Martin Hoffmann

Considering Criteria for Model Modification and Theory Change in Psychology

Commentary on Reinhold Kliegl and Ralf Engbert

1 Introduction

In contemporary philosophy of science, there are two main approaches to reconstructing the relation between models and theories. According to the so-called semantic view of theories, theories are just families of models. It was Patrick Suppes who declared that “a theory is a linguistic entity consisting of a set of sentences and models are non-linguistic entities in which the theory is satisfied” (Suppes 1960, p. 290); and this basic idea was elaborated by Sneed (1971), van Fraassen (1980), Stegmüller (1986), Suppe (1989), and others. Recently, this approach was criticised by some authors because it cannot account for the complex role that models play in scientific practice (Morrison and Morgan 1999; Suárez 1999). These authors propose a second, alternative approach to the relation between models and theories which regards models as almost independent from theories. Models are “autonomous agents” (Morrison 1999) that mediate between the level of general theories and concrete empirical data, and fulfil a variety of functions: models are useful instruments to represent reality, but also to test, explore, and elaborate theories.

Kliegl and Engbert outline a picture of modeling that fits the second approach of the model/theory-relation. In their paper they discuss many interesting and innovative ideas how to enhance the methods and how to expand the criteria for model evaluation in psychology. The main aim of this endeavour is to gain a powerful tool which will enable us to identify the best model amongst the many alternatives suggested in the present psychological discussion.

Kliegl and Engbert’s account is inspired by a radical critique of the current practices of model evaluation in psychology, which have been formulated by Roberts and Pashler (2000, 2002). The main point of Roberts and Pashler’s challenge is the following: Many theories in psychology are primarily tested and confirmed by their ability to fit the data. The methodological discussion focuses on developing methods to determine the goodness of fit. But, say Roberts and Pashler, goodness of fit alone is not sufficient to determine the empirical validity and the explanatory power of a model. They propose three additional criteria, namely model strictness, reliability of data and unexpected predictions. In their paper, Kliegl and Engbert apply these criteria in a subtle way to different versions of their SWIFT model. They show that model strictness and reliability of data can be measured at least in principle, and that they can help to compare the adequacy of SWIFT with competing accounts. They identify the third criterion of Roberts and Pashler's as the main difficulty: unexpected, but correct model predictions. On the one hand, Kliegl and Engbert qualify this criterion as the "gold standard for a model" (section 5.2). On the other hand, they conclude that the SWIFT model meets this criterion only in some cases. It is more common that, in the beginning, there are surprising experimental findings. Then these findings are compared with model predictions, and one has to test whether the model can fit the data – which often requires a suitable modification. The central question is this: is such a modification of the model legitimate? Or should its failure to generate a correct prediction be regarded as a strong reason to abandon the model? I can only see two possible solutions: either the third criterion by Roberts and Pashler is unreasonably strict, or it is methodologically problematic to adhere to a model which generates wrong predictions.

My aim in this commentary is to discuss this tension on the basis of concrete examples from the SWIFT model. But before I will do that, it is necessary to say something about the epistemological reasons for the third criterion.

2 Why are unexpected, but correct model predictions important at all?

Roberts and Pashler take the idea for their third criterion from Imre Lakatos' philosophy of science. Kliegl and Engbert quote Lakatos' claim that not all predictions, but only the "dramatic, unexpected, stunning predictions" can corroborate the theory in question (Lakatos 1978a, p. 6). But why should it be important that an empirical result is novel, unexpected, or even stunning? These seem to be merely psychological categories, and it is not at all clear why such emotive responses should have any impact on the corroboration of theories and models. In order to see why these features are of any epistemological importance, it is necessary to make some remarks about Lakatos' philosophy of science.

