19 From Popper to Lakatos: A Case for Cumulative Computational Modeling

Ardi Roelofs
Max Planck Institute for Psycholinguistics and
F.C. Donders Centre for Cognitive Neuroimaging,
Nijmegen, The Netherlands

An important problem with several modeling enterprises in psycholinguistics is that they are not cumulative, unlike successful experimental research. For example, in the field of language production, quite a few models focus on a few findings only instead of trying to account simultaneously for a wide range of data. Even worse, some investigators treat their models like their toothbrushes by using them only for their own data. There is no guarantee that these micromodels can be integrated into a single comprehensive macromodel, because micromodels are often mutually incompatible. Moreover, experimental tests of models developed by others are often conducted in the world of a misinterpreted Popper, where testing models is like skeet shooting. [1] The aim is to shoot down

[1] Lakatos (1970) distinguished three Poppers: Popper_0, Popper_1, and Popper_2. "Popper_0 is the dogmatic falsificationist who never published a word: he was invented—and 'criticized'—first by Ayer and then by many others. Popper_1 is the naive falsificationist, Popper_2 the sophisticated falsificationist. The *real* Popper developed from dogmatic to a naive version of methodological falsificationism in the twenties; he arrived at the *'acceptance rules' of sophisticated falsificationism* in the fifties. Thus the real Popper consists of Popper_1 together with some elements of Popper_2" (p. 181). Skeet shooting is often defended by referring to the mythical Popper_0.

models with falsification bullets. Alternatively, Lakatos proposed to treat models like graduate students. Once admitted, one tries hard to avoid flunking them out (of course, not at all costs) and one spends much time and effort on their development so that they may become long-term contributors to science (cf. Newell, 1990).

In this chapter, I make a case for Lakatos-style or cumulative computational modeling and model testing. This involves working with a single model that accounts for a wide range of existing data and that is incrementally extended and tested on new data sets. First, I contrast cumulativeness in relation to modeling with the noncumulative toothbrush and skeet shooting approaches. Next, I describe the cumulative modeling approach in which models are treated like graduate students. Finally, I demonstrate the cumulative modeling approach by describing the scientific career of one of my own model graduate students, namely the WEAVER++ model of spoken word production.

TOOTHBRUSHES, SKEET SHOOTING, AND GRADUATE STUDENTS

Cumulativeness in relation to modeling means that in developing models one builds on earlier modeling results, just as one does in cumulative experimental research. Cumulativeness in relation to modeling is not always seen as a virtue. For example, a goldfield for modeling in psychology is the literature on the color-word Stroop task (Stroop, 1935), one of the most widely used tasks in academic and applied psychology (between 1965 and 2003, some 2000 articles appeared on the task, partly reviewed by MacLeod, 1991). The task requires naming the ink color of written color words or reading the words aloud. The basic finding is that participants are much slower and make more errors in naming the ink color of an incongruent color word (e.g., saying "red" to the written word BLUE in red ink) than the ink color of a congruent word (the word RED in red ink). When the task is to read aloud the words and to ignore the ink colors, there is no congruity effect. Despite the extensive accumulating literature on this phenomenon, Stroop modeling has not been cumulative.

Since the early 1990s, the literature on Stroop has been dominated by the model of Cohen, Dunbar, and McClelland (1990). This feedforward model was discarded by its main designer, Cohen, in the mid-1990s (Cohen & Huston, 1994) in favor of a similar interactive model. However, the new model was not tested against all the data that motivated the construction of the old model. Moreover, no experiments were run that tested the new against the old model. Rather, it seems that the old feed-

forward model was dismissed only because interactiveness had become part of the Zeitgeist. So, it is unclear whether the new interactive model represented any improvement over the old feedforward model.

Although Cohen et al. (1990) did not conduct any new experiment to empirically test their Stroop model against extant models in the literature, there was at least an attempt to provide an account of a wide range of existing data. Unfortunately, this is not even attempted in two popular approaches to modeling and testing models in psycholinguistics, namely the toothbrush and the skeet shooting approaches.

The toothbrush approach involves constructing a model for your own data only. Success for the model is claimed by pointing to the fit of the model to the data it was designed to explain. For example, Cutting and Ferreira (1999) and Starreveld and La Heij (1996) reported new data on word production together with new models that were designed to account for these data. The toothbrush approach is popular with several journals, because it leads to self-contained publications. The article reports new data and a model that accounts for the new data. The model is often very simple, because the only thing it has to do is to account for the reported data and nothing else. Regularly, the approach is defended as an application of Ockham's razor: Accept the simplest model that works for the reported data. It is thereby forgotten that Ockham's rule does not apply to both model and data. Ockham understood his principle as recommending models that make no more assumptions than is necessary to account for the phenomena. But he did not advocate to keep the number of phenomena to a minimum. Ockham's rule is an important guiding principle in model construction (do not introduce any needless assumptions in your model) and a last resort in testing between models. It applies when two models make identical predictions or when there are no more phenomena to use as a test between models. But the latter is almost never the case in psycholinguistics.

