CHAPTER

4

Features of Response Times: Identification of Cognitive Mechanisms through Mathematical Modeling

Daniel Algom, Ami Eidels, Robert X. D. Hawkins, Brett Jefferson, and James T. Townsend

Abstract

Psychology is one of the most recent sciences to issue from the mother-tree of philosophy. One of the greatest challenges is that of formulating theories and methodologies that move the field toward theoretical structures that are not only *sufficient* to explain and predict phenomena but, in some vital sense, *necessary* for those purposes. Mathematical modeling is perhaps the most promising general strategy, but even under that aegis, the physical sciences have labored toward that end. The present chapter begins by outlining the roots of our approach in 19th century physics, physiology, and psychology. Then, we witness the renaissance of goals in the 1960s, which were envisioned but not usually realizable in 19th century science and methodology. It could be contended that it is impossible to know the full story of what can be learned through scientific method in the absence of what cannot be known. This precept brings us into the slough of model mimicry, wherein even diametrically opposed physical or psychological concepts can be mathematically equivalent *within specified observational theatres*! Discussion of examples from close to half a century of research illustrate what we conceive of as unfortunate missteps from the psychological literature as well as what has been learned through careful application of the attendant principles. We conclude with a statement concerning ongoing expansion of our body of approaches and what we might expect in the future.

Key Words: parallel processing, serial processing, mimicking, capacity, response times, stochastic processes, visual search, redundant targets, history of response time measurement

From Past to Future: Main Currents in the Evolution of Reaction Time as a Tool in the Study of Human Information Processing

If time has a history (Hawking, 1988), the timing of mental events certainly does. The idea that human sensations, feelings, or thoughts occur in real time seemed preposterous less than two centuries ago. When the idea has finally gained traction, its gradual acceptance in psychology has often been accompanied by much rancor that continued well beyond the development of the first attempts at measurement. After some early progress that had been made in harnessing latency or reaction time (RT) to the study of psychological processes, Titchener (1905, p. 363) was still pondering whether "we have any right to speak of the 'duration' of mental processes." Putting the term *duration* in inverted commas indicates the recent origin of usage of the term as well as Titchener's own doubts about its validity or serviceability.

Thirty years later, Robert Sessions Woodworth in his celebrated Experimental Psychology argued against acceptance of the first method to use reaction time. In a section poignantly titled, "Discarding the subtraction method" (Woodworth 1938, p. 309), Woodworth expressed broader and deep seated reservations, observing that because "we cannot break up the reaction into successive acts and obtain the time for each act, of what use is reaction time?" Even more recent is Johnson's (1955, p. 5) assertion that, "The reaction-time experiment suggests a method for the analysis of mental processes that turned out to be unworkable."

An onerous history granted, the use of RT is firmly established in modern cognitive psychology not least due to the general conceptual framework provided by the domain known as the information-processing approach. Within this framework, RT is used in a systematic, theoretically guided fashion in the quest to isolate the underlying processes and their interactions activated by a given experimental task (cf. Laming 1968; Luce 1986; Townsend and Ashby 1983; Welford 1980).

Nevertheless, we would be remiss if we did not examine, if in passing, the essence of Woodworth's reasoning. Woodworth's concerns hark back to the forceful argument on the continuity of consciousness offered by William James in his seminal *Principles of Psychology* (see in particular, James 1890, Vol. 1, p. 244). In the chapter on the stream of thought, James contends that, due to its absolute continuity, thought or consciousness cannot be divided up for analysis. His attack is directed against the possibility of introspecting minute mental experiences, but the objection is equally cogent with respect to RT. When obtaining a value of RT, one measures the duration between two markers in time, usually that between some specified signal and the observer's response. The RT is taken then to represent the time consumed by an internal process needed to perform a mental task. However, if mental processes are not amenable to partition, any pair of markers must be considered arbitrary. On a deeper level, the situation is a replica or subspecies of the relationship between nature and language as discussed by Friedrich Nietzsche (1873). Nature might well comprise a continuous whole, but human language (used to describe nature) is always discrete. How does one treat a continuous variable with discrete tools? Without dwelling on this issue in any depth, the upshot is clear. A fundamental, yet heretofore unarticulated assumption underlying all RT-based models, serial or parallel, is this: Natural mental functioning can be divided into separate, psychologically meaningful acts.

Returning to history, why did the idea that mental acts occur in real, hence measurable, time seem so incredible less than 200 years ago? The physiology of the human nervous system had made startling advances just around that time, but for many centuries the main thrust of attempts to understand the system along with the attendant sensations fell under the rubric of "vitalism." Vitalism is the doctrine that there is a fundamental difference between living organisms and nonliving matter because the former entail something that is missing from the latter. Pinpointing just what this "something" was has proved elusive, yet the doctrine enjoyed widespread influence from antiquity (the Greek anatomist Galen held that vital spirits are necessary for life) to the 19th century (for all his great contributions to physiology, the towering figure of Johannes Müller subscribed to vitalism) to our own time (Freud's "psychic energy," "emerging property," or even "mind" itself come to mind). Vitalism is best understood as opposition to the Cartesian extension of mechanistic explanations to biology (Bechtel and Richardson 1998; Rakover 2007). It is on this background of the strong influence of vitalism that researchers at the time believed that nerve conduction was instantaneous (in the order of the speed of light or faster) and that, in any rate, it was too fast to be measured.

Hermann von Helmholtz's Measurement of the Speed of the Nerve Impulse

Therefore, Hermann von Helmholtz (1821–1894) along with his fellow students at Johannes Müller's Berlin Institute of Physiology had to summon their best judgment and blood (signing their antivitalism oath) to rebuff their teacher, and espouse a strictly mechanistic position. Under the circumstances, it was a bold move on the part of Helmholtz and his peers to consider the moving nerve impulse as (merely) an event in space-time on a par with, say, that of a moving locomotive. Devising an ingenious method for measuring time, Helmholtz proceeded to measure the speed of the former. He stimulated a motor nerve in a frog's leg and found that the latency of the muscular response depended on the distance of the stimulation from the muscle: the smaller the distance, the faster the response. Helmholtz's calculations showed that the propagation of the impulse down the nerve was surprisingly slow, between 25 to 43 meters a second. Regardless of the value, it became evident that the speed of nerve conduction was finite and measurable! More boldly yet, Helmholtz turned to humans, asking participants to push a button when they felt stimulation in their leg. Predictably enough, people reacted to stimulation in the toe slower than to stimulation in the thigh. Helmholtz estimated the speed of nerve conduction in humans to be between 43 and 150 meters per second. The large range is notable, attesting to considerable variability. It was this variability, within-individuals as well as between-individuals, that discouraged Helmholtz from further pursuing RT research as a reliable means of psychological investigation.

The last point is also notable because individual differences was the subject of a now-famous incident at the Greenwich observatory, which occurred half a century before Helmholtz's measurements. Assistant astronomer David Kinnebrook was relieved of his job by his superior, Nevil Maskelyn, due to disagreement in reading the time that a star crossed the hairline in a telescope. The superior found that his assistant's observations were a fraction of a second longer than his own. Twenty years later, this little-noticed incident (at the time) came to the attention of the German astronomer F. W. Bessel, who started to compare transit times by various astronomers. This first RT study revealed that all astronomers differed in their recordings. In order to cancel out individual variation from the astronomic calculations, Bessel set out to construct "personal equations" as a means to correct or equate differences among observers. Notice that the concept of "personal equation" assumes small (to nil) intra-individual variability in tandem with stable interindividual differences. Neither notion proved to be correct as Helmholtz witnessed with his observers. It turns out that variability, whether of intra- or inter-individual species, is a fixture of RT measurement. It is at this juncture that models developed within the generic framework of human information processing become truly valuable, attempting to disentangle the various sources of RT variability.

