumentation as presented in Kertész & Rákosi (2012, 2014) and applied to diverse fields of linguistic research.

- *control experiments* intended to rule out possible systematic errors in the original experiment or in one of its modifications,
- *counter-experiments* which make the most radical revision to the original experiment by applying a different method (experimental paradigm) to the same stimulus material in order to provide evidence against the research hypothesis at issue,
- a *wider set of perceptual and experimental data*,
- *diverse perspectives* by adherents of different theories,
- *different versions of the research hypothesis*, but also
- *conflicts* emerging from different evaluations of the outcome of the original experiment,
- different kinds of *problems* as well as *solution attempts*;
- *a process of plausible argumentation* that re-evaluates the earlier experimental results in the light of the newer experiments in the experimental complex and tries to resolve the inconsistencies between them.[3]

Experimental complexes have a basically cyclic structure:

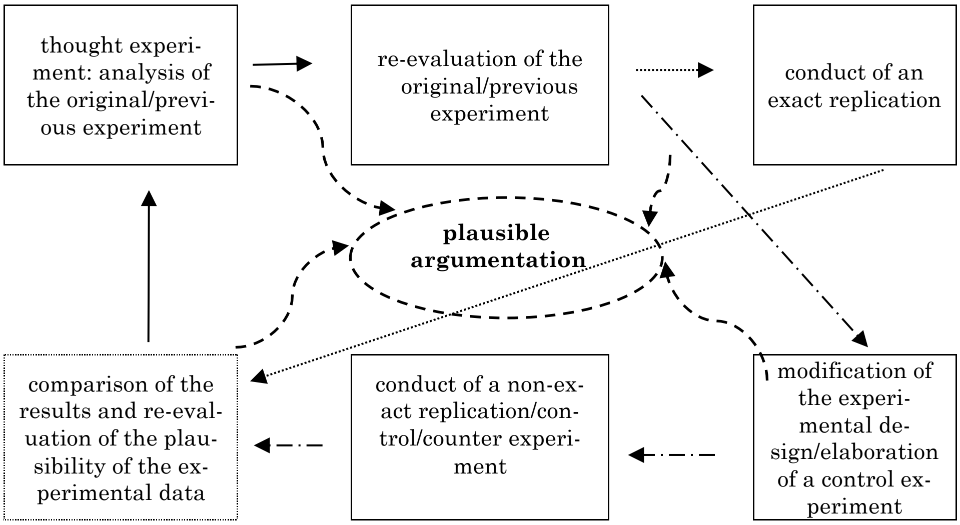
thought experiment: analysis of the original/previous experiment
re-evaluation of the original/previous experiment
conduct of an exact replication
plausible argumentation
comparison of the results and re-evaluation of the plausibility of the experimental data
conduct of a non-exact replication/control/counter experiment
modification of the experimental design/elaboration of a control experiment

Figure 1: The structure of experimental complexes

[3] For a more thorough analysis of the argumentative aspects of psycholinguistic experiments, see Rákosi (2012, 2014), Kertész & Rákosi (2012, Part IV).

The aim of these *cyclic re-evaluations* is the elaboration of an experiment that is, at least temporarily, stable and generally accepted by the members of the given research field:

(D2) An experiment is a *limit* of an experimental complex, if
(a) it evolved from the original experiment through a series of non-exact replications (that is, it results from the gradual modifications of the original experiment),
(b) it has at least one successful exact replication (that is, it is reliable), and
(c) it does not contain unsolved problems, so that the elaboration of further non-exact replications seems to be unmotivated (that is, it can be regarded as valid in the given informational state).

It is always the limit that provides the *most plausible* experimental data within the given experimental complex, because it is free of known problems and is also reliable.[4]

The limit of an experimental complex can be reached, or at least approached, with the help of more and more elaborated non-exact replications of the original experiment. The effectiveness of this process may result from the requirement that every non-exact replication has to solve at least one unsolved problem of the original experiment or the previous members of the chain of experiments. That is, non-exact replications have to be progressive:

(D3) A non-exact replication is *progressive* if it eliminates at least one systematic error or other problem of its predecessors and/or refines the research hypothesis by taking into consideration more relevant factors. If a non-exact replication is not progressive, then it is *stagnating*.

In the light of these concepts, we might be tempted to transform question (Q) raised in Section 1 into (Q'):

[4] Nevertheless, we should not forget that *convergence is mostly only a temporary characteristic of experimental complexes, and it is always relative to a certain informational state and research community*. That is, an experimental complex can arrive at a limit and come to a stop only pro tem and not permanently.

(Q') What is the limit of the experimental complex evolving from the set of experiments presented in Thibodeau & Boroditsky (2011)?

Albeit answering (Q') is a prerequisite for an answer to (Q), it does not necessarily lead to a solution to (Q). In order to solve (Q'), we have to check the progressivity and the problematicness of the non-exact replications of the original experiment in Thibodeau & Boroditsky (2011). The situation, however, may be much more complex. Namely, it is possible that an experimental complex has not reached a limit so far, or it may also have more than one limit at the same time. If there is no limit, the question arises of whether and how a limit could be reached. If there are two limits, then we seem to face an unresolvable contradiction (at least, on the basis of the information at our disposal).

Moreover, there are further difficulties one has to deal with during the problem solving process. For instance, it is not the case that every progressive replication produces more plausible experimental data. The reason for this lies in the circumstance that any modification may not only rule out possible systematic errors but can also lead to the emergence of new ones, which, in addition, may be more serious than the resolved problem, or may even turn out to be fatal. Thus, a progressive replication may solve a problem but also induce a dead end at the same time. Moreover, it is not always the case that non-exact replications provide more and more similar results: quite often the opposite of this happens and the conflicts deepen and multiply. From this, however, it would be premature to conclude that replications were ineffective tools of problem solving. The point is that *effectiveness – in contrast to progressivity – can be judged only in the long run.* This means that we need a methodological tool which makes it possible to *describe and evaluate different strategies of inconsistency resolution.*

2.2 Strategies of inconsistency resolution

The above definition of experimental complexes does not exclude the possibility that within an experimental complex, two chains of non-exact replications (or non-exact replications and counter-experiments) lead to conflicting results. These contradictions cannot be resolved simply by a mechanical comparison of the plausibility value of the last member of the chains of experiments. *It is primarily not the current state of the cyclic process of re-evaluation that is decisive, but the assessment of future prospects.*

Therefore, the first thing to do is to *reconstruct the structure of the experimental complex*, that is, to identify the limit-candidates as well as the chains of non-exact replications, control- and counter-experiments which produce them. The second step consists of *re-evaluating the problem solving process* within the chains of experiments, and then comparing them. If the inconsistencies cannot be resolved on the basis of the information at hand, then the third step should be the *determination of the directions of the continuation of the cyclic process of re-evaluation*.[5] Basically, two strategies are possible in such situations.