The biggest problem with the toothbrush approach is that it cuts on both the number of theoretical assumptions and the number of phenomena. Moreover, there is no attempt at snowballing, that is, to build on earlier empirical and modeling results. Ultimately, however, we want to have unified theories explaining how language works (i.e., how language is acquired and used in production and comprehension). The toothbrush approach commonly leads to several micromodels each capturing a different aspect of reality but together not giving a consistent picture.

For example, motivated by empirical phenomena suggesting interaction, but perhaps also partly inspired by the Zeitgeist, most existing computationally implemented models of spoken word production are interactive (e.g., Cutting & Ferreira, 1999; Dell, Schwartz, Martin, Saffran,

& Gagnon, 1997; Starreveld & La Heij, 1996). However, the design characteristics of these models differ greatly. For example, the model of Cooper and Ferreira (1999) assumes inhibitory interactions, whereas the models of Dell et al. (1997) and Starreveld and La Heij (1996) do not. Also, the empirical domains of the models differ. For example, the model of Dell et al. (1997) was designed to explain speech errors, whereas the model of Starreveld and La Heij (1996) was designed to explain production latencies. Because the design characteristics and domains of the models differ, collectively they do not make up a single interactive model. Therefore, what counts as empirical success for one interactive model does not automatically count as success for all the other models. For example, Dell et al. (1997) addressed interactive effects on segmental speech errors and Starreveld and La Heij (1996) addressed interactive effects in the picture-word interference task. However, the model of Starreveld and La Heij cannot account for the interactive effects on segmental speech errors, simply because it has no segmental level of representation. Moreover, the model of Dell et al. cannot account for the interactive effects in the picture-word interference task, simply because it cannot account for latencies at all. Thus, it makes little sense to point to the success of *the* interactive approach by referring to the success of the various interactive models. Instead, to make a convincing case for an interactive account, a unified interactive model is required that can account for a wide range of findings both on production errors and latencies.

In the skeet shooting approach, the aim of the experimenter is to collect data that blast models. If the collected data disagree with one or more models, the mission is accomplished and the data are published with the recommendation that a completely new model is developed. For example, Caramazza and Costa (2000) reported a series of experiments that tested the response set assumption made by the WEAVER++ model of spoken word production (Levelt, Roelofs, & Meyer, 1999; Roelofs, 1992). According to Caramazza and Costa (2000), the outcomes of their experiments were problematic for WEAVER++ and they demanded a fundamental modification of the model: "It is not obvious that minor changes to the model—that is, changes that do not alter the fundamental architecture of the model— would be successful in this regard" (p. B61). Therefore, Caramazza and Costa (2000) took it that their study "undermines the model as a whole" (p. B61). They concluded that "if one were willing to drop the response set principle used in WEAVER++, the *new* model would have to be able to account for the data reported here and the various other data that were previously used to support the old WEAVER++ model" (p. B61). Although the response set assumption was assumed to

be refuted by Caramazza and Costa (2000), an alternative was not considered.

The skeet shooting approach is also popular with journals, because it gives the impression that we are making scientific progress. After all, we have eliminated a model or a class of models. But usually, no answer is given to the critical question: What next? The problem is that models may be wrong for various reasons. For example, models may be incomplete. In the latter case we only need to extend the model rather than construct a completely new one. Alternatively, only a small change to an existing model may be required to remedy the problem. For example, in response to Caramazza and Costa (2000), I argued that there is no need for a fundamental change of the WEAVER++ model (Roelofs, 2002). Instead, the supposedly problematic findings of Caramazza and Costa (2000), and all previous findings that support the model, could be explained by assuming that a response set is only marked in memory when the number of responses is small and can be kept in short-term memory. Thus, a small change in an assumption of the model could do the job. There is a motto in politics saying that you cannot beat something with nothing. You cannot beat a candidate simply by pointing to inadequacies, but you must offer an alternative. The same applies to testing and modeling. But the skeet shooting approach fails to point to new directions.