Studies of Reaction Time in Wundt's Laboratory: Moving from the Periphery to the Center

Note that for all his pioneering contribution, Helmholtz's measurements were restricted to the periphery of the nervous system, to sensory and motor nerves transmitting impulses toward or from the brain (Fancher 1990). Even this result, as we recounted, was achieved after travelling a torturous road. Nevertheless, barely a decade after Helmholtz's measurements in 1850, the following intriguing question was posed (separately) by Wilhem Wundt (1832–1920) and Franciscus Donders (1818–1889). Could RT measurement be refined to gauge duration of central processes, presumably reflecting mental activity in the brain itself?

Wundt approached the question experimentally by probing the simultaneity of stimulus appearance in the conscious mind. Do stimuli presented at exactly the same (physical) time evoke similarly simultaneous sensations? In a simple experiment performed in his home in 1861, Wundt attached a calibrated scale to the end of the pendulum of his clock so that pendulum's position at any time could be determined with precision. A needle fastened to the pendulum perpendicularly at its middle would strike a bell at the very instant that the pendulum reached a predefined position on the scale. Using this makeshift (yet accurate for the time) instrument (Figure 4.1), Wundt was observing his own mind: Hearing the sound of the bell, Wundt did *not* perceive the pendulum to be in the predetermined position but always away from there. Calculation based on the *perceived* distance of the pendulum from its original position showed the perceived time difference to be at around one-tenth of a second. Inevitably, Wundt concluded, people do not consciously experience the visual and auditory stimuli simultaneously, despite the fact that these stimuli occur at the same time.

Encouraged by such data, Wundt subsequently attempted to measure specific central processes. A favorite topic was "apperception," an early term for what is now known as attention. Wundt found that the RT to a given stimulus was shorter by one-tenth of a second if the observer concentrated on the response rather than on the stimulus. The

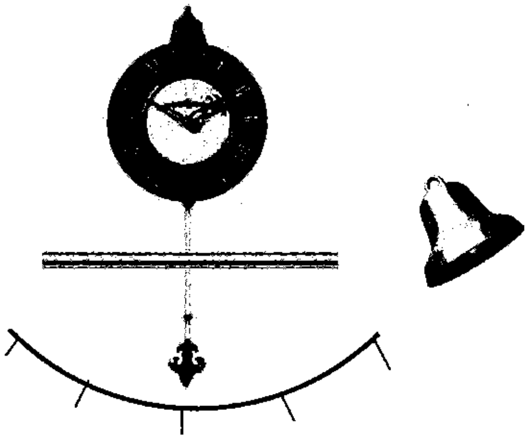

Fig. 4.1 Schematic of Wundt's thought meter.

reason is that one has first to perceive the stimulus and then to apperceive it, that is, to decide whether it is the appropriate one for responding. When focusing on the response, the second of these processes is gratuitous. Consequently, Wundt proposed that apperception takes about one-tenth of a second. Regardless of the particular results, the significance of Wundt's early foray into RT measurement lies in his bold thrust to probe the duration of mental processes of consequence to cognitive science and everyday life alike. Cognizant of its potential, Wundt's home apparatus has been depicted as a "thought meter," and the title of his own report (including subsequent data) aptly read, "*Die Geschwindigkeit des Gedankens*" (The speed of thoughts; Wundt 1892).

Important work in Wundt' laboratory was carried out on a related subject, the number of stimuli noticed simultaneously during a short glance. James McKeen Cattell (1860–1944), Wundt's American student and assistant, first employed RT in the study of the visual *span of attention* or span of apprehension. However, a true pioneer in this domain was the Scottish philosopher, Sir William Hamilton, whose observations are reported in a posthumous book published in 1859. Hamilton spread out marbles on the ground and concluded that, on average, the span of visual attention is limited to 6–7 items. However, if the marbles are arranged in groups (of say two, three, or four marbles a group) the person can comprehend many more marbles because the mind considers each group as a unit.

These results and conclusions anticipated those of George Miller a century later in his famous article on the "magical number seven" and on the effects of "chunking" (Miller, 1956). The power of grouping was expounded by Cattell himself who found that whole words could replace single unrelated letters, leaving invariant the number of units noticed within the span. Modern studies on the span of attention use short exposure times (at around 50 ms) in order to avoid eye movements and counting. As a result, observers actually report the contents of their short-term or "iconic" memory. George Sperling (1960), reviving interest in the subject in his groundbreaking studies on the information contained in brief visual presentations, concluded that the span was much larger than previously thought (in the order of 12–16 letters), but that it was also short lived. The very report by the observer can conceal the true size of the span; larger estimates are found when the deleterious effects of reporting are circumvented.

We surely have come a long way from Hamilton's informal surmises. Nevertheless, his observations brought to the fore the idea of limited capacity (resources or attention) and even the idea of parallel processing. Murray (1988, p. 159), ever the keen reader, concluded that "Hamilton perceived consciousness as a kind of receptacle of limited capacity." Needless to add, capacity and parallel processing are key concepts in the current approach known as human information processing.

Donders' Complication Experiment and Method of Subtraction

We already mentioned Franciscus Donders, the true pioneer of RT measurement in psychology. This Dutch physiologist (founder of modern ophthalmology among sundry achievements) developed the first influential, hence lasting procedure for measuring the duration of specific mental processes. Donders devised an experimental setup known as the *complication experiment* with an assorted method of RT measurement called the *method of subtraction*. The idea was to present tasks of increasing complexity and to subtract then the respective RTs in order to identify the duration of the added processes. The technique is best illustrated by the procedures used by Donders himself (Donders, 1868; we follow Murray's 1988 depiction). In one variation, the *a-method*, a sound such as *ki* is presented by the experimenter and the observer reproduces it orally as quickly as possible (one should note that Donders was the first experimenter to use human [his own] voice in RT studies). The a-task is a simple reaction time experiment, recording the time it takes the observer to react to a predetermined stimulus by a predetermined response. In the *b-method*, one of several sounds is presented on a trial, and the observer repeats the sound as fast as possible. This variation is dubbed choice reaction time: Several different stimuli are presented and the observer responds to each of them differently. In the *c-method*, several sounds are given again, but the observer imitates only one of them and remains silent when the others are presented (this variation is now known as the go/no-go procedure). The differences between the respective RTs reflect the duration of the psychological processes involved. For example, the RT for the b-procedure entails both discrimination (or identification) of the stimulus presented and the selection of the appropriate response, whereas that for the c-procedure entails merely discrimination

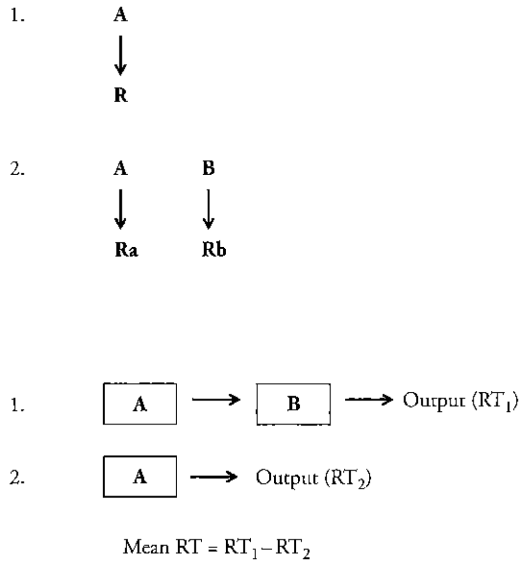

Fig. 4.2 Illustration of the complication experiment and analysis by the method of subtraction. Top: A simple RT experiment (a single predetermined response made to a single predetermined stimulus) is complicated into a choice RT experiment (two different stimuli with a different response made to each). Bottom: The time it takes to perform the mental act of choice is estimated by subtracting the mean RT of the simple RT experiment from the mean RT of the choice RT experiment.