The **first strategy** consists of a separate continuation of the chains of experiments by conducting further non-exact replications, counter- or control experiments, comparing the results and coming to a decision. An analogue of this method was called the "*Contrastive Strategy*" in Kertész & Rákosi (2012, 2014). There are three basic situations:

- If the elaboration of further non-exact replications of one of the chains terminates and leads to a limit of the experimental complex in the sense of (D2), while the other chain comes to a dead-end, then the conflict can be resolved in such a way that the limit is kept, while the rival chain is rejected. This means that the elaboration of the first chain of experiments was a case of an effective problem solving process, while the second was an ineffective one.
- If no limit can be achieved by continuing all the chains, then the experimental complex is not capable of reaching a limit and the problem solving process is ineffective.
- It may also occur that both chains of experiments evolving from the same original experiment lead to a limit. In such cases, it would not be reasonable to give up either of them. Thus, this inconsistency has to be (at least temporarily, in the given informational state) tolerated by the application of the second strategy.

A **second strategy** is based on the elaboration and conduct of further experiments involving a refinement of the research hypothesis and experimental design in such a way that all factors found relevant so far are taken into consideration. The analogue of this method was called the "Combinative Strategy" in Kertész & Rákosi (2012, 2014). This method may make it possible to resolve contradictories between exper-

[5] See also Rákosi (2017b) on this.

iments conducted by researchers committed to rival approaches by integrating their results.

In the next sections, we will apply the model delineated in Sections 2.1-2.2 to the experiments in Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014), and Reijnierse et al. (2015). In Section 3, the structure of the experimental complex evolving from Thibodeau & Boroditsky (2011) will be reconstructed and the progressivity of the non-exact replications judged, while Section 4 intends to provide an evaluation of the effectiveness of the problem solving process. That is, Section 3 will focus on the relationship among the experiments. The analyses will try to reveal whether there is at least one problem which is unsolved by an experiment but is solved by its successor, and choose the most elaborated version (limit-candidate) within experiments belonging to the same paper. Section 4 intends to go further and identify remaining unsolved problems which burden the limit-candidate, and evaluate the whole problem solving process.

3 Reconstruction of the structure of the experimental complex

3.1 The original experiment

OE (Thibodeau & Boroditsky 2011, Experiment 1): Participants were presented with one version of the following passage:

> Crime is a **{wild beast preying on/virus infecting}** the city of Addison. The crime rate in the once peaceful city has steadily increased over the past three years. In fact, these days it seems that crime is **{lurking in/plaguing}** every neighborhood. In 2004, 46,177 crimes were reported compared to more than 55,000 reported in 2007. The rise in violent crime is particularly alarming. In 2004, there were 330 murders in the city, in 2007, there were over 500.

Then, they had to answer the open question of what, in their opinion, Addison needs to do to reduce crime. The answers were coded into two categories on the basis of the results of a previous norming study: 1) diagnose/treat/inoculate (that is, they suggested social reforms or revealing the causes of the problems) and 2) capture/enforce/punish (that is, they proposed the use of the police force or the strengthening of the criminal justice system). The researchers found that there was a remarkable difference between the answers of participants who obtained the crime-as-beast metaphorical framing and those who read the crime-as-virus framing: the former preferred enforcement significantly more frequently than the latter group (74% vs. 56%).

3.2 Non-exact replications of the original experiment and control experiments

The experimental complex evolving from OE involves several non-exact replications (NR) and control experiments (CON). See Figure 2 for its structure.

In order to provide a common basis for the comparison of the experiments, we will characterise the non-exact replications with the help of 5 parameters:

1) number of stories;
2) metaphorical content;
3) task;
4) coding system;
5) statistical tools applied.

In the case of the OE, this means the following:

OE (Thibodeau & Boroditsky 2011, Experiment 1):

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 3 metaphorical expressions belonging to one of the two metaphorical frames;
3) Task: suggesting a measure for solving the crime problem;
4) Coding: binary (social reform vs. enforcement), based on the authors' intuitions;
5) Statistical tools: chi-square test, without controlling for other possibly relevant factors such as age, political views, education, etc.