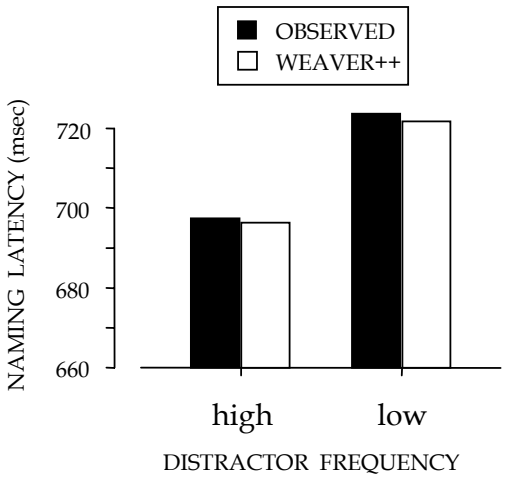

FIG. 19.1 The effect of high- versus low-frequency distractors in picture naming: Observed data (Miozzo & Caramazza, 2003) and WEAVER++ simulation results.

Moreover, when proposing an alternative, it is important to make sure that the alternative is really warranted. When a model is rejected and a new assumption is considered as the starting point of a new model, it should be excluded that the rejected model with the new assumption would fit the data equally well. For example, Miozzo and Caramazza (2003) observed that high-frequency distractor words yielded less interference than low-frequency distractor words in picture naming and they argued that "it is clear that the distractor frequency interference effect seriously challenges a popular model of lexical access", namely WEAVER++ (p. 249). To account for their novel finding, Miozzo and Caramazza (2003) proposed a new frequency-sensitive mechanism by which distractors are actively blocked. But Roelofs (2003) and Roelofs and Hagoort (2002) proposed exactly such a blocking mechanism, namely production rules blocking out distractors, although they did not *explicitly* assume that the blocking rules are frequency sensitive. However, given that *all* production rules in WEAVER++ are frequency sensitive (as acknowledged by Miozzo & Caramazza, 2003), frequency-sensitive blocking of distractors is entailed. Figure 19.1 shows that when the frequency of the blocking rules is manipulated, WEAVER++ fits the data of Miozzo and Caramazza (2003) without difficulty. To conclude, in rejecting a model and proposing a new assumption as the starting point of a new model, one should not be blind to the possibility that making the new assumption for the rejected model would fit the data equally well. If the latter is the case, the data require a model patch rather than a construction from scratch.

As an alternative to the toothbrush and skeet shooting approaches, I propose to treat models like graduate students. Once admitted, you spend time and effort on their development in the hope that they become long-term contributors to psycholinguistics. You extend their theoretical content and empirical coverage by confronting them with new data sets. Of course, they are flunked out when they fail too many tests or when they are not productive for a long period of time.

Treating models like graduate students represents a more conservative approach to model testing than skeet shooting. The conservative protectiveness is not unreasonable. In an empirical science like psycholinguistics, we try hard to achieve approximate truths. It would be a mistake to believe that we can find a single simple model that captures the whole truth and nothing else. Instead, we hope to see the light by a strategy of continual approximations. It is said that Thomas Edison ran more than two thousand experiments before he got an adequately working light bulb. When asked how he felt about having failed so many times, Edison replied "I never failed once. It just happened to be a 2000 step process".

Moreover, we try to avoid the mistake of Jorge Luis Borges' (1985) cartographers, who constructed a map that was as big and detailed as the country itself—capturing most of reality but being completely useless. In order to be useful, models have to simplify reality. When we find discrepancies between model and data, it is therefore reasonable to first try to patch rather than to rebuild from scratch. As with training real graduate students, constructing a new model is a costly project, taking much time and effort. Moreover, when discrepancies between model and data appear, it is often not immediately obvious where the difficulty lies. It may be located in a fundamental assumption of the model, but it may as well be merely a defect in one of the simplifying assumptions, auxiliary hypotheses, or measurement assumptions that had to be made in order to connect the model with data. Increasing complexity or revising the auxiliary hypotheses or measurement assumptions may be sufficient to save the model.

The critical importance of localizing the fault rather than just noting that there exists a discrepancy was pointed out by Popper and Lakatos. Whereas nineteenth century philosophers of science tended to stress the importance of justifying a model, Popper stressed the importance of finding and understanding discrepancies. Discrepancies can only arise when models stick out their neck by excluding certain data patterns ("No guts, no story"). Models should be falsifiable. According to Popper, we can only make scientific progress when there are discrepancies between model and data. A discrepancy is not necessarily a falsification. As indicated, the trouble may be located in a fundamental assumption of the model, but it may as well be merely a shortcoming of an auxiliary hypothesis or a measurement assumption. A discrepancy only leads to scientific progress if it shows the way to a new theoretical claim, either in terms of a revision of theory or model, a revised auxiliary hypothesis, a revised measurement assumption, or a new theory or model.