Lakatos has formulated a re-statement of Popper's well-known idea of falsification as a criterion for the validity of empirical theories. Popper himself thought that success in science is primarily determined by strict tests of theories. A theory is falsified if a conflict with empirical data is indicated. In Lakatos' view, this version of falsificationism is naïve, so he puts his own version of sophisticated falsificationalism forward. Lakatos' main criticism is that Popper has construed the central concepts of theory acceptance and falsification by employing two-place relations: observational data on the one hand, and the theory on the other hand. (By the way, in this respect there is a resemblance between Popper's ideas and the criterion of goodness of fit – which means fitting just one theory or model to a data structure). Lakatos thinks that this reconstruction is inadequate because scientific progress is only possible if a falsified theory is replaced by a promising successor. For this reason, in Lakatos' view, the competition between rival theories is neither an accidental property of science nor an indicator of a crisis; it is rather an essential element of fruitful theory development. Popper's two-place relation has – in Lakatos' conception of sophisticated falsificationalism – been replaced by a three-place relation between observational data and at least two rival theories T and T′.

> A theory T is falsified if and only if a different theory T′ is proposed which exhibits the following characteristics:
> (1) T′ has excess empirical content over T …;
> (2) T′ explains the previous success of T …
> (3) some of the excess content of T′ is corroborated. (Lakatos 1978b, p. 32)

The requirement that T′ has to be in accordance with novel, unexpected facts is essential for condition (1): "Excess empirical content" does not only mean that the area of application is broader, or the class of predictions bigger, than that of T. What is important here is that T′ predicts facts that are improbable or forbidden according to T. This is precisely what constitutes the distinction between well-known and novel, unexpected facts relevant in the present context. Known and expected facts are facts that are in accordance with both theories T and T′. These facts are uninteresting for testing the new theory, because on their basis alone no decision between T and T′ is possible. In contrast to that, reference to novel and unexpected facts (like the return of Halley's Comet) is decisive, because their confirmation allows for a justified choice between T and T′. The confirmation of correct but unexpected predictions is the decisive criterion to identify the theory with excess empirical content. But if one changes the content of the theory T′ in the light of new data, the decision between T and T′ becomes arbitrary, since it is possible to immunize every theory against conflicting data by making ad hoc assumptions.

For this reason, modifications in reaction to new data are problematic. If one allows for modifying T′ in a way to accommodate the new data, Lakatos' theory of justified theory choice no longer works. Kliegl and Engbert point out that model modifications in fact do play an important role in current model evaluation. So the question arises whether one can give a rational reconstruction of this methodological procedure. Let us have a closer look at Kliegl and Engbert's examples.

3 Examples

Kliegl and Engbert present an example that fulfils the third criterion: concerning refixation probabilities after skipped content and function words, the SWIFT model generates unexpected but correct predictions (section 5.2.3). But they themselves admit that this example is "not representative of normal model development" (section 5.2). Usually, a process of model modification is initiated in the light of new experimental data.

Kliegl and Engbert present two examples to illustrate this process (sections 5.2.1 and 5.2.2). They focus on the fixation durations prior to skipped words: in the majority of cases word skipping generates a *skipping cost*, that means an increase of the fixation duration before skipped words compared to the fixation duration before fixated words. But experimental findings surprisingly show that there are also *skipping benefits* (that is *decreased* fixation durations before skipped words), which occur when the words are short or occur with high frequency.

These skipping benefits conflict with the predictions of the initial SWIFT model, presented by Engbert et al. (2005). But in contrast to alternative models for eye movements during reading (like the E-Z-Reader model), the SWIFT model's predictions can be altered by varying the value of a free model parameter. The SWIFT model can account for a dynamical modulation of the perceptual span depending on word length. This modification allows for fitting the model to the experimental data. The modified SWIFT model is in accordance with the experimental data and predicts skipping benefits when the words in question are short. This methodological move deserves careful interpretation. Let me analyse in more detail which function the model modification might have in this particular context.