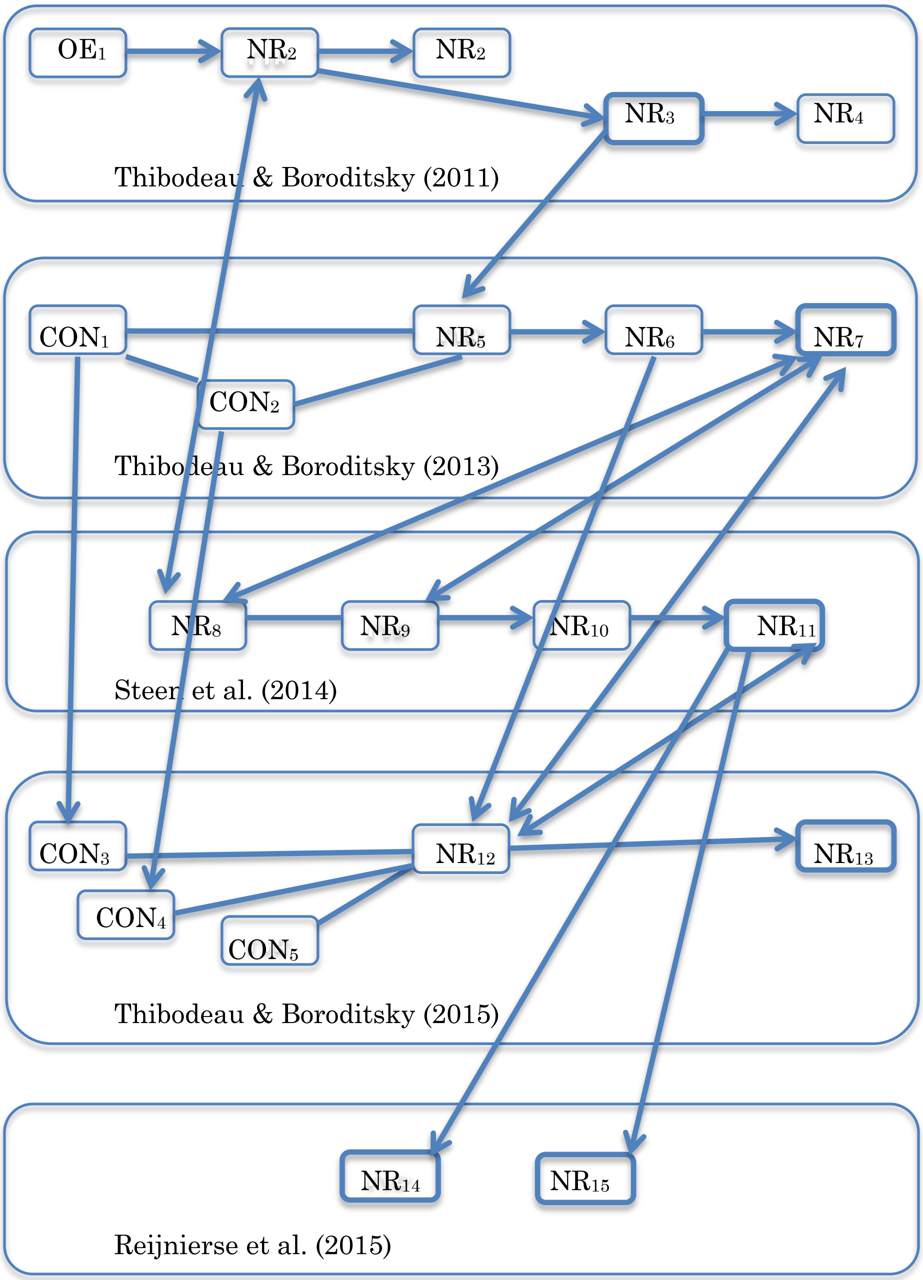

Figure 2: The structure of the experimental complex evolving from Thibodeau & Boroditsky (2011)

3.2.1 Thibodeau & Boroditsky (2011)

The first step of our reconstruction is the description of the experiments along the 5 parameters. We will provide a full characterisation only of the original experiment and the limit-candidate; in all other cases, only modifications carried out to the predecessor of the given experiment will be highlighted.

NR$_1$ (Thibodeau & Boroditsky 2011, Experiment 2), compared to OE:

2) Metaphorical content: *1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions*;
3) Task: suggesting a measure for solving the crime problem + *explaining the role of the police officers*;
4) Coding: binary (social reform vs. enforcement) *with both tasks and averaging the two values*, based on the authors' intuitions.

The first modification is motivated by a case of informational underdetermination insofar as on the basis of the data obtained from OE, one cannot decide whether a metaphorical framing effect can be triggered by metaphorical expressions belonging to the same frame, or a single metaphor would suffice. The second modification is an improvement of the experimental design aiming at disambiguating the relatively frequent answer "increase the police force". The third modification is a consequence of the second change.

NR$_2$ (Thibodeau & Boroditsky 2011, Experiment 3), compared to NR$_1$:

2) Metaphorical content: *0 metaphor*;
3) Task: *providing synonyms for the words 'virus' or 'beast'*, suggesting a measure for crime reduction and explaining the role of police officers.

These changes are motivated by a case of informational underdetermination, too, because OE and NR$_1$ do not make it possible to rule out the possibility that even a single word might suffice to cause a metaphorical framing effect.

NR$_3$ (Thibodeau & Boroditsky 2011, Experiment 4), compared to NR$_1$ – limit-candidate(?):

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames, presented at the beginning of the passage and further ambiguous metaphorical expressions;
3) Task: *selecting 1 crime-related issue from a range of 4 for further investigation*;
4) Coding: binary (social reform vs. enforcement), based on the authors' intuitions;
5) Statistical tools: chi-square test, without controlling for other possibly relevant factors such as age, political views, education.

The only change in comparison to NR$_1$ pertains to the type and focus of the task: instead of the application of an open question about the most important/urgent measure, participants had to choose one issue for further investigation from a 4-member list. This means two things. First, this version may be suitable for reducing informational underdetermination pertaining to the question of whether metaphorical frames can influence people in a similar manner if they have a broader range of possibilities to choose from. Second, asking for possible further investigations may go beyond people's spontaneous decisions and reveal the long term influence of metaphorical frames.

NR$_4$ (Thibodeau & Boroditsky 2011, Experiment 5), compared to NR$_3$:

2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames, presented *at the end* of the passage, and further ambiguous metaphorical expressions.

Moving the metaphor to the end of the passage to be read might help to find out whether metaphors have an effect in isolation or make their impact by guiding and organising knowledge acquisition.

Summary: Every step of the problem solving process is progressive in Thibodeau & Boroditsky (2011), because each non-exact replication provides a solution for at least one problem of its predecessor. This means in most cases, the elimination of informational underdetermination. Nonetheless, it is important to realise that while NR$_1$ is a revised version of OE, which replaces the latter, the relationship between

NR$_1$-NR$_4$ is rather a complementary one. Jointly, they provide evidence for the hypothesis that even a single metaphor can organise the reception of a text in such a way that it influences both direct and long term decisions, while lexical activation of a metaphorical term cannot fulfil this function. Indeed, it is NR$_3$ that seems to be viewed by the authors as a limit-candidate within this chain of experiments. For the reasons for this, see the summary of Subsection 3.2.2.

3.2.2 Thibodeau & Boroditsky (2013)

CON$_1$ (Thibodeau & Boroditsky 2013, Experiment 1), control experiment:

1) Number of stories: 1 story in 1 version (without metaphors);
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames, presented after reading the passage;
3) Task: ordering 1 measure each from a list of 4 to each metaphorical frame;
4) Coding: number of congruent choices (+2, 0, -2);
5) Statistical tools: chi-square test

CON$_2$ (Thibodeau & Boroditsky 2013, norming study), control experiment:

1) Number of stories: –;
2) Metaphorical content: –;
3) Task: rating the 5 measures on the basis of their reform/enforcement-orientedness;
4) Coding: analysis with the help of a 101-point scale, separately for each measure;
5) Statistical tools: t-test

CON$_1$ and CON$_2$ are control experiments. Their function is to check the correctness of the coding system applied in the main experiments.

NR$_5$ (Thibodeau & Boroditsky 2013, Experiment 2), compared to NR$_3$:

3) Task: *selecting the most effective crime-reducing measure from a range of 4*;
5) Statistical tools: chi-square test, logistic regression, *also with control for political views*.

The wording of the task was modified substantially in order to touch upon participants' attitude towards crime reducing measures directly. Several potentially relevant factors were taken into consideration during the statistical analyses.

NR$_6$ (Thibodeau & Boroditsky 2013, Experiment 3), compared to NR$_5$:

3) Task: selecting the most effective crime-reducing measure *from a range of 5.*

The only change to NR$_5$ was the extension of the selection of measures with the 'neighbourhood watches' option, whose evaluation was not unanimous, according to CON$_1$.

NR$_7$ (Thibodeau & Boroditsky 2013, Experiment 4), compared to NR$_6$, limit-candidate:

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions;
3) Task: *ranking 5 crime-reducing measures according to their effectiveness*;
4) Coding: binary (social reforms vs. enforcement), based on CON$_1$ and CON$_2$;
5) Statistical tools: chi-square test, logistic regression, also with control for political views.

There was only a slight difference between this experiment and its predecessor: the technique the participants used to rank the 5 measures was modified.