For Popper, falsification concerned a relation between model and data, although "in most cases we have, before falsifying a hypothesis, another one up our sleeves" (Popper, 1959, p. 87). For Lakatos, there *must* be an alternative, that is, a presumed new insight: "There is no falsification before the emergence of a better theory. ... Refutation without an alternative shows nothing but the poverty of our imagination in providing a rescue hypothesis" (Lakatos, 1970, pp. 119-120). In cumulative computational modeling, there is, by definition, always an alternative.

CUMULATIVE COMPUTATIONAL MODELING

A computational model is a formalization of a theory in terms of a computer program (unfortunately, in practice, computational models are frequently constructed without a theory, which holds especially for many connectionist models, see Norris, this volume). Computational models have many advantages over verbal models. Computational models guarantee the sufficiency and internal consistency of a theory. By running computational models as computer simulations, one can assess whether the theoretical assumptions are sufficient to explain the data. Moreover, computer simulations reveal whether the theoretical assumptions are mutually consistent, because inconsistencies will stop a simulation. Another advantage of computational models over verbal models is that they generate precise predictions.

A disadvantage of computational models compared to verbal models is that in order to make the model run as a computer simulation, sometimes assumptions have to be made that were not part of the theory. Thus, a computational model may be more specific than its theory. This complicates the testing of model and theory. When we find discrepancies between model and data, the trouble may lie in the specific assumptions of the model or in the assumptions of the theory that it implements. When the problem lies in the model-specific assumptions, revision of these assumptions may be sufficient to save both theory and model. Of course, when the trouble lies in the assumptions of the theory that the model implements, revision or rejection of the theoretical assumptions is necessary to save theory and model. Figure 19.2 illustrates the empirical cycles.

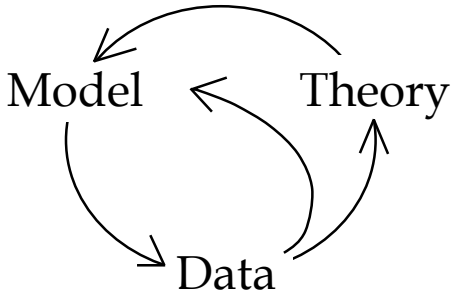

FIG. 19.2 The two main empirical cycles involved in constructing and testing a model for a theory.

Given that the ultimate goal of psycholinguistic research is to obtain comprehensive theories of how language works, it makes little sense to develop models that focus on a few findings only instead of trying to account concurrently for a wide range of data. Moreover, it makes no sense to construct models for your own data only. There is no warranty that these micromodels can be integrated into a single comprehensive macromodel, because micromodels are often irreconcilable. Moreover, it makes little sense to test models with the only aim to obtain a mismatch between model and data. A discrepancy should be a new beginning of theorizing. Thereby, theorizing and modeling should be cumulative, just like successful experimental psycholinguistic research.

Cumulativeness in relation to modeling can take a number of forms. The best known form is probably *nested modeling* (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Jacobs & Grainger, 1994). In nested modeling, a more extensive version of a model is tested against a more restricted version of the same model to see which model version gives a better fit to a particular set of data. A pitfall is overfitting. The more complex model may provide a better fit only because it has more parameters and therefore can fit not only the main trend but also some of the noise (random error) in the data. The remedy is to test for generalizability, that is, to test the simple and complex versions on other relevant sets of data (e.g., Pitt & Navarro, this volume). For example, overfitting of a model of spoken word recognition may be prevented by testing it not only on data obtained by lexical decision, but also on data from phoneme monitoring, word spotting, eye tracking, and so forth. If the more complex model did better than the simple model on the lexical decision data because it fitted some of the noise in the data, it most likely does worse on the wider range of data sets.

Nested modeling by itself does not lead to comprehensive models of how language works. A model of word recognition tested on data from lexical decision and phoneme monitoring remains a model of word recognition regardless of whether it is also tested on word spotting and eye-tracking data. In order to attain comprehensive models of how language works, one needs to extend models beyond the empirical domain for which they were originally developed. For example, to attain a model of spoken word recognition *and* word production, one needs to extend the model of spoken word recognition by including assumptions about word production, or vice versa, and test the extended model on relevant data. The incremental extension of a model to a new empirical domain outside its current scope is *incremental modeling*. Note that an incremental extension of a model also implies an incremental extension of the corresponding theory. Extending a model of spoken word recognition by including

assumptions about word production implies making theoretical assumptions about word production. Every extension should rule out certain data patterns.