(or recognition, see Luce, 1986, p. 213). The mean difference (c—a) was taken by Donders to measure the duration of recognition, whereas that of (b—c) estimated the time consumed by the need to make a choice between responses (see Figure 4.2 for an outline of the Donders experiment and for the logic of the method of subtraction).

In the scheme developed by Donders, there is a chain of discrete nonoverlapping processing systems. The duration of each process is measurable, assuming that each added experimental task uniquely taps one and only one of the processing systems. If the assumptions hold, the procedure succeeds in inferring the duration and eventually the attendant architecture of the psychological system under test. Consequently, the idea of subtraction has exerted a profound influence on RT theory and experimentation. Townsend and Ashby (1983) paid well-deserved homage to Donders by designating psychological processes carried out in a serial fashion (i.e., sequential and without overlap in processing time) as *Dondersian systems*. This much granted, closer scrutiny of the method (in particular, its underlying assumptions) uncovered several problems, so that the method has not been wholeheartedly accepted by students of RT. The main criticisms are easily summarized because they are interconnected in final analysis.

First, the experimental data collected by different investigators or by an individual investigator at different times proved extremely variable. For example, Donders (1868), Laming (1968), and Snodgrass, Luce, & Gealanter (1967) reported vastly different RTs for (c—a) and (b—c). Over and above the variability, the order of the differences is not preserved: Donders found (b—c) longer than (c—a), but those subsequent investigators found the opposite pattern. Wundt, an early champion of the method, was so discouraged by the large intra-individual variability that he abandoned his RT studies altogether.

Second, the method requires that the added experimental task has no influence on any of the other tasks. The assumption of "pure insertion" (Sternberg, 1969a,b) asserts that the previous processes unfold in time precisely in the same fashion regardless of whether another process is inserted into the chain. If pure insertion is impossible in general or does not hold in particular cases, the assumptions of additivity and independence of the processes are also compromised. To compound the problem, the assumption of pure insertion is untestable with mean statistics although it might be with distributional statistics (Ashby & Townsend, 1980). The issue is not fully settled (cf. Luce, 1986, p. 215), and it is moot whether it can be fully settled with any mathematical or statistical test.

The third criticism is even more fundamental. It concerns the relationship between the experimental task and the unobservable psychological process or subprocesses that the task is supposed to tap. It is not *prima facie* clear that by calling a task "response choice/selection" or "stimulus discrimination" the underlying psychological process is that of choice or discrimination. It is not even clear that the task taps a single process, excluding all sorts of subprocesses.

The *raison d'etre* of the complication experiment is minimum complication, so that a single well-defined process is probed with each addition. This minimal-addition- or single-process principle is not readily testable (certainly not at the level of the mean) and it is even more difficult to satisfy in experimental practice. After all, how can one decide that the added task comprised the smallest complication possible (Külpe, 1895; Sternberg, 1969a,b). Symptoms of the problem have recurrently surfaced in the century following the Donders experiment. Where Donders called a given task "discrimination," Wundt called the same task, "cognition." Donders' c-task was conceived to tap stimulus recognition, but already in 1886 Cattell questioned its validity, arguing that the task entails processes beyond identification or recognition. More recently, Welford (1980), echoing Cattell's concerns, concluded that the difference between the b-task (originally thought to tap response selection) and the c-task is one of degree and that both entail choice of the response. Wundt, acutely aware of the problem, conceived a new task, the d-procedure (meant to be a pure measure of recognition), to no avail. More than linguistic indeterminism is at stake. G. A. Smith (1977), for one, obtained data showing choice to be faster than recognition! How does one make choices among stimuli that one does not recognize? In the absence of a definite task-process association and theory, we cannot know with certainty the identity and order of the pertinent psychological processes. Given the problems, the method of subtraction was out of favor for many years with students of RT.

The succeeding section will bring us into the modern era of cognitive research. Subsequent sections will revisit many of the concepts with more quantitative detail, but still with emphasis on a friendly style.

Saul Sternberg's Revival of the Donders Project: Inaugurating the Modern Study of Human Information Processing

Reminiscent of the tale of Sleeping Beauty, Dondersian procedures were lying dormant for over a century. The prince-investigator reviving the technique was Saul Sternberg (1966, 1969a,b), and the magic kiss awakening renewed interest was his memory scan experiment. The participants are first shown a number of items. Then, they decide whether a test item was or was not present in the set just shown. Prototypical results are given in Figure 4.3. Two features of the data are noteworthy. First, RT is a linear function with a positive slope of the size of the memory set shown. Adding a single member to the memory set increases RT by the same constant amount. Second, targets and foils produce the same increment in RT, so that the slope of the function is the same for yes and for no responses (in Figure 4.3, the intercept, reflecting stimulus encoding, base and residual time, incidentally is also the same; however, the important feature is the parallelism of the target-present and target-absent functions). Sternberg interpreted the linear function with the positive slope to reflect serial processing such that the test item is compared with the memory representation of each of the items in the positive set – one item a time. He interpreted the parallelism of the slopes to mean that the search continues until the entire memory is exhausted even if an early item in the positive set matches the probe stimulus. Sternberg's interpretation of his data is now known as the standard *serial exhaustive* search model. If search ceases as soon as a probe item is located, the process is said to *self-terminate.* Sternberg's original (1966) analyses were stronger than many of the scores of studies that followed, due not only to invoking several control conditions but also in helping to rule out an important class of parallel models. Again, we will discuss this matter as well as other topics in this section in more quantitative detail subsequently.

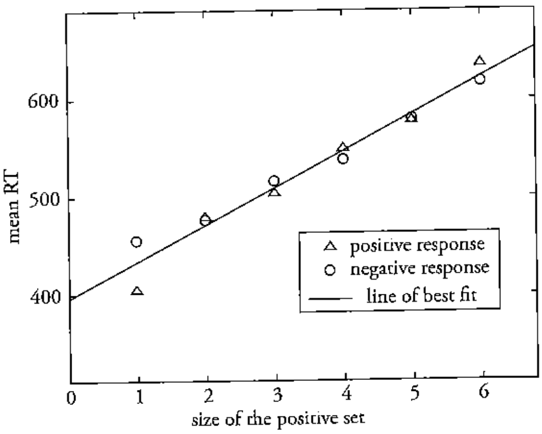

Fig. 4.3 Prototypical results of Sternberg's memory scan experiment.

Sternberg's conclusions seem compelling, but, as subsequent research has revealed, neither is forced by the data. The positive slope appears to have all the earmarks of serial processing, but a moment of reflection suffices to show that the same result follows in a natural fashion from parallel processing. Think of horse races (actual ones, not modeling metaphors) with a different number of horses in each race. The referee reports back to the organizer once each race is over (i.e., when the slowest horse crosses the finish line). Clearly, each race is parallel and exhaustive. It requires only a little intuition to conclude that the larger the number of horses, the longer the expected duration between the common start and the finishing time by the slowest horse (i.e., the RT-set size function has a positive slope). Now, if every horse runs just as fast and with the same random variation no matter how many other horses are present, then it can be shown that the increasing duration for all the horses to finish bends over (i.e., increases by less and less an amount as the number of horses increases; see Townsend & Ashby 1983, p. 92 for a proof) rather than being straight. Such a system, whether run by horses or by parallel perceptual or cognitive channels, is said to be *unlimited capacity* (e.g., Townsend, 1974; Townsend & Ashby 1978). Sternberg's (1966) analyses did rule out *this* variety of parallel processing.