Summary: From the set of experiments NR$_1$-NR$_4$, only NR$_3$ has been continued in Thibodeau & Boroditsky (2013). Earlier experiments with a negative outcome seem to be regarded by the authors as completed, and the only line of research which was followed was one which entices us with positive results. Thus, the scope of the investigations has been narrowed down. An important improvement, however, is that the assignment of the crime-reducing measures to the metaphorical frames is no longer based on the intuition of the authors but has been checked with the help of two control experiments. The role of potentially relevant further factors was investigated, and the task given to participants was varied, too – more precisely, the formulation of the task was closer to the versions used in OE-NR$_2$. In this case, the relationship between the members of the chain of experiments NR$_5$-NR$_7$ is rather a linear one: each non-exact replication seems to be an improved version of its predecessor. Therefore, this is a progressive series of non-exact replications, too, with NR$_7$ as its limit-candidate.

3.2.3 Steen et al. (2014)

NR$_8$ (Steen et al. 2014, Experiment 1, compared to NR$_7$):

1) Number of stories: *1 story in 3 versions (no-metaphor/'beast'/'virus' frame) in Dutch*;

2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions *vs. 1 metaphor belonging to one of the two metaphorical frames without metaphorical support*;

3) Task: ranking 5 crime-reducing measures according to their effectiveness *before and after reading the passage about crime*;

4) Coding: *+2 (two enforcement-oriented choices in the first two places) / +1 (one enforcement-oriented and one social reform oriented choice / 0 (two social reform-oriented choices), based on the authors' intuitions and/or Thibodeau & Boroditsky (2011, 2013)*;

5) Statistical tools: *ANOVA, logistic regression*, also with control for political views, age, etc.

The authors tried to improve on the earlier versions along all 5 dimensions. They added

– a no-metaphor version, in order to provide a neutral point of reference,
– a version without further metaphorical expressions (a 'without support' version), and
– the task of providing a ranking before reading the stimulus material, too.

They modified the coding system, and the method of the control for further possibly relevant factors, as well as the applied statistical tools. For instance, they took into consideration the first two choices instead of only the first one, and coded them in such a way that they obtained a 3-point scale instead of a purely binary classification.

NR$_9$ (Steen et al. 2014, Experiment 2, compared to NR$_8$)

1) Number of stories: 1 story in 3 versions (no-metaphor/'beast'/'virus' frame) *in English*;

Only the language was changed to NR$_8$. This kind of repetition provides at least as strong a check of the reliability of the results as an exact replication would do.

NR$_{10}$-NR$_{11}$ (Steen et al. 2014, Experiments 3-4, compared to NR$_9$), limit-candidate:

1) Number of stories: 1 story in 3 versions (no-metaphor/'beast'/ 'virus' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions vs. 1 metaphor belonging to one of the two metaphorical frames without metaphorical support;
3) Task: ranking 5 crime-reducing measures according to their effectiveness *only after reading the passage about crime*;
4) Coding: +2 (two enforcement-oriented choices in the first two places) / +1 (one enforcement-oriented and one social reform oriented choice / 0 (two social reform-oriented choices), based on the authors' intuitions and/or Thibodeau & Boroditsky (2011, 2013);
5) Statistical tools: ANOVA, logistic regression, also with control for political views, age, etc.

One of the modifications of NR$_8$-NR$_9$, namely, pre-reading evaluation of the measures, was rejected. The only difference between NR$_{10}$ and NR$_{11}$ was the number of participants: NR$_{11}$ applied a higher number of participants so as to have the power to detect small effects, as well.

Summary: Each non-exact replication is a clearly progressive step in Steen et al. (2014). Interestingly, however, NR$_{10}$ and NR$_{11}$ resolve problems which emerged in the previous members of this chain of experiments. Thus, they provide a kind of self-correction, and can be regarded as the limit-candidates within this chain of non-exact replications. Contrasting a 'without metaphorical support' with a 'with metaphorical support' condition also means a return to NR$_1$, although with a contradictory result.

3.2.4 Thibodeau & Boroditsky (2015)

CON$_3$ (Thibodeau & Boroditsky 2015, norming task 1), control experiment, compared to CON$_1$):

1) Number of stories: 1 story in 1 version (without metaphors);
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames, presented after reading the passage;
3) Task: choosing 1 measure each *from a list of 5* that is most consistent with the given frame;
4) Coding: *analysis separately for each measure*;
5) Statistical tools: *logistic regression*

CON$_4$ (Thibodeau & Boroditsky 2015, norming task 2), control experiment, compared to CON$_2$:

1) Number of stories: –;
2) Metaphorical content: –;
3) Task: rating the 5 measures on the basis of their reform/enforcement-orientedness;
4) Coding: analysis with the help of a 101-point scale, separately for each measure;
5) Statistical tools: t-test

CON$_5$ (Thibodeau & Boroditsky 2015, norming task 3), control experiment:

1) Number of stories: 1 story in 4 versions ('beast', 'virus', 'problem', 'horrific problem');
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames and two non-metaphorical counterparts;
3) Task: ranking the 4 story versions according their severity, metaphoricity, and conventionality on a 101-point scale, and choosing the best one.
4) Coding: analysis separately for each measure;
5) Statistical tools: t-test

The three control experiments contribute to the inter-subjectivity of the results of NR$_7$ and NR$_8$ to a considerable extent.

NR$_{12}$ (Thibodeau & Boroditsky 2015, Experiment 1), compared to NR$_{7}$, limit-candidate:

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions;
3) Task: ranking 5 crime-reducing measures according to their effectiveness;
4) Coding: binary (social reforms vs. enforcement), *based on CON$_1$ and CON$_2$, respectively*, and *also separate analyses for each measure*;
5) Statistical tools: chi-square test, logistic regression, also with control for political views *and other possibly relevant factors such as age, education, etc.*

Due to the two modifications and the application of the three control experiments, this non-exact replication is progressive. Both the separate statistical analysis of the full distribution of the first ranked choices and the deeper analysis of the role of several possibly relevant factors are seminal innovations.

NR$_{13}$ (Thibodeau & Boroditsky 2015, Experiment 2), compared to NR$_{12}$:

3) Task: *choosing between 2 crime-reducing measures.*

The novelty of this member of the experimental complex is that it reduces the impact of the binary coding of the five measures in such a way that only the two most prototypical choices are offered for participants to decide between.

Summary: NR$_{12}$ and NR$_{13}$ add new elements to the experimental designs and rely on carefully elaborated and improved control experiments. At the same time, however, they do not react directly with counter-experiments on the modifications initiated by NR$_{8}$-NR$_{11}$.

3.2.5 Reijnierse et al. (2015)

NR$_{14}$ (Reijnierse et al. 2015, Experiment 1, compared to NR$_{11}$):

1) Number of stories: *1 story in 2 versions (no-metaphor/‘virus’ frame)*;
2) Metaphorical content: *0-1-2-3-4 metaphorical expressions*;
3) Task: *evaluating 4+4 crime-reducing measures according to their effectiveness on a 7-point Likert-scale*;
4) Coding: *average of the enforcement-oriented vs. reform-oriented values*;
5) Statistical tools: *one- and two-way ANOVA*, both with and without control for political affiliation, etc.