Unfortunately, compared to nested modeling, incremental modeling further complicates the testing of model and theory, which is the price paid for achieving comprehensive coverage. When we find discrepancies between the extended model and data, the trouble may lie in the assumptions of the extension or in assumptions of the original model and theory. Given that there are more possible loci of trouble, cumulative modeling and testing might seem to be a hopelessly complicated endeavor. However, in practice, this is not the case, especially not if one extends a model in a modular fashion by adding theoretical assumptions without changing existing ones. This guarantees that the fits of the original model are preserved.

A SKETCH OF THE SCIENTIFIC CAREER OF WEAVER++

In this section, I demonstrate the incremental approach by describing the scientific career of one of my own models, namely WEAVER++. I describe some of the major steps in developing WEAVER++. The steps range from WEAVER++'s origin as a model designed to explain chronometric findings on lemma retrieval from picture-word interference experiments to its current state as a comprehensive model of the various processes underlying word production, including its relation with spoken and visual word recognition, their attentional control, the self-monitoring for speech errors, and the relation between self-monitoring and speech comprehension. Whereas the original model was designed to explain chronometric data, recently WEAVER++ has been extended to eye-tracking, electrophysiological, and neuroimaging data. Stroop-like tasks have run as a continuous thread trough WEAVER++'s career and they are therefore used for illustrative purposes.

Figure 19.3 gives an overview of all the processing components assumed by the current version of WEAVER++. The architecture of the model is derived from Levelt's (1989) blueprint of the speaker. The blueprint embeds the architecture in the general context of sentence and discourse production. The architecture distinguishes between conceptual preparation, lemma retrieval, and word-form encoding, with the encoding of forms further divided into morphological, phonological, and phonetic encoding. Information is retrieved from a lexical network by spreading activation. During conceptual preparation, concepts are flagged as goal concepts. In lemma retrieval, a goal concept is used to

retrieve a lemma from memory, which is a representation of the syntactic properties of a word, crucial for its use in sentences. For example, the lemma of the word *red* says that it can be used as an adjective. Lemma retrieval makes these properties available for syntactic encoding processes. In word-form encoding, the lemma is used to retrieve the morpho-phonological properties of the word from memory in order to construct an appropriate articulatory program. For example, for *red* the morpheme <red> and the speech segments (e.g., /r/) are retrieved and a phonetic plan is generated. Finally, articulation processes execute the motor program, which yields overt speech.

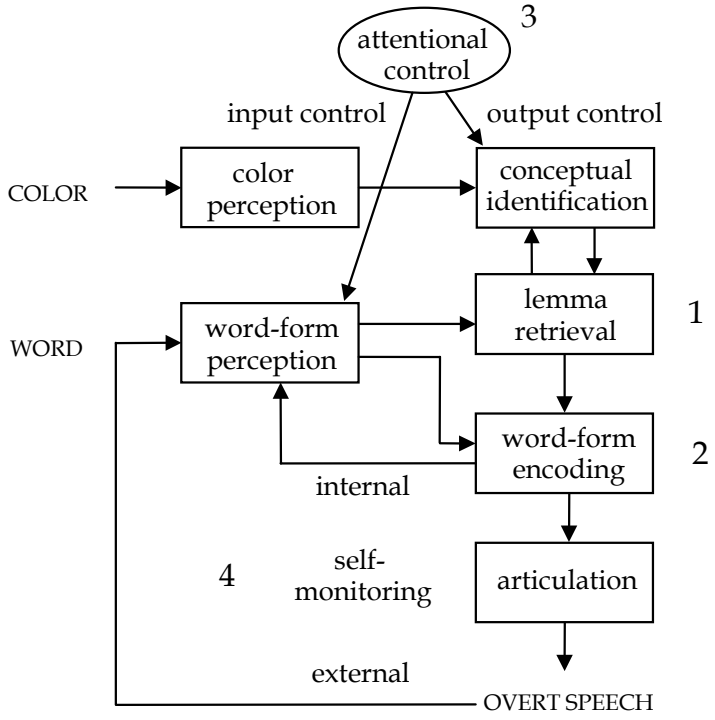

FIG. 19.3 The architecture of WEAVER++. The numbers indicate major steps in the incremental development of the model: (1) lemma retrieval, (2) word-form encoding, (3) attentional control, and (4) self-monitoring and its relation with speech comprehension.