Formal models of memory- or perceptual-scanning have introduced the notion of *limited capacity* in performing the comparison process. In Townsend's capacity reallocation model (Townsend, 1969, 1974; Townsend &Ashby, 1983; see also, Atkinson, Holmgren, & Juola, 1969), a finite amount of capacity is redistributed after completing the comparison of each item, the processing itself is always a parallel race between the remaining items. Such limited capacity, parallel exhaustive search models yield precisely the same predictions as Sternberg's original model (e.g., positive parallel slopes for target-present and target-absent processing, absence of a serial position effect, and linear growth of variance with the numbers of items), some of which are not generally confirmed by experimental data. Following Townsend's early development (1969, 1971), several classes of parallel models have been shown to predict Sternberg's results (Corcoran, 1971; Murdock, 1971; Townsend, 1969, 1971a,b, 1972, 1974; Townsend & Ashby, 1983). Moreover, Sternberg's data can be predicted by self-terminating rather than exhaustive search whether in parallel (e.g., Ratcliff, 1978) or even serial (e.g., Theios Smith, Haviland, Traupmann, & Moy 1973) models. The reader should consult Section 5 as well as Van Zandt and Townsend (1993) and Townsend and Colonius (1997) for more details on the topic of testing self-terminating versus exhaustive processing in parallel and serial models.

The interrogation of Sternberg's results entailed also (slight) experimental modifications. For example, the memory set can follow rather than precede the probe stimulus thus initiating what are usually termed *visual search* (or *early target*) experiments. Early examples of these designs are found in the studies by Estes and Taylor (1969), Atkinson et al. (1969), and van der Heijden (1975).

A more consequential manipulation entails the inclusion of more than a single replica of the target stimulus in the search list. RT is found to decrease with the number of redundant targets (e.g., Baddeley & Ecob, 1973; Egeth, 1966; in bimodal perception, see Bernstein, 1970), a result inconsistent with the prediction of the standard serial exhaustive model. Regardless of this particular result (the violation can be dealt with fairly easily by slight modification of the pertinent models), redundant target designs proved a powerful tool in revealing virtually all aspects of human information processing.

Sternberg revived Donders' method of subtraction in a further profound way. In his *method of additive factors* (Sternberg, 1966, 1969a,b), one does not eliminate or bypass a stage (as in the method of subtraction) but rather affects it selectively. Think of the standard memory scan experiment for an illustration. In the additive factors scheme, the operation of comparison comprises a single stage affected by the factor of size of the search set. Suppose that one adds another stage, stimulus encoding, affected by degrading the quality of the visual presentation. The logic of the method is as follows. .Varying the number of stimuli in the search set affects comparison (and response) processes, whereas degrading the quality of the stimuli affects perceptual encoding. Additivity (of the mean RTs) holds if indeed the manipulations influence the respective processes selectively. If one further assumes independence, the incremental effects of added stages should be additive over accumulated RTs, too. The expected result in this two-stage serial model is shown in Figure 4.4. The influence of set size is revealed by the positive slopes of the RT curves and that of visual degradation by the longer RTs. Critically, the two factors do not interact as is evident in the parallelism of the slopes.

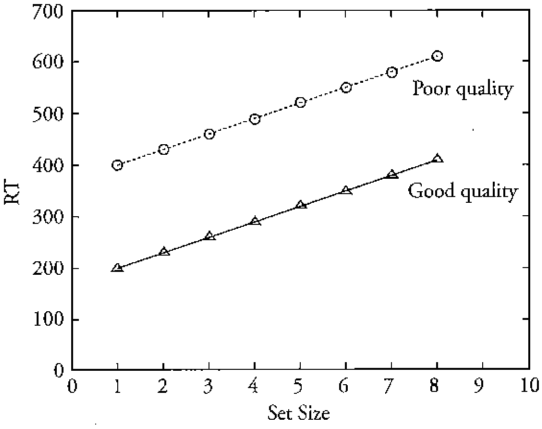

Fig. 4.4 Hypothetical results in an additive factors experiment in which additivity is seen to hold.

The additive factors method, like the memory scan experiment, has engendered a very large amount of research, producing a wealth of valuable theorems (e.g., the independence of additivity and stochastic independence) and theoretical insights (e.g., success and failure in mimicry of serial systems by parallel systems). The last point will be particularly appreciated by those experiencing the frustration in convincing a graduate student (or a seasoned researcher!) that a positive slope does not, ipso facto, imply serial processing (Feature Integration Theory [Treisman & Gelade, 1980] is a poignant case in point). Criticisms and generalizations of the method unearthed further important information. For example, additivity does generally support separate processing stages, but interaction does not necessarily support a single stage. Statistical properties of analysis of variance (ANOVA) might compromise, to an extent, its value as the (sole) diagnostic tool (cf. Townsend, 1984). A really consequential feature of the method in virtually all modifications and generalizations (but see Schweickert, 1982) is that it tells us nothing about the order of occurrence of the various stages (or underlying processes). The ensuing problems were already noticed with respect to the original method by Donders, but they are equally serious with the method of additive factors.

Sternberg's landmark studies, along with the almost concomitant works by Sperling, Estes, Nickerson and Egeth and others, inaugurated the human information-processing approach in earnest. Where Donders, in his subtraction method, changed the nature of the tasks as well as the number of stimuli, Sternberg, in his memory experiment, did not change the task, only added items. It is easier to subtract numerical values of RT than entire psychological processes (cf. Marx & Cronan-Hillix, 1987). In his additive factors method, Sternberg showed that it was not even necessary to subtract processes, only to affect them experimentally in a selective way. Within a decade of Sternberg's seminal contribution, virtually all students of RT and roughly half the community of cognitive psychologists (Lacmann, Lachmann, & Butterfield 1979) were conducting research employing or testing some aspect of Sternberg's theory and methodology.

Basic Issues Expressed Quantitatively

In the previous sections we surveyed some of the history of mental chronometry. Several key issues were highlighted from historical and philosophical perspectives, all related to the notion of time and the role it plays in mental processes. First, mental events—our feelings, thoughts, and decisions—take time, and this time can be measured. Second, internal subprocesses can take place one at a time (and are hence called *serial processes*), or at the same time (*parallel processes*). Third, when several subprocesses take place, the system must await the completion of each and every one of these subprocesses before moving on to respond (*exhaustive processing*), or conversely, it can finish before that, say, upon the termination of any one of the subprocesses (*minimum-time*).[1] Fourth, subprocesses may be *independent* from one another (or not), and so do the time durations taken to complete each subprocess. And finally, we introduced the idea that people may have a *limited capacity*—limited amount of resources (attention)—and hence can deal effectively with a limited amount of processing at any given time. In what follows we provide a formal treatment of each of these basic issues, along with illustrative examples.

The first issue, regarding the temporal modeling of information processing, is ubiquitous in theoretical approaches to human cognition. We see this affirmed in several chapters of this book, such as Chapter 3 (Modeling Simple Decisions and Using a Diffusion Model) and Chapter 6 (A Past, Present, and Future Look at Simple Perceptual Judgment). Many models of perception and decision making are based on the premise that information, or evidence toward some target behavior is accumulated over time. Thus, to answer Titchener's (1905) question, we have both the right and the *obligation* to speak about duration of mental processes.

The remaining basic issues are discussed next in greater detail; the reader may find the following example helpful throughout this discussion. Suppose that you are a driver approaching an intersection. The sight of a red light or the sound of a policeman's whistle signals you to stop and give way. One can think of the visual signal and the auditory signal as being processed in separate subsystems, which we call *channels*. We denote the time to process and detect a signal in each of the channels by t_A (for the visual channel) and t_B (for the auditory channel). We further make the assumption that both signals are presented at exactly the same time (we can relax this assumption subsequently). What can we learn about the time course of information processing? What can we learn about the relationship between the information-processing channels?