NR$_{15}$ (Reijnierse et al. 2015, Experiment 2, compared to NR$_{14}$)

1) Number of stories: *1 story in 2 versions (no-metaphor/’beast’ frame)*;

NR$_{14}$ and NR$_{15}$ could be combined to make one experiment. The experimental design was improved at several points. Both the application of different numbers of metaphorical expressions and the modification of the task are innovative steps. The use of a Likert-scale is a more sensitive and informative tool than ranking the options and the binary coding of the first choice or the first two choices.

Summary: This pair of experiments is highly progressive, not only in comparison to its immediate predecessors but also because it might be suitable for reducing the informational underdetermination mentioned in relation to NR$_{1}$-NR$_{4}$.

4 Evaluation of the problem solving process

4.1 Thibodeau & Boroditsky (2011)

As we have seen in Section 3.2.1, all members of the chain of experiments in Thibodeau & Boroditsky (2011) are progressive non-exact replications, because they provide a solution for at least one problem of their predecessors. Despite this, each of them remains multiply problematic, that is, they are burdened with problems which are associated with all parameters:

1) Number of stories: On the basis of solely one pair of metaphors, it is unfounded to generalise the research hypothesis to all metaphors. Moreover, it might, for example, be the case that it is not the metaphors themselves that make people prefer certain measures, but the fact that newspapers, Internet sources, politicians, etc. could have used a metaphor and associate it with a certain style of argumentation or policies. Such bias can be ruled out only with the help of corpus linguistics and, more importantly, with the involvement of several different topics and metaphors in the experiments.

2) Metaphorical content: As Steel et al. (2014: 4) also remark, the difference between the two versions of the stimulus material used in NE_1, NE_3 and NE_4 does not only lie in the word 'beast'/'virus', because the text contains further idiomatic expressions that can be interpreted differently in the two metaphorical frames. It is also debatable whether the phrases "was in good shape" or "the city's defence systems have weakened" are equally easily and naturally paired with both metaphors.

3) Task: One measure had to be named, one issue had to be chosen, etc. by participants. Therefore, the analysis of their behaviour is reduced to the choice of one measure. A second concern is that the task of selecting a crime-related issue for further investigation in NR_4 and NR_5 approaches peoples' opinion about the efficacy of the possible measures in a considerably more indirect way than earlier and later formulations of this task, leaving room for other interpretations by the participants.

4) Coding: The binary coding (social reforms vs. enforcement) is considerably less sensitive and informative than coding all possible answers separately, and it is based on a categorization which originates solely in the authors' intuitions.

5) Statistical tools: The first concern is that several possibly relevant factors such as age, political views, and education were taken into consideration only in subsequent statistical analyses. Secondly, and more importantly, the effect size, as both Cramér's V and the odds ratio values in Table 1 show, was small.

experiment	**OE**		**NR$_1$**		**NR$_3$**	
condition	**enforce**	**social**	**enforce**	**social**	**enforce**	**social**
beast	1.59 (0.1)	*-2.17* *(0.03)*	1.22 (0.22)	-1.55 (0.12)	1.56 (0.12)	-1.03 (0.3)
virus	-1.61 (0.1)	*2.20* *(0.028)*	-1.1 (0.27)	1.4 (0.16)	-1.46 (0.15)	0.96 (0.34)
Cramér's V	0.18 (p = 0.00013)		0.171 (p = 0.009)		0.192 (p = 0.014)	
odds ratio	2.15		2.05		2.32	
rate of congruent choices	59%		57%		60%	

Table 1: Standard residuals (and significance), effect sizes, rate of congruent choices in Thibodeau & Boroditsky (2011)

Effect size should be viewed as equally important as significance in the interpretation of the results. Therefore, it is highly questionable whether it is justifiable to maintain the (universal) hypothesis that metaphors influence people's opinion if this influence is very limited in its magnitude and/or extent. Thirdly, if we break down the significant chi-square tests with standardized residuals, then we have to confront a further issue. Namely, the standardised residuals in the congruent cells (beast and enforce, virus and social) should be positive and significant, indicating that these cells contribute significantly to the chi-square value (and complementary, the incongruent cells should have significant minus values). Except for the social type answers in the original experiment, the values reveal that the response frequencies do not differ significantly from their expected values in the individual cells. This finding suggests that the differences are in the right direction, but that they are not strong enough. Moreover, since it is only OE that produced a result which is, at least in the case of one condition, in perfect harmony with the predictions, the authors' decision to continue solely with NR$_3$ in their later publications can be questioned. More specifically, the deeper statistical analysis of the perceptual data indicates that raising open questions as a task should not be abandoned, and the application of several metaphorical expressions belonging to the given frame should be investigated again.

A further interesting point is, as Table 2 shows, that there were changes in the proportions of the answers of the types 'enforce' and 'social'.

experiment	**OE**	**NR$_1$**	**NR$_2$**	**NR$_3$**	**NR$_4$**
enforce	65%	62%	64%	*30%*	*33%*
social	35%	38%	36%	*70%*	*67%*

Table 2: Count proportions in Thibodeau & Boroditsky (2011)

According to the authors' explanation, this shift is due to the application of a closed list of possibilities instead of open questions.[6] On the basis of later developments (see Section 4.2), however, this explanation seems to be insufficient.

From these considerations it follows that none of the experiments in Thibodeau & Boroditsky (2011) can be regarded as the limit of this experimental complex, because they are not free of problems.

4.2 Thibodeau & Boroditsky (2013)

1) Number of stories: No improvement was made in comparison to Thibodeau & Boroditsky (2011).
2) Metaphorical content: No improvement was made in comparison to Thibodeau & Boroditsky (2011).
3) Task: The progressivity of this chain of experiments is to a considerable extent due to the more refined formulation of the tasks.
4) Coding: CON$_1$, that is, Experiment 1 in Thibodeau & Boroditsky (2013) is a control experiment, intended to test the hypothesis that people "can extract the metaphorical entailments of the two metaphors when they have an opportunity to compare the two frames explicitly" (Thibodeau & Boroditsky 2013: 4). According to the authors, from this "we should expect people to associate enforcement-oriented programs with the beast metaphor and reform-oriented programs with the virus metaphor." This means that this experiment intends to check the correctness of the stimulus material and coding system of Experiments 2-

[6] Cf. "Laying out four possible approaches to crime shifted the overall likelihood that people wanted to pursue social reform. It seems that explicitly seeing the space of possible responses makes people more likely to attempt reducing crime through reform than enforcement. However, we still found that peoples' responses were influenced by the frame that they read." (Thibodeau & Boroditsky 2011: 8)

4. It is questionable, however, that this aim has been achieved. The decisive point is the statistical evaluation of the perceptual data. Namely, the authors conducted a chi-square test that showed that significantly more participants gave two congruent responses and significantly fewer participants provided two incongruent responses than expected by chance. If, however, we take into consideration that not all measures must have been assigned to the two metaphorical frames, but that participants had to choose only 1 measure each for both frames, then it seems to be more appropriate to accept only responses with 2 congruent solutions. To put it differently, it seems to be reasonable to collapse the answers into two categories (acceptable, i.e., 2 congruent answers vs. non-acceptable with 1 or 0 congruent answer), and require that at least 66% of participants gave an acceptable answer. This was, however, not the case. A binomial test indicated that the proportion of acceptable answers of 57% was significantly lower than expected, $p = 0.003$ (1-sided).