Assume a speaker wants to refer to the ink color of the word BLUE in red ink. This involves the conceptual identification of the color based on the perceptual input and its designation as goal concept (i.e., RED(X)), the retrieval of the lemma of the corresponding word (i.e., *red*), and the encoding of the form of the word. The final result is a motor program for the word "red", which can be articulated. In performing the color-word Stroop task, aspects of word planning are under attentional control. The system has to achieve color naming rather than word reading ("output control") and the irrelevant input (the word in color naming) has to be suppressed ("input control"). Moreover, speakers monitor their performance. In Stroop's (1935) original experiments, participants had to repair their errors, and this still holds for most psychometric applications of the task.

WEAVER++'s career started about a decade ago as an anonymous computational model of lemma retrieval (Roelofs, 1992). Although the model was developed within the theoretical framework of Levelt's blueprint of the speaker, it did not simply implement the theoretical assumption about lemma retrieval in the blueprint. Instead, the model instantiated a new set of assumptions. To highlight that computational models implement a theory, the first publication on the model was called "A spreading-activation *theory* of lemma retrieval in speaking" (Roelofs, 1992).

As a next step, a computational model of word-form encoding was developed. The editor of the journal of the first publication on the word-form encoding model, David Balota, suggested that I choose a name for it (Roelofs, 1996). I decided for the name WEAVER, which is an acronym of Word-form Encoding by Activation and VERification. The acronym intended to capture the fact that words are encoded in the model by activating, selecting, and connecting types of verbal information. Unlike the lemma retrieval model, WEAVER largely followed the theoretical assumptions of the blueprint. A full description and motivation of WEAVER was published under the title "The WEAVER model of word-form encoding in speech production" (Roelofs, 1997).

The lemma retrieval model and the WEAVER model of word-form encoding were subsequently combined into a single model of word planning. This model was published as an implementation of a general theory of lexical access (Levelt et al., 1999). To highlight the incremental nature of the modeling, the combination of models was called WEAVER++. The ++ refers to the ++-operator in the C programming language, meaning "incremental extension". Thus, WEAVER++ means "incremental extension of WEAVER". Moreover, WEAVER++ plans words incrementally. Lemmas are selected for lexical concepts, morphemes for lemmas, seg-

ments for morphemes, and syllable programs for syllabified segments. Also, syllabification of segments proceeds incrementally from the beginning of the word to its end.

A combination of models may be more than the sum of the component models, because the combination may include claims about the relation between the components. Roelofs (1992) proposed an interactive model for lemma retrieval and Roelofs (1997) proposed a feedforward model for word-form encoding. In these articles, no claim was made concerning the relation between lemma retrieval and word-form encoding. Levelt et al. (1999) made the claim that only selected lemmas activate their speech segments and this was implemented by WEAVER++.

In recent years, WEAVER++ has been further extended to other domains. In addition to language, numerals constitute the second most important symbolic system employed by humans. A WEAVER++ implementation has been made for naming dice, digits, and number words. Moreover, the model has been used to address the issue of how two languages are represented and controlled in bilingual individuals. Simulations have been run for English-Spanish Stroop task performance (Roelofs, 2003). Moreover, to examine the issue of similarities and differences in word-form encoding between languages, a WEAVER++ implementation has been made (by Train-Min Chen) for a language that is very different from Dutch and English, namely Mandarin Chinese, the language with the largest number of native speakers in the world.

A further extension of WEAVER++ concerned making assumptions about the relationship between spoken word production and word recognition, assumptions about self-monitoring for speech errors, and assumptions about the relation between self-monitoring and speech comprehension (Roelofs, 2004). Moreover, WEAVER++ has been extended to the domain of attentional control. In their classic paper "Attention to action: Willed and automatic control of behavior", Norman and Shallice (1986) made a distinction between "horizontal threads" and "vertical threads" in the control of behavior. Horizontal threads are strands of processing that map perceptions onto actions and vertical threads are attentional influences on these mappings. Behavior arises from interactions between horizontal and vertical threads. WEAVER++ implements specific claims about how the horizontal and vertical threads are woven together in planning spoken words. A central claim embodied by WEAVER++ is that the control of word perception and production is achieved symbolically rather than purely associatively. WEAVER++'s lexical network is accessed by spreading activation while condition-action rules determine what is done with the activated lexical information depending on the task. When a goal symbol is placed in working

memory, the attention of the system is focused on those rules that include the goal among their conditions (e.g., those for color naming rather than reading in the Stroop color naming task).