First of all, it has to be said that the better fit to the data generated by the modification contributes nothing to the corroboration of the model. Its initial predictions went wrong and it is re-stated given the conflicting data. So it does not exceed the empirical content of the old model in this respect. But nevertheless there is an important difference to a mere ad hoc hypothesis: the model assumptions are not just relaxed, but changed. This means that the old model predictions are partly replaced by new ones. Even if this does not corroborate the model, it may turn out theoretically fruitful: the new predictions can at least potentially be confirmed by new, unexpected empirical data. In fact, this is the case in Kliegl and Engbert's second example. They report that the invention of a dynamic perceptual span in the SWIFT model does not affect the goodness of fit in one other important respect (section 5.2.2). In a sense, new and successful predictions can outweigh the model relaxation caused by the model modification.

But perhaps the focus on the predictive power of models is too narrow. There may be other aspects that have to be taken into account to describe a model's methodological role. In order to explain this, I would like to draw attention to the distinction introduced at the beginning: the distinction between models and more general and unified theories. It is important to notice that Lakatos only considers research programmes on the level of *theories*. On this general level, the methodological aim is to replace one theory by a progressive successor. But it is questionable whether the relation between a model and its successor has to be reconstructed in the same way. Following Morrison (1999), models are "autonomous agents" and for this reason it is an oversimplification to identify series of models with series of succeeding theories. According to this view, models are specific, flexible instruments that can be characterized independently of the underlying theories and that can be used for a variety of methodological purposes. But this type of "autonomy" of models does not imply that the theories themselves are irrelevant. On the contrary, the development of models is no end in itself. Models are rather designed as instruments or tools for developing more unified theories. They simply do not only serve to corroborate the theory, but also to explore the theory, develop new predictions, apply the theory to special and new areas, etc. So in order to integrate models into Lakatos' picture of the development of research paradigms, it is important to clarify the relations that obtain between models and general theories.

Therefore, in the present context the following question becomes central: what *is* the theory in question that should be refined by the modifications of the model? Kliegl and Engbert do not say much about the relation between the SWIFT model and a corresponding theory. But I think that the conflict between the two following theories lies at the heart of their project: their aim is to confirm the theory of parallel or distributed processing and to argue against a theory of serial or sequential processing. Serial processing is characterized by two assumptions: (i) attention is focused on just one word at a time and (ii) attention shifts mandatorily from one word to the next. Distributed processing is characterised by loosening both assumptions: (i′) attention is a process allocated parallel on different words, which is called an activation field, (ii′) attention shifts are explained by different patterns. These general theories are not restricted to eye movement control, but claim to give unified accounts of motor behaviour in general. Only one of these theories corresponds to the SWIFT models: the theory of distributed processing. For this reason the implementation of a dynamical perceptual span is possible in the SWIFT model, but it is not in line with models like the E-Z-Reader, formulated on the basis of the serial processing theory. Against this background, we can interpret the central function of the model modification in question –

while fitting the data – to generate new predictions for future confirmations of the theory.

Because of the complex relation between the SWIFT model and the distributed processing theory, it is difficult or impossible to define strict normative standards for particular model changes at the present time. But perhaps one can formulate two modest requirements for the relation between theory and models instead. First, it is of crucial importance that every model modification remains in accordance with the core assumptions of the theory. Second, every model modification should be defined such that the predictions of the new model still contradict the predictions of the competing theory. Only if these conditions are fulfilled, the modified model remains a useful instrument for developing the theory in question. In this particular case the model predictions are indeed in conflict with the predictions of the competing theory of serial processing, because the skipping benefits are extremely difficult to explain assuming serial processing. Serial processing theory explains word skipping as a termination of the saccade program if the next word is recognized. In this case the saccade program is cancelled and restarted to fixate the next but one word. But this process can only lead to skipping costs, not to skipping benefits. So the results of the modified SWIFT model, which predicts skipping benefits, contradict the theory of serial processing, but are perfectly in line with the theory of distributed processing.