The critical properties of architecture, stopping rule, and independence will now be introduced with only little mathematics. A rigorous mathematical statement regarding architecture (i.e., parallel and serial processes) appears in Section 4 of this Chapter. For more quantitative detail on these features, the reader should consult Townsend and Ashby (1983) or Townsend and Wenger (2004b, for a more recent statement).

Architecture: Parallel Versus Serial Processing

As mentioned, two or more subprocesses can take place one at a time (*serial*), or at the same time (*parallel*). Figure 4.5 illustrates these modes of processing, where each arrow corresponds to a particular channel. It is convenient to consider the way the system operates—its *architecture*—through the prism of the time it takes to complete the processing of *both* signals. Suppose that the driver is unwilling to hit the brakes unless both signals are spotted, that is, she processes the two signals exhaustively. In the serial case (Panel a), the time to process both signals is the sum of the durations needed to process each channel, such that total $t_{serial} = t_A + t_B$. In the parallel case, this time equals that needed to process the slower of the two processes, $t_{parallel} = \max(t_A, t_B)$. It is tempting to think that parallel processing will yield a faster braking response compared with serial processing (and more generally that parallel processing is more efficient than serial processing), given that $\max(t_A, t_B) < t_A + t_B$, for any $t_A, t_B > 0$. This intuitive notion is true only as long as we assume that t_A (and similarly t_B) is the same in the serial and the parallel cases.[2]

Is it realistic to expect our driver to bring her car to a stop only after she detects both sources of information? On intuitive grounds, one would prefer to act quickly on the basis of only one signal, whichever signal is detected first as a sign of danger. This issue is considered next.

Exhaustive versus Minimum-Time Stopping Rule

Awaiting the completion of two subprocesses is referred to as *exhaustive processing*. The processing durations, t_{serial} and $t_{parallel}$ for that strategy were given earlier. It is also possible to stop as soon as the first process is completed; in our example, as soon as the driver detects the red light *or* hears the policeman's whistle. This strategy is referred to as *minimum-time processing*. The overall time it takes for a parallel system with a minimum-time rule is given by $t_{parallel} = \min(t_A, t_B)$. For a serial system, the total duration depends on the order of processing, $t_{serial} = t_A$ if A is first, and $t_{serial} = t_B$ if B is processed first. Needless to add, in a serial system that stops as soon as the first channel completes (as soon as the first signal is detected), the second channel will not have a chance to operate at all. Although other stopping rules are also possible, the exhaustive and minimum-time stopping rules are of particular interest. They are illustrated in Figure 4.5 (Panels a–d). Processing times for the different systems are summarized in Table 4.1.

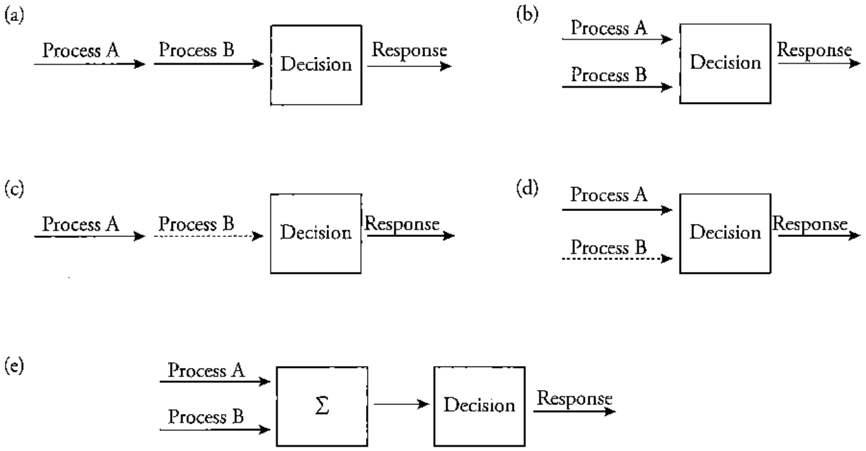

Fig. 4.5 Illustrations of serial (Panels a, c), parallel (b, d), and coactive (e) systems. Panels a and b demonstrate exhaustive processing, where both processes A and B must finish before a decision and response can be made. Panels c and d show minimum-time processing, where processing ceases once process A is completed (but B had not finished, as indicated by the broken line). Panel e illustrates a coactive mode of processing, where activation from two channels is summed before the decision stage.

Table 4.1. Summary of overall completion times for the various models. t_A and t_B denote the time to process signals in channels A and B, respectively.

Model and stopping rule	Overall completion time
Parallel exhaustive	$\max(t_A, t_B)$
Serial exhaustive	$t_A + t_B$
Parallel minimum time	$\min(t_A, t_B)$
Serial minimum time	
if channel A is processed first and B second	t_A
if B is processed first and A second	t_B

Stochastic Independence

Two events are said to be statistically independent if the occurrence of one does not affect the probability of the other. For example, height and SAT score (standardized test score for college admissions in the United States) are independent if knowing the height of a person tells nothing about his SAT score. In the context of processing models, total completion-times of channels A and B are independent if knowing one does not tell us a thing about the value of the other.

Our discussion of the architecture and stopping rule was simplified by the fact we assumed that processing is *deterministic*, rather than *stochastic* (probabilistic). A deterministic process always yields a fixed result, such that the effect or phenomenon we observe has no variability. For example, a deterministic process predicts that the time taken to drive from Sydney to Newcastle is always fixed, or that the time to choose between chocolate and vanilla flavors is the same every time we stop at the ice-cream parlor. Under this assumption, we were able to represent the time for processing in channels A and B by the fixed values, t_A and t_B.

However, observations of human performance (and Sydney's traffic) lead to the conclusion that behavior is quite variable and that it can probably be better described as a stochastic process. If so, processing time in any particular channel can no longer be characterized by a fixed value, but is represented by a *random variable*. A random variable does not have a single, fixed value but can rather take a set of possible values. These values can be characterized by probability distributions. The *probability density function* (pdf) is defined by $f(t) = p(T = t)$, and gives the likelihood that some process, which takes random time T to complete, will actually be finished at time t.

We can use $f(t)$ to define stochastic independence. In probability theory, two random variables are independent if knowing the value of one tells nothing whatsoever about the values of the other (e.g., Luce, 1986, chapter 1). In processing models, total completion times of channels A and B are independent if knowing one, say t_A, tells us nothing about the likelihood of various values of t_B. Thus, we can express independence in terms of the joint pdfs, $f_{AB}(t_A, t_B) = f_A(t_A) \cdot f_B(t_B)$, which means that the joint density of processes A and B both finishing at time t is equal to the product of the probability of A finishing at time t_A and the probability of B finishing at time t_B.

Workload Capacity and the Capacity Coefficient

We recounted earlier that the time to process multiple signals depends on the stopping rule and mode of processing (serial, parallel). Notably, processing also depends on the amount of resources available for processing, a notion that we call *capacity*. One may think of the cognitive system as performing some work, and the more subprocesses (channels) are engaged the greater the amount of work there is to perform. We define *workload capacity* as the fundamental ability of a system to deal with ever heavier task duties (Townsend & Eidels, 2011; see also Townsend & Ashby, 1978, 1983). A ready example is the increase in load from processing one signal to processing two or more signals.