CON_2 is a control experiment, too. Here, the relatively low number of participants and the high standard deviations can be regarded as weak points. From this point of view, the evaluation of the "neighbourhood watches" option is pivotal, because it was only slightly above the midpoint of the scale. This finding and the large standard deviation indicate that the judgement of this option was rather equivocal. The authors' decision to dichotomize the results and force this option into the enforcement-oriented category exerted a decisive influence on the interpretation of the experimental data obtained in CON_1 and NR_5-NR_7, too. Moreover, the "neighbourhood watches" option was not included in CON_1; thus, its assignment to the 'enforcement' category is even more questionable.

To sum up, a detailed re-analysis of the data for each option separately with both metaphors in CON_1 could be highly beneficial (see CON_3 on this). A further possibility could be the application of the numerical values obtained in CON_2 instead of the binary coding in the statistical evaluation of the results of CON_1 and the further experiments.

5) Statistical tools: The extension of the statistical analyses to the investigation of the impact of the political affiliation of participants in the main analyses is an important step. The problems mentioned in relation to OE-NR_4 in Section 4.1, however, remain unsolved. What is more, NR_6 produces only marginally significant results ($\chi^2 = 3.761$, $p = 0.058$). See Tables 3 and 4.

experiment	NR$_6$		NR$_7$	
condition	enforce	social	enforce	social
beast	0.72 (0.47)	-1.2 (0.23)	1.1 (0.27)	-0.9 (0.37)
virus	-0.69 (0.49)	1.15 (0.25)	-1.17 (0.24)	0.93 (0.35)
Cramér's V	0.148 (p= 0.058)		0.111 (p = 0.049)	
odds ratio	1.99		1.58	
rate of congruent choices	56%		55%	

Table 3: Standard residuals (and significance) and effect sizes in Thibodeau & Boroditsky (2013)

experiment	NR$_5$	NR$_6$	NR$_7$
enforce	19%	76%	39%
social	81%	24%	61%

Table 4: Count proportions in Thibodeau & Boroditsky (2013)

If we compare the data in Table 4 with those in Table 2, it becomes clear that the authors' explanation for the finding that the rate of enforcement-oriented and social reform-oriented answers changes drastically among experiments cannot be sustained. Thibodeau & Boroditsky (2013: 5f.) identified two possible causes: the number of the measures from which participants could chose (2+2 vs. 3+2), and their political affiliation. These factors, however, do not seem to provide a satisfactory answer, for example, for the differences between NR$_6$ and NR$_7$.

A further issue needing a closer look is the choice of the statistical tools. First, the authors used logistic regression in their analyses. Since all data are categorical in NR$_5$-NR$_7$, chi-square test and loglinear analysis could be better choices, or, at least, it seems to be reasonable to use them as control analyses. Second, there are further alternatives which seem to be worth investigating. They are based on the abandonment of the questionable binary coding of the measures into reform- and enforcement options. This, as we have already mentioned, could happen in two ways, pointing in opposite directions.

a) *Analysing the relationship between metaphorical frames and the five response options directly*. With NR$_6$, a chi-square test indicated no effect of the frames: χ^2 (4) = 6.94, p = 0.141. Similarly, a chi-square test indicated no effect of the frames in the case of NR_7, either: χ^2 (4) = 5.876, p = 0.21. Tables 5 and 6 make it possible to reveal the enormous differences between the percentages and standardized residuals of the measures in NR_6 and NR_7, respectively:

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	2.4% 0.3	17.1% -1.2	46.3% 1.1	8.5% -0.8	25.6% 0.4
virus	3.4% 0.2	29.2% 1.1	31.5% -1.1	14.6% 0.8	21.3% -0.4
total	2.9%	23.4%	38.6%	11.7%	23.4%

Table 5: Count proportions in Thibodeau & Boroditsky (2013, Experiment 3)

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	42.1% -0.9	14.2% -0.3	14.2% 0.9	17.5% 1.1	12% -0.1
virus	51.2% 0.9	15.9% 0.3	9.4% -0.9	11.2% -1.1	12.4% 0.1
total	46.5%	15%	11.9%	14.4%	12.2%

Table 6: Count proportions in Thibodeau & Boroditsky (2013, Experiment 4)

b) Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON_2. Instead of dichotomising the responses, we might try to apply a finer scale with different values for each response. That is, the application of the ratings collected in CON_2 might represent the enforcement- vs. reform-orientedness nature of the measures in a better way. The analyses show that there is an effect of the frames – although the results are more convincing with NR_7. In the case of NR_6, a Mann-Whitney U test showed that the beast frame was significantly more enforcement-oriented (mean rank = 93.54) than the virus frame (mean rank = 79.06), U = 3031, p = 0.046 (two-sided). The mean enforcement value was 66.68 for the beast frame and 59.06 for the virus frame. A Kruskal-Wallis test reinforced the result that the enforcement-orientedness was significantly affected by the choice of the metaphorical frame; H(1) = 3.989, p = 0.046 (two-sided). As for NR_7, a Mann-Whitney U test showed that the beast frame was significantly more enforcement-oriented (mean rank = 187.83) than the virus frame (mean rank = 165.34), U = 1357.5, p = 0.028 (two-sided). The mean enforcement value was

44.5 for the beast frame and 37.05 for the virus frame. A Kruskal-Wallis test produced a similar result; H(1) = 4.813, p = 0.028 (two-sided).

These analyses should have produced similar results in the sense that they should be in harmony (that is, both should be either significant or non-significant). On the basis of the above considerations, none of these non-exact replications can be regarded as a limit of this experimental complex, either.