The fruitfulness of the incremental modeling approach was recently demonstrated by WEAVER++'s successful simulation of 16 classic data sets on Stroop-like performance, mostly taken from the review by MacLeod (1991), including incongruency, congruency, reverse Stroop, response set, semantic gradient, time course, stimulus, spatial, multiple task, manual, bilingual, training, age, and pathological effects (Roelofs, 2003). With only 3 free parameters taking 2 values each to accommodate task differences (color naming, picture naming, word reading, manual responding), the model accounts for 96% of the variance of the 16 studies. In addition, new empirical work refuted a rescue hypothesis for the model of Cohen et al. (1990), supported an assumption of WEAVER++, and confirmed a critical prediction of the model.

The functional architecture of WEAVER++ has also successfully been used in analyses of data on word production from neuroimaging and electrophysiological studies. For example, Indefrey and Levelt (2000) used the functional architecture in a meta-analysis of 58 brain imaging studies on word production in the literature. The studies included picture naming, verb generation (generating a use for a noun, e.g., saying "hit" to HAMMER), word reading, and pseudoword reading. The lower panel of Figure 19.4 relates the word planning stages to areas of the human brain. Moreover, WEAVER++ successfully simulated data from functional magnetic resonance imaging (fMRI) studies, in particular, the fMRI BOLD (Blood Oxygen Level Dependent) response in different subregions within Wernicke's area during speech production and comprehension tasks. Whereas left perisylvian areas, including the areas of Broca and Wernicke, map colors and words onto the corresponding articulatory programs, the anterior cingulate cortex (on the medial surface of the human brain) and the dorsolateral prefrontal cortex subserve attentional control. The upper panel of Figure 19.4 relates attentional control processes to areas of the human brain. Evidence suggests that the dorsolateral prefrontal cortex serves to maintain the goals in working memory. WEAVER++ instantiates the view that the anterior cingulate achieves input- and output control. WEAVER++ successfully simulated the fMRI BOLD response in the anterior cingulate during Stroop task performance (Roelofs & Hagoort, 2002).

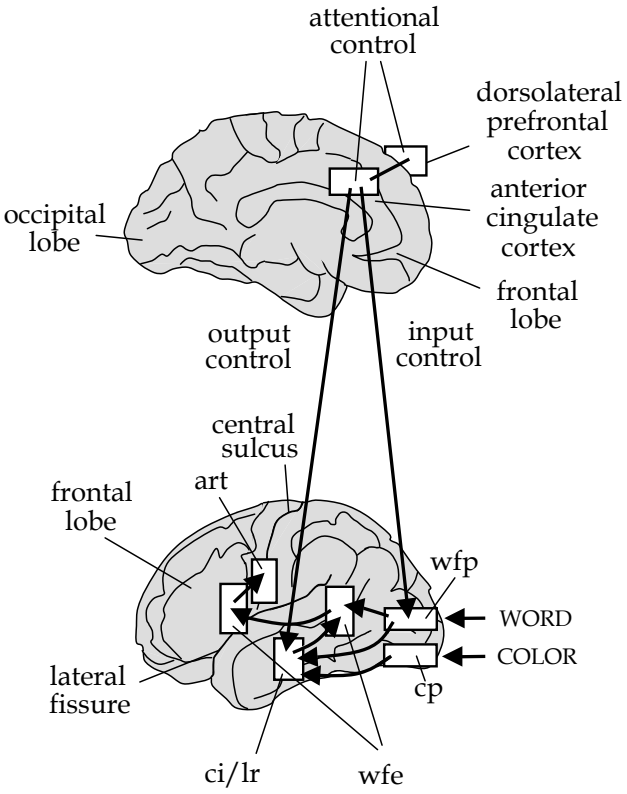

FIG. 19.4 The neural correlates of word planning and attentional control in the Stroop task. Medial view (upper panel) and lateral view (lower panel) of the left hemisphere of the human brain. The word planning system achieves color naming through color perception (cp), conceptual identification (ci), lemma retrieval (lr), word-form encoding (wfe), and articulatory processing (art); word-form perception (wfp) activates lemmas and word forms in parallel. Word reading minimally involves word-form perception (wfp), word-form encoding (wfe), and articulatory processing (art). The attentional control system achieves output and input control.