One may find it useful to think about work and capacity in terms of metaphors such as water pipes filling a pool, or tradesmen building a house. Suppose that the tradesmen operate in parallel (and, for illustration, deterministically) and that there is an *infinite* amount of resources (tools, building materials)—*unlimited capacity*. In that case, a twofold increase in the number of workers will cut to half the amount of time needed to build the house (assuming all tradesmen have the same workrate). Critically, adding more workers does not affect the labor rate of each individual worker. In a similar vein, increasing load on the cognitive system by increasing the number of to-be-processed items does not have an effect on the efficiency and time of processing each item alone. The time to process the visual signal (red light) when it is presented alone should be the same as the time to process the same signal when it is presented in tandem with the auditory signal (whistle by the policeman),

$t_{A|A} = t_{A|AB}$. To clarify the notation, the subscript $A|A$ indicates processing of signal A given that only signal A is present, whereas $A|AB$ indicates processing of signal A when A and B are both present. If several channels are working toward the same goal and capacity is unlimited, then adding more channels should facilitate processing.

It is possible however, that capacity is *limited.* In one special case, the overall amount of processing resources, *X*, can be a fixed value. With more and more channels coming into play, fewer resources can be allocated to each channel, and, consequently, the time to complete processing within each channel increases. So, for example, the time to process the visual signal is longer when the auditory signal is also present. Using the same notation as before, we can express this as $t_{A|A} < t_{A|AB}$ and $t_{B|B} < t_{B|AB}$. Under limited capacity, performance with a given target is impaired as more targets are added to the task. Metaphorically, this is tantamount to tradesmen who are trying to work in parallel but share one set of tools. Worker A cannot work at the same rate that she did alone if she needs to await her partner handing over the hammer. Given that multiple workers or channels operate toward the same goal, a limited-capacity system can still complete processing faster than (or at least as fast as) any single channel alone (depending on the severity of the capacity limitation). However, a limited-capacity system cannot be faster than an otherwise identical unlimited-capacity system.

A third and at first curious case is that of *super capacity*. It is possible in principle that as more and more channels are called for action, the system recruits more resources (àla Kahneman, 1973) and is able to allocate to each of the channels more resources than what each channel originally had when it was working alone. In this case, $t_{A|A} > t_{A|AB}$ and $t_{B|B} > t_{B|AB}$, and moreover, the more signals (and channels) there are, the faster the system completes processing. Under supercapacity, performance with a given target is *improved* as more targets are added to the task.

We can model super capacity by way of a system in which channels A and B pool their activation into a single buffer, in which evidence is then compared against a single criterion. In that sense, processing channels can also join efforts to satisfy a common goal as could be the case in the tradesmen example. This mode of processing is often referred to as *coactivation* (e.g., Colonius & Townsend, 1997; Diederich & Colonius, 1991; Miller, 1978, 1982; Schwarz, 1994; Townsend & Nozawa, 1995; Townsend & Eidels, 2011) and is illustrated in Figure 4.5e. Clearly, this type of model benefits from an increase in the number of relevant signals. With auditory and visual signals contributing to a single pool, evidence accumulates more quickly, and will surpass threshold faster. Thus, a coactive model is a natural candidate for supercapacity. However, it is not the only way supercapacity can be achieved in parallel systems as we shall see (Eidels, Houpt, Altieri, Pei, & Townsend 2011; Townsend & Wenger, 2004a).

Townsend and Nozawa (1995) offered a measure of workload capacity known as the *capacity coefficient*:

$$C_{OR}(t) = \frac{\log[S_{AB}(t)]}{\log[S_A(t) \cdot S_B(t)]}. \qquad (1)$$

$S_A(t)$ and $S_B(t)$ are the survivor functions for completion times of processes A and B, and tell us the probability that channels A and B, respectively, did not finished processing by time *t*. $S_{AB}(t)$ is the survivor function for completion times of the system when channels A and B are both at work (e.g., when two targets are being processed simultaneously). We have already defined the pdf, $f(t) = p(T = t)$, as the likelihood that a process that takes random time T to complete will actually be finished at time *t*. We can also define the probability that the process of interest is finished before or at time *t*, known as the *cumulative distribution function* (cdf), $F(t) = p(T \leq t)$. The *survivor function* is the complement of the cdf, $S(t) = 1 - F(t) = p(T > t)$, and tells us the probability that this process had not yet finished by time *t*.

The capacity coefficient, $C_{OR}(t)$, allows to assess performance in a system that processes multiple signals by comparing the amount of work done by the system when it processes two signals with the amount of work it does when each of the signals is presented alone. The subscript OR indicates that processing terminates as soon as subprocess A *or* subprocess B finishes (i.e., minimum-time termination). Townsend and Wenger (2004a) developed a complimentary capacity coefficient for the AND design, where the system can stop only after the two processes, A *and* B, are both finished:

$$C_{AND}(t) = \frac{\log[F_A(t) \cdot F_B(t)]}{\log[F_{AB}(t)]}. \qquad (2)$$

Equations 1 and 2 both apply to two channels, but the C(*t*) index can be easily generalized to account for more than two processes (Blaha & Townsend, 2006). The interpretation of $C_{OR}(t)$

and $C_{AND}(t)$ is the same, so that $C(t)$ refers to both indices. Parallel-independent models are characterized by unlimited capacity, $C(t) = 1$. Capacity is $C(t) < 1$ in a limited capacity model, and it is $C(t) > 1$ with super capacity in force. Architecture (serial, parallel), stopping rule, and potential dependencies can also affect the capacity coefficient. For the effect of architecture, consider a serial model, which processes channel A first and then processes channel B. This model will take more time to complete, on average, than an otherwise identical parallel model in which processes A and B occur simultaneously. The former also results in $C(t) < 1$ – limited capacity. Breakdown of independence across channels also affects $C(t)$ in a predictable manner. Townsend and Wenger (2004a) and Eidels et al. (2011) have shown that positive dependency (one channel "helps" the other) can lead to supercapacity, $C(t) > 1$, whereas negative dependency (one channel inhibits the other) can lead to limited capacity, $C(t) < 1$. The capacity coefficient is discussed further in the later section, Theoretical Distinctions, along with an illustrative example from the Stroop milieu. The interpretation of $C(t)$ is particularly revealing when discussed with respect to the benchmark model that we describe next.

The Benchmark Model: Parallel, Independent, Unlimited Capacity

The standard parallel model can be considered as the "industry's standard" in response-time modeling. This model is characterized by unlimited capacity and independent, parallel processing channels (attributes that yield the acronym UCIP, e.g., Townsend & Honey, 2007). If we further assume that the model can stop as soon as either one of the channels completes processing, we end up with an *independent race model*, illustrated earlier in Figure 4.5(d). Formally, the stochastic version of this model can be written as

$$S_{AB}(t) = S_A(t) \cdot S_B(t). \qquad (3)$$

$S_A(t)$ and $S_B(t)$ are again the survivor functions for completion times of processes A and B and tell us the probability that channels A and B, respectively, did not finish by time t. Consider a model that stops processing as soon as either channel finishes (*minimum-time processing*), but will otherwise not stop as long as process A is still going on *and* process B is still going on (i.e., as long as both processes "survive," hence the term *survivor function*). Because processing-channels A and B are independent, we can multiply the probabilities so that the probability that the entire system does not stop by time t, $S_{AB}(t)$, is given by the product of the probabilities of A and B not finishing (see Eq. 3 again).[3] We note that this equation describes a model with only two channels, but it can be generalized to any number of channels. The probability that an independent race model with n parallel channels does not complete by time t is given by the product of the probabilities of neither channel finishing,

$$S_{mimimum-time}(t) = S_1(t) \cdot S_2(t) \cdot \ldots \cdot S_n(t) = \prod_{i=1}^{n} S_i(t). \qquad (4)$$

Given a parallel model, it is possible that the system stops only when *all* of its channels had completed processing (*exhaustive processing*). In the example, the system will stop only when both channel A and channel B stop. Assuming again that the channels are independent, the probability that the model completes processing by (at or before) time t is equal to the product of the probabilities of channels A and B finishing,

$$F_{AB}(t) = F_A(t) \cdot F_B(t) \qquad (5)$$

and in the more general form, with n channels,

$$F_{exhaustive}(t) = F_1(t) \cdot F_2(t) \cdot \ldots \cdot F_n(t) = \prod_{i=1}^{n} F_i(t) \quad (6)$$

Two well-known RT inequalities also define the benchmark model. Miller (1978, 1982) proposed an upper bound for performance in the OR design ("respond as soon as you detect A or detect B"), the *race model inequality*:

$$F_{AB}(t) \leq F_A(t) + F_B(t). \qquad (7)$$

The inequality states that the cumulative distribution function for double-target displays, $F_{AB}(t)$, cannot exceed the sum of the single-target cumulative distribution functions if processing is an ordinary race between parallel and independent channels. Violations of the inequality imply supercapacity of a rather strong degree (Townsend and Eidels 2011; Townsend and Wenger 2004a).