4.3 Steen et al. (2014)

1) Number of stories: The most problematic point of OE-NR_7, namely, the use of only one pair of metaphors in the stimulus materials, questions the generality of the results of NR_8-NR_{11}, too. On the basis of only one pair of metaphors, one can draw neither positive nor negative conclusions about the research hypothesis.
3) Task: NR_8 and NR_9 cannot be regarded as data sources providing plausible experimental data, because raising the same questions before and after the presentation of the stimulus material could have influenced participants' decisions insofar that they might have stuck with their first decision. This could have diminished or masked the influence of the stimuli.
4) Coding: The assignment of the 5 measures to the two metaphors was not controlled for. Thus, the coding system is less reliable than it was in Thibodeau & Boroditsky (2013), because it is based either on the researchers' intuitions or was simply taken from earlier experiments.
5) Statistical tools: The authors applied ANOVA to Likert-type items, which is controversial. Thus, it seems to be advisable to repeat the statistical analyses with the help of tests allowing the dependent variable to be ordinal. Such tests are, for instance, Ordinal Logistic Regression or Optimal Scaling (Categorial Regression). Nevertheless, these tests reinforce the results of the authors: no metaphorical support can be identified. The same result was found with analyses narrowed down to the first chosen options.

We might also try the alternative analyses conducted with NR_6 and NR_7 in the previous subsection in this case, too.

a) *Analysing the relationship between metaphorical frames and the five response options directly*. With NR_{11}, a three-way loglinear analysis resulted in a model with a likelihood ratio of χ^2 (0) = 0. It indicated no three-way interaction between response, metaphorical frame and metaphorical support: χ^2 (8) = 8.228, p = 0.412, and no two-way interactions were found, either: χ^2 (14) = 15.072, p = 0.373. As Table 7 shows, the data produce a different pattern from the data obtained in earlier experiments; moreover, in several cases, their direction (sign) and/or their value is in sharp conflict with the predictions:

		measure				
		economy	**education**	**patrols**	**prison**	**watch**
neutral	**no support**	22.8% 0.1	21% -0.7	18% -0.5	4.8% 0.3	33.5% 0.8
	support	23.9% -0.3	20.9% 0.3	22.1% 0.3	1.2% -1.2	31.9% 0.2
beast	**no support**	*25.6%* *0.9*	20.6% -0.8	*21.1%* *0.4*	*3.9%* *-0.3*	28.9% -0.3
	support	*29.8%* *1.2*	18.5% -0.4	*21.3%* *0.1*	5.1% 1.9	*25.3%* *-1.4*
virus	**no support**	*18.6%* *-1.0*	29.3% 1.5	*19.8%* *0.0*	4.2% -0.1	28.1% -0.5
	support	*22%* *-0.9*	*20.2%* *0.1*	*19.7%* *-0.4*	1.7% -0.8	*36.4%* *1.2*

Table 7: Count proportions in Steen et al. (2014, Experiment 4)

Nonetheless, if we reduce our analyses to the 'with metaphorical support' version and focus solely on the comparison of the 'beast' and 'virus' frames, the results are marginally significant: χ^2 (4) = 8.684, p = 0.069. It is questionable, however, whether this result provides any support to the research hypothesis, because there should be differences between the 'virus' frame and the 'neutral' condition, as well as between the 'beast' frame and the 'neutral' condition, and these differences should point in opposite directions. This was, however, not the case.

b) Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON$_4$. When the first two choices were taken into consideration, a multiple regression found no effect of the frames or the presence of metaphorical support on enforcement-orientedness, F(2) = 0.525, p = 0.592, R^2 = 0.01. On a second attempt, only the first choice of participants was investigated. This analysis led to the same results, F(2) = 0.13, p = 0.988, R^2 = 0.00002. Similarly negative results were produced by an analysis which used a non-parametric test, omitted the variable 'metaphorical support', and took into consideration only the data of participants who received the text with metaphorical support.

Summing up our analyses, we may conclude that no member of this chain of experiments can be regarded as the limit of the experimental complex, because each of them remained multiply problematic.

4.4 Thibodeau & Boroditsky (2015)

1) Number of stories: The same pair of metaphors was used in one story. Thus, there is no progress in this case, either.
2) Metaphorical content: Since no no-metaphor version was used and the number of metaphorical expressions was not varied, in this respect, this rather counts as a relapse.
5) Statistical tools: Since there are no significant differences between the two conditions in respect to participants' age, political affiliation and gender in the two experiments, it is possible to check the relationship between frames and responses directly. A chi-square test showed no significant effect of the frames in NR_{12}, χ^2 (1) = 1.432, p = 0.241. NR_{13} produced marginally significant results: χ^2 (1) = 3.322, p = 0.075. Table 8 helps us to compare the data with the outcomes of OE, NR_1, NR_3, NR_6 and NR_7:

experiment	**NR_{12}**		**NR_{13}**	
condition	**en-force**	**social**	**pa-trols**	**educa-tion**
beast	0.7	-0.5	0.8	-0.9
virus	-0.6	0.5	-0.9	1.0
Cramér's V	0.052 (p = 0.241)		0.080 (p = 0.068)	
odds ratio	1.24		1.38	
rate of congruent choices	53.4%		54.7%	

Table 8: Standard residuals and effect sizes in Thibodeau & Boroditsky (2015)

Alternative analyses:

a) *Analysing the relationship between metaphorical frames and the five response options directly.* With NR_{12}, a chi-square test indicated a significant effect of frames on the choice of the measures: χ^2 (4) = 13.748, p = 0.008. As Table 9 shows, however, the only category with significant differences was the response option 'watch'.

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	19.9%	24%	33.3%	6.1%	*16.7%*
	0.9	0.4	1.2	-0.7	*-2.1*
virus	15.2%	21.6%	25.9%	8.5%	*28.7%*
	-0.9	-0.4	-1.1	0.7	*2.0*
total	17.4%	22.7%	29.4%	7.4%	23.1%

Table 9: Count proportions in Thibodeau & Boroditsky (2015, Experiment 1)

b) *Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON$_4$.* An analysis making use of the ratings collected in CON$_4$ showed no effect of the frames. According to a Mann-Whitney U test, there is no significant difference between the beast frame (mean rank = 259.79) and the virus frame (mean rank = 268.61), U = 35844.5, p = 0.496 (two-sided). The mean enforcement value was 47.05 for the beast frame and 46.07 for the virus frame. A Kruskal-Wallis test produced the same results; H(1) = 0.464, p = 0.496 (two-sided).

As for NR$_{13}$, a chi-square test showed only a marginally significant effect of the metaphorical frame: χ^2 (1) = 3.322, p = 0.075. A loglinear analysis indicated a clearly significant interaction between political affiliation and response: χ^2 (2) = 13.203, p = 0.001, a marginal interaction between response and frame: χ^2 (1) = 3.235, p = 0.072, and no three-way interaction among these factors: χ^2 (2) = 0.24, p = 0.887.

This means that there is no unproblematic non-exact replication in Thibodeau & Boroditsky (2015), either.

4.5 Reijnierse et al. (2015)

1) **Number of stories**: Similarly to NR$_5$-NR$_{13}$, there was only one story, although in two slightly different versions (crime described as a long-term problem vs. a short-term problem) in NR$_{14}$ and NR$_{15}$, respectively. Thus, only two sets of metaphors were used again.