How does WEAVER++ simulate data? In all simulations, WEAVER++ ran through time in discrete steps, each of which was assumed to correspond to 25 milliseconds in real time. On every time step, activation spread from node to node in the network and the rules tested their conditions or they performed an action. I go through a simulated color-word Stroop trial to illustrate this. Assume that the color has to be named of a red color patch on which the word BLUE is superimposed, whereby the word is presented 100 milliseconds before the color patch (the stimulus onset asynchrony or SOA is –100 milliseconds). The simulation starts with the lemma node of *blue* receiving external activation. Activation then spreads through the network, with the lemma node of *blue* sending a proportion of its activation to the concept node BLUE(X). This node in its turn sends activation to other concept nodes. After the number of time steps that is the equivalent of 100 milliseconds (the SOA), the concept node RED(X) receives external input from the color patch. On the next time step, the production rule for the selection of RED(X) fires and RED(X) becomes flagged as goal concept. Simultaneously, activation spreads from RED(X) to *red*. After the selection threshold of the lemma of *red* is exceeded (i.e., *red* should be more active than *blue* by a certain amount), the production rule for the selection of *red* fires. Although the selection threshold has been reached for the lemma of *blue* earlier because of the preexposure of the word BLUE, the production rule for *blue* did not fire because BLUE(X) was not flagged as the goal concept.

By following this simulation procedure, lemma retrieval times for different experimental conditions may be obtained. Assume it takes 7 time steps in the model (which would map onto 175 milliseconds real time) to retrieve the lemma of *red* in naming a red patch with BLUE superimposed. This retrieval time may then be compared with the time it takes to retrieve lemmas for other stimuli, such as a red patch without a word superimposed. Assume it takes 5 time steps (i.e., 125 milliseconds) to retrieve the lemma *red* for this stimulus. The simulated Stroop interference effect would then be 2 time steps or 50 milliseconds. By comparing simulated and observed effects, the fit between model and data may be determined. Glaser and Glaser (1982) observed 45 milliseconds Stroop interference for this particular situation, so the simulated effect would be in close agreement with the real observation.

SUMMARY AND CONCLUSIONS

I made a case for cumulative computational modeling and testing. This involves working with a single model that accounts for a wide range of existing data and that is extended and tested on new data sets. I first pit-

ted cumulative modeling against two popular methods in psycholinguistics that are not cumulative, namely the toothbrush and skeet shooting approaches. Next, I described the cumulative approach in which models are treated like graduate students. Finally, I demonstrated the productivity of the cumulative approach by describing the scientific career of my own model graduate student WEAVER++. Cumulative modeling does not guarantee success, but it is also not a blind alley, unlike the other approaches. The basic problem with the other approaches is that they do not commit themselves to a strategy of continual approximation. Once started, they do not take any further steps. However, if it took Edison more than two thousand cumulative steps to see the light, we cannot expect to arrive any quicker at a comprehensive understanding of how language works.

REFERENCES

Borges, J. L. (1985). On rigor in science. In J. L. Borges, *Dreamtigers* (p. 90). Austin, Texas: University of Texas Press.

Caramazza, A., & Costa, A. (2000). The semantic interference effect in the picture-word interference paradigm: Does the response set matter? *Cognition*, *75*, B51-B64.

Cohen, J., Dunbar, K., & McClelland, J. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332-361.

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umiltà & M. Moscovitch (Eds.), *Conscious and nonconscious information processing: Attention and Performance XV* (pp. 453-476). Cambridge, MA: MIT Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204-256.

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 318-344.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801-838.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 875-894.

Indefrey, P., & Levelt, W. J. M. (2000). The neural correlates of language production. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 845-865). Cambridge, MA: MIT Press.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and*

Performance, *20*, 1311-1334.
Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91-196). Cambridge University Press: Cambridge, UK.
Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-38.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163-203.
Miozzo, M., & Caramazza, A. (2003). When more is less: A counterintuitive effect of distractor frequency in the picture-word interference paradigm. *Journal of Experimental Psychology: General*, *132*, 228-252.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwarts, & D. Shapiro (Eds.), *Consciousness and self-regulation. Advances in research and theory, Vol. 4* (pp. 1-18). New York: Plenum Press.
Popper, K. (1959). *The logic of scientific discovery*. Basic Books: New York
Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107-142.
Roelofs, A. (1996). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language*, *35*, 854-876.
Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249-284.
Roelofs, A. (2002). Set size and repetition matter: Comment on Caramazza and Costa (2000). *Cognition*, *80*, 283-290.
Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, *110*, 88-125.
Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). *Psychological Review*, *111*, 561-572.
Roelofs, A., & Hagoort, P. (2002). Control of language use: Cognitive modeling of the hemodynamics of Stroop task performance. *Cognitive Brain Research*, *15*, 85-97.
Starreveld, P. A., & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 896-918.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.