Grice, Canham, & Gwynee (1984) introduced a bound on limited capacity, often referred to as the *Grice inequality*:

$$F_{AB}(t) \geq MAX[F_A(t), F_B(t)]. \qquad (8)$$

This inequality states that performance on double-target trials, $F_{AB}(t)$, should be *faster* than (or at least as fast as) that in the faster of the single-target channels. If this inequality is violated, the simultaneous processing of two target signals is highly inefficient and the system is very limited capacity. An implication is that there is "no savings" or gains in moving from a single target to multiple targets (in OR designs). In Section 5 we shall demonstrate the use of the three assays of capacity in an OR design – *C(t)* and inequalities (7) and (8). Colonius and Vorberg (1994) proposed upper and lower bounds appropriate for AND tasks ("respond if you detect target A *and* target B"), which are analogous to OR tasks in the sense that their violations indicate supercapacity and limited capacity. Our benchmark model is, therefore, useful in serving as a gold standard against which performance can be compared and interpreted.

Conclusion

Information-processing models can be characterized by the following four features referring to the relations among processing channels: architecture (serial, parallel), stopping rule (minimum-time, exhaustive), capacity (limited, unlimited, super), and stochastic (in)dependence. Most of these properties are latent and cannot be observed directly. Response times are useful tools in uncovering these properties, but in some cases the result is not unique. Model mimicry is thus the focus of the upcoming section. The caveats granted, recent advances in response-time modelling of cognitive processes proved useful in addressing some of the mimicking challenges (allowing researchers to identify critical features of human information-processing). The later section on Theoretical Distinctions outlines some of the advances, followed by applications of novel techniques from empirical literature. The reader might have noticed that some interesting topics such as the stochastic form of serial models were excluded from our discussion due to lack of space.[4] However, the topics included in this chapter should give the reader a good understanding of elementary information-processing theory and a solid preparation for more specialized reading. Box 1 gives a practical illustration of the outstanding issues.

Model Mimicry

Possessing the building blocks (architecture, stopping rule, capacity, and independence), we now can expand purview to establishment of classes of models characterized by those properties. For example, exhaustive stopping rule in a serial model with independent identically distributed processing times, will have a mean response time equal to the sum of the mean response times for each *channel*

$$E[RT] = E[RT_{Channel\ 1}] + E[RT_{Channel\ 2}] + \cdots + E[RT_{Channel\ n}] + nE[T_0],$$

Box 1 Is human capacity limited?

We noted in this section that workload capacity—as measured by the capacity coefficient—could theoretically be limited, unlimited, or super, depending on whether the efficiency of processing decreases, is left unchanged, or increases with additional load (e.g., more signals to process). Cumulative evidence suggests that human capacity is limited (Kahneman, 1973), yet important and frequent situations of modern life, such as driving a car, require simultaneous processing of multiple signals. Therefore, a key question is whether human capacity is, in fact, limited, and what might be the consequences of such limitations in our everyday life.

Strayer and Johnston (2001) studied the effects of mobile-phone conversations on performance in a concurrent (simulated) driving task. They found that conversations with either a hand-held or a hand-free mobile phone while driving resulted in a failure to detect traffic signals and in slower reactions to these signals when they were detected. The findings clearly suggest that human capacity is *limited*. However, in a more recent driving-simulator study Watson and Strayer (2010) have been able to identify a group of individuals—referred to as "supertaskers" who can perform multiple tasks without observed detriments. Although the majority of the participants showed significant performance decrements in the dual-task conditions (compared with a single-task condition of driving without distraction), a small minority of 2.5% showed no performance decrements. These supertaskers can be best characterized as having *unlimited* capacity (and possibly even *super*capacity). The simulated-driving studies by Strayer and colleagues highlight some practical implications of uncovering latent mental constructs (capacity, in this example).

where T_0 is the base time to respond. Thus, for each channel added, we simply add its mean response time for the total average response time. But, is this the only model with such a prediction?

In this section, we provide instances of overlap in predictions that arise from assuming various models. When one model can predict the results of another model, we face an instance of *model mimicry*. Though perhaps an obvious platitude, investigators rarely seem to concern themselves with the specter of mimicry. In this discussion, we emphasize total mimicry, that is, the existence of mathematical functions carrying the structure of one model to another in such a way as to render them completely equivalent. The upshot is that no data expressed at the same level as the mimicking equations can decide between competing models. Mimicry at other levels will be considered as well as some remedies to parallel-serial dilemma (in the following section).

Mean Response Time Predictions

Recall that mean RT has been a useful tool in helping to determine (or eliminate) models best suited for data. Sternberg (1966), discussed in Section 2, supported a positive linear relationship between mean RT and set size. An early extension of this paradigm to conditions where the items were on display (instead of being stored in memory) was carried out by Atkinson et al. (1969) with largely similar results. The evidence for exhaustive processing was supported by the lack of an effect for the serial position of the target in the list. On the other hand, Nickerson (1966) argued that these data could be taken to favor self-terminating processing. In a seminal research with a different type of visual paradigm, same-different matching design with multiple targets, the data were interpreted as supporting a serial self-terminating process (Egeth, 1966). Even within the visual search paradigm, sometimes a self-terminating stopping is found and sometimes an exhaustive stopping is concluded. [See section, Theoretical Distinctions for further discussion of assessing the decisional stopping rule].

However, in none of these pioneering studies was the potential for confounding by other processing characteristics, especially capacity, taken into account. As we recounted, the early standard model was a *serial, exhaustive model with equal mean processing times for every item*. If one additionally assumes that each item or stage possesses the same actual processing distribution (thus producing the equal mean processing times a fortiori) and that they are also independent, then one has the complete*standard serial model* as outlined earlier.

For simplicity, assume that the mean processing time for each of the single items are all equal. Assume further that the target has equal probability of appearing in any of the n positions. On target-present trials, participants process $\frac{n+1}{2}$ items on average (yielding a positive linear relationship between mean RT and set-size). On target-absent trials, participants have to process the entire list, so that the average RT is n times the mean RT for a single item. Therefore, on both target-present and target-absent trials, there is a positive linear relationship between mean RT and set size.

As we alluded in Section 2, it can be shown that *unlimited capacity, independent parallel models* do not generally make this prediction. These models, when using an exhaustive stopping rule, produce logarithmic-like functions that increase with set size, but not in a linear fashion (see Townsend and Ashby 1983, p.92). In the case of minimum time (i.e., race) stopping, they yield curvilinear decreasing mean RT functions. Interestingly, single-target self-terminating processing reveals a flat, straight-line mean RT function for these models. Yet, the linear prediction of the standard Sternberg model is not unique to the serial class of models. Next, we introduce a particular parallel model, where the rate of processing depends on the number of items to be processed, that does yield the linear increase prediction. This model is just one of a multitude of models that can predict the linear relationship found in the data.