3) **Task**: Participants had to evaluate 8 crime-reducing measures according to their effectiveness on a 7-point Likert-scale. This step could produce more sensitive measures and lead to more valuable experimental data than was the case in the previous experiments.

The authors, however, presented the measures not in a random order for each participant but showed the frame-consistent 4 measures first and the other 4 measures second. This might lead to a bias which seriously calls into question the validity of the results, because the skewing effect of the presentation order could not be eliminated.

4) **Coding**: Besides the basically binary coding (average of the enforcement-oriented vs. reform-oriented values), a comparison of the values separately for each measure could also be informative.
5) **Statistical tools**: Similarly to NR_{10}-NR_{11} in Steen et al. (2014), the application of ANOVA to Likert-scale items is debatable.

Despite the innovative character of the experimental design in Reijnierse et al. (2015), both experiments remained problematic.

5 The answer to (Q)

Now, we can address the question we raised in Section 1:

(Q) How can the cumulative contradictions between Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014) and Reijnierse (2015) be resolved?

In Section 2.1, we transformed this question into (Q'):

(Q') What is the limit of the experimental complex evolving from the set of experiments presented in Thibodeau & Boroditsky (2011)?

On the basis of the re-evaluation of the non-exact replications in Subsections 4.1-4.5, the following answer to (Q') presents itself:

(A') The experimental complex evolving from the set of experiments presented in Thibodeau & Boroditsky (2011) has not reached a limit in Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014) or Reijnierse et al. (2015).

Clearly, (A') is not sufficient to provide a fully-fledged answer to (Q). Therefore, the task to be undertaken is to determine the directions of the continuation of the cyclic process of re-evaluation:

(A) Since no non-exact replication was superior in all respects to its rivals, the application of the Combinative Strategy seems to be more appropriate. The process of re-evaluation should be continued by collecting and systematising all relevant and workable elements of the experiments conducted within this experimental complex, and new limit-candidates could be elaborated along the following lines:

- **Number of stories**: The most important change could be to increase the number of stories.[7] One cannot draw general conclusions on the basis of only one topic and two metaphorical frames. For example, it might be the case that politicians and the press made use of a metaphorical frame, which, as a result, can be associated with a certain political standpoint. Such influences can be circumvented only if the experiments make use of several different stories and a great variety of metaphors.
- **Metaphorical content**: It remained an open question as to whether only one metaphorical expression might influence participants' decisions or a series of expressions belonging to the same metaphorical frame are needed. A no-metaphor control seems to be warranted. In addition, it could be also examined whether novel and conventional metaphors have the same effect.
- **Task**: There were plenty of more or less different versions of the task. The most sensitive and informative seems to be the application of Likert-scales, but the use of an open question (such as suggesting a measure for solving the crime problem in OE) in a first set of experiments could be beneficial, too, in order to provide a comprehensive and varied set of options to participants.
- **Coding system**: Both the basically binary coding (average of the enforcement-oriented vs. reform-oriented values), and an analysis of the values separately for each measure would be beneficial, providing information from different points of view.
- **Applied statistical tools**: The applied statistical tools should be chosen in such a way that their applicability to

[7] Thibodeau (2016) made important steps into this direction.

> diverse variable types is taken into consideration. The impact of possible relevant factors such as political affiliation, age, education, etc. should be controlled for properly. A further important task evolves from the small effect size – provided that the further non-exact replications will also find small effect sizes. Namely, one should attempt to narrow down the investigations to subgroups so that the factors characterising people who are responsive to the influence of metaphors can be identified.

The most striking feature of our answer to (Q) is that it does not decide between the two standpoints and tell us who was right. Secondly, our analyses showed that the later experiments conducted by the same authors do not provide converging evidence for the authors' standpoint. That is, we cannot interpret the situation in such a way that the plausibility values of the experimental data would add up to continuously higher values. Instead, the relationship among these experiments is determined by the operation of *recurrent re-evaluation*: the newer, more refined and revised versions replace the earlier ones. Thirdly, the high quality of the experiments analysed in this paper does not allow an easy and clear-cut decision, but is a rich source of inspiration and guidance for further progress. That is, the experiments conducted by both parties motivate and provide starting points for the continuation of the re-evaluation process in order to elaborate a limit for this experimental complex.

References

Kertész, A. & Rákosi, Cs. (2012): *Data and evidence in linguistics: A plausible argumentation model.* Cambridge: Cambridge University Press.

Kertész, A. & Rákosi, Cs. (2014): The p-model of data and evidence in linguistics. In: Kertész, A. & Rákosi, Cs. (eds.): *The evidential basis of linguistic argumentation.* Amsterdam & Philadelphia: John Benjamins, 15-48.

Rákosi, Cs. (2012): The fabulous engine: strengths and flaws of psycholinguistic experiments. *Language Sciences* 34, 682-701.

Rákosi, Cs. (2014): On the rhetoricity of psycholinguistic experiments. *Argumentum* 10, 533-547. http://argumentum.unideb.hu/2014-anyagok/angol_kotet/rakosicsi.pdf.

Rákosi, Cs. (2017a): Replication of psycholinguistic experiments and the resolution of inconsistencies. *Journal of Psycholinguistic Research* 46(5), 1249-1271.

Rákosi, Cs. (2017b): 'Experimental complexes' in psycholinguistic research on metaphor processing. *Sprachtheorie und germanistische Linguistik* 27(1), 3-32.

Reijnierse, W.G., Burgers, C., Krennmayr, T., & Steen, G.J. (2015): How viruses and beasts affect our opinions (or not): The role of extendedness in metaphorical framing. *Metaphor and the Social World*, 5, 245-263. doi:10.1075/msw.

Steen, G.J., Reijnierse, W.G., & Burgers, C. (2014): When do natural language metaphors influence reasoning? A follow-up study to Thibodeau and Boroditsky (2013). *PLOS ONE*, 9(12), e113536. DOI: 10.1371/journal.pone.0113536.

Thibodeau, P.H. (2016): Extended metaphors are the home runs of persuasion: Don't fumble the phrase. *Metaphor and Symbol* 31(2), 53-72.

Thibodeau, P.H., & Boroditsky, L. (2011): Metaphors we think with: The role of metaphor in reasoning. *PLOS ONE*, 6(2), e16782. DOI: 10.1371/journal.pone.0016782.

Thibodeau, P.H., & Boroditsky, L. (2013): Natural language metaphors covertly influence reasoning. *PLOS ONE*, 8(1), e52961. DOI: 10.1371/journal.pone.0052961.

Thibodeau, P.H., & Boroditsky, L. (2015): Measuring effects of metaphor in a dynamic opinion landscape. *PLOS ONE*, 10(7), e0133939. doi:10.1371/journal.pone.0133939.

Dr. Csilla Rákosi
MTA-DE-SZTE Research Group for Theoretical Linguistics
University of Debrecen
Pf. 400
H-4002 Debrecen
rakosics@gmail.com