4 Conclusion

To sum up, I am in favour of the following position concerning the plausibility of Roberts and Pashler's third criterion: prediction of unexpected, but correct empirical results. Applied to models, it is obviously too strong in its unrestricted formulation. It would be even irrational to require the fulfilment of this restrictive criterion as a necessary condition for accepting a model as part of a progressive research programme. However, this does not speak against Lakatos' ideas about strict theory testing. One has to consider that Lakatos applies this criterion only to the level of general *theories*, not to models. For this reason it is perfectly in line with Lakatos' account to maintain the third criterion as a plausible methodological rule for theories, and to abandon it for models.

In Lakatos' account it remains underdetermined how to react to conflicting data on the level of models. Kliegl and Engbert present many appealing and original ideas concerning model modifications in the particular case of the SWIFT model, but it remains difficult to derive more general rules from these. For now, it is merely possible to formulate one modest methodological restriction: model modifications, which *only* relax the empirical content of a model, are problematic, because of their ad hoc character. The class of empirical data, which are in accordance with the model after its modification, should be outweighed by a class of data which were compatible with the previous model, and are not in line with its predictions after modification. This restriction is important to prevent mere ad hoc modifications which might render the model compatible with *every* experimental result.

So, primarily, this commentary is not intended to be a criticism of Kliegl and Engbert's application of the criteria by Roberts and Pashler, but rather a criticism of the application of methodological criteria for strict theory testing in the evaluation of modern, complex models. If one adopts the view of models as autonomous agents – suggested by Morrison, Morgan and others –, it remains an important task for future research in philosophy of science to clarify the complex relations between theories and models in more detail and to define precise methodological rules for how to modify models in the light of new data.

References

Engbert, Ralf, Nuthmann, Antje, Richter, Eike M., & Kliegl, Reinhold (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review* 112. 777–813.

Fraassen, Bas C. van (1980). *The Scientific Image*. Oxford: Oxford UP.

Lakatos, Imre (1978a). Introduction: Science and Pseudoscience. In: Currie, John Worrall Gregory (ed.). *The Methodology of Scientific Research Programmes. Philosophical Papers* 1. Cambridge: Cambridge UP. 1–7.

Lakatos, Imre (1978b). Falsification and the Methodology of Scientific Research Programmes. In: Currie, John Worrall Gregory (ed.). *The Methodology of Scientific Research Programmes. Philosophical Papers* 1. Cambridge: Cambridge University Press. 8–101.

Morrison, Margaret (1999). Models as Autonomous Agents. In: Morgan, Mary S. & Morrison, Margaret (ed.). *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press. 38–65.

Morrison, Margaret & Morgan, Mary S. (1999). Models as Mediating Instruments. In: Morgan, Mary S. & Morrison, Margaret (ed.). *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press. 10–37.

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

Roberts, Seth & Pashler, Harold (2002). Reply to Roders and Rowe. *Psychological Review* 109. 605–607.

Sneed, Joseph D. (1971). *The Logical Structure of Mathematical Physics*. Dordrecht: Reidel.

Stegmüller, Wolfgang (1986). Theorie und Erfahrung: Dritter Teilband. Die Entwicklung des neuen Strukturalismus seit 1973. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie* II/3. Berlin: Springer.

Suárez, Mauricio (1999). Theories, Models, and Representations. In: Magnani, Lorenzo, Nersessian, Nancy J. & Thagard, Paul (ed.). *Model-Based Reasoning in Scientific Discovery*. New York; Boston; Dordrecht: Kluwer. 75–83.

Suppes, Patrick (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese* 12. 287–301.

Suppe, Frederick (1989). *The Semantic View of Theories and Scientific Realism*. Urbana; Chicago: University of Illinois Press.

Dr. Martin Hoffmann
University of Hamburg
Department of Philosophy
Von-Melle-Park 6
20146 Hamburg
Germany
martin.hoffmann@uni-hamburg.de