Mean response times are a common measure used in determining the processing mechanisms in a task. Although illuminating with respect to the manipulated variables, the model conclusions made from such observations must consider the possibility of mimicry.

Supporting Mathematics: Serial Model

Recall from the previous section, Basic Issues Expressed Quantitatively, that, in a serial model, items are processed one at a time. In minimum-time processing the target may appear in any of the available positions and processing stops when the target is found. As standard practice, $E[I_i]$ denotes the mean processing time of the ith item. Then, using mathematical induction, for target present trials, one has

$$\begin{aligned} &E(\text{Response Time for n positions}) \\ &= \frac{1}{n}E[I_1] + \frac{1}{n}E[I_1 + I_2] + \cdots + \frac{1}{n}E[I_1 + \cdots + I_n] \\ &= \frac{1}{n}E[I_1] + \cdots + \frac{1}{n}\left[E[I_1] + \cdots + E[I_n]\right] \\ &= \frac{1}{n}E[I_1] + \frac{2}{n}E[I_1] + \cdots + \frac{n}{n}E[I_1] \\ &= \frac{(n+1)}{2}E\left[I_1\right] \end{aligned}$$

whereas on target-absent trials the result is simply

$$\begin{aligned} &E(\text{Response Time for } n \text{ items}) \\ &\quad = E[I_1 + \cdots + I_n] = nE[I_1]. \end{aligned}$$

There is, thus, a linear relationship between number of items and mean response times for positive and negative responses. Of course, the minimum time serial prediction is simply *E[I₁]*, a flat straight line.

Supporting Mathematics: Parallel Model

For clarity, a "stage" of processing is the time from one item finishing processing to the next item finishing processing. For example, in an exhaustive model with three items to be processed, any channel will have three stages: the time from start until the first item is processed, the time after the first item is processed to the time the second item is processed, and the time from the second item's completed processing until the remaining item is finished processing. In a parallel model, the distribution of stage processing time takes the form of a difference between item processing times usually conditioned on channel information. Within-stage independence is defined as the statistical independence of stage processing times across two or more channels in the same stage, *j*. Across-stage independence assumes the independence of these times occurs within the same channel, but for different stages. Consider the within-stage independent parallel model with each item having a processing time following an exponential distribution with a rate inversely proportional to the number of items, *n*. In other words, the more items to be processed, the longer the actual processing time of each item will be. Thus, let $g_{a_{ij}} = \exp\left(-\frac{\lambda}{n-j+1}\right)$ be the processing density for the i^{th} item in stage *j* of processing. For example, stage-one processing on all items is $g_{a_1 1} = g_{a_2 1} = \cdots = g_{a_n 1} = \exp\left(-\frac{\lambda}{n}\right)$, whereas stage-two processing has density function

$$g_{a_1 2} = g_{a_2 1} = \cdots = g_{a_{n-1} 1} = \exp\left(-\frac{\lambda}{n-1}\right).$$

We omit the reasoning due to space limitations, but the average processing time for each stage is $\frac{1}{\lambda}$. So, the mean processing time for *n* items is $\frac{n+1}{2}\left(\frac{1}{\lambda}\right)$ (positive response) and $n\left(\frac{1}{\lambda}\right)$ (negative response). So, for $\lambda = \frac{1}{E[I_1]}$ this parallel model gives the same predictions as the aforementioned serial model for mean response times as functions of the number of items.

Intercompletion Time Equivalence

We refer to the time required for a stage of processing as the intercompletion time. So in a serial model, the intercompletion times are just the processing times. We now examine the issue of model mimicry with respect to the distribution of the intercompletion times. We will show cases in which equivalence can occur between two common models, the across-stage independent serial model and a large class of parallel models that assume within-state independence. Across-stage independence is defined as the property that the probability density function of two or more stages of processing is the product of the component single stage density functions. Consider the case in which there are two channels, *a* and *b*, each dedicated to processing a particular item.

To make the equivalence easy to follow, we write the serial model on the left side of the equations and the parallel model on the right. We use *f* for the pdf of the serial model, and *g* for the parallel model. *p* denotes the probability that *a* is processed first in the serial model. $f_{a1}(t_{a1})$is the probability that it takes *a* the exact time of $t_{a1}$to finish in the first stage of processing. *G* is the cumulative distribution function of the respective subscript (for a parallel model). So $\overline{G_{b1}}(t_{a1})$ denotes the probability that the first stage of processing for *b* will fail to finish before the time t_{a1}in a parallel model.

Then, for the independent serial model to mimic the independent parallel model on all response time measurements it is necessary that:

$$\begin{aligned} &pf_{a1}(t_{a1})f_{b2}(t_{b2}|t_{a1}) \\ &\qquad \equiv g_{a1}(t_{a1})\overline{G_{b1}}(t_{a1})g_{b2}(t_{b2}|t_{a1}), \end{aligned} \tag{9}$$

$$\begin{aligned} &(1-p)f_{b1}(t_{b1})f_{a2}(t_{a2}|t_{b1}) \\ &\qquad \equiv g_{b1}(t_{b1})\overline{G_{a1}}(t_{b1})g_{a2}(t_{a2}|t_{b1}), \end{aligned} \tag{10}$$

For mimicry on the level of intercompletion times, we need equivalence for each stage of processing. For example, in the case in which where *a* is processed first (preceding Eq. (9)) one needs to define f and p so that

$$pf_{a1}(t_{a1}) \equiv g_{a1}(t_{a1})\overline{G_{b1}}(t_{a1}).$$

The three "equal" signs simply indicate that this equation must be true for all values of t_{a1}.

This turns out to be readily done. Thus, there is a serial model that can completely mimic response time predictions from any given independent parallel model. This shows us that response time measurements are not enough to prove that there is a unique model for the processes involved in a task. Fortunately, there are distributions for the serial model that make parallel mimicry impossible. The upshot here is that this serial class of models is more general than that of the parallel models—the parallel class is *mathematically* contained within the serial class. This result provides one potential avenue for assessing architecture: Try to determine from the experimental data and appropriate statistics if processing satisfies serial but not parallel processing. If parallel models pass the tests, then these particular tests cannot discriminate (for that task) serial versus parallel architectures.

The Math Beneath the Mimicry

Note that by integrating with respect to t_{b2}, (Eq. 9) reduces to

$$pf_{a1}(t_{a1}) \equiv g_{a1}(t_{a1})\overline{G_{b1}}(t_{a1})$$

$$\text{[FirstStageProcessing]}$$

$$f_{b2}(t_{b2} \mid t_{a1}) \equiv g_{b2}(t_{b2} \mid t_{a1}).$$

$$\text{[SecondStageProcessing]}$$

The same conclusions hold for Eq. (10) by integrating with respect to t_{a2}. This means that if there is intercompletion time equivalence, then there is total model equivalence.

Proposition 1. *Given any within-stage independent parallel model there is* ***always*** *a serial model that is completely equivalent to it.*

Proof. This proof generalizes to cases where there are more than two processing positions (Townsend 1976a). Consider the following within-stage independent parallel model:

$$g_{a1,b2}(t_{a1}, t_{b2}, <a,b>) = g_{a1}(t_{a1})\overline{G_{b1}}(t_{a1})g_{b2}(t_{b2}|t_{a1})$$

and

$$g_{b1,a2}(t_{b1}, t_{a2}, <b,a>) = g_{b1}(t_{b1})\overline{G_{a1}}(t_{b1})g_{a2}(t_{a2}|t_{b1}),$$

where $<a, b>$ denotes that a finishes before b. To show equivalence one needs to define $f_{a1}, f_{b1}, f_{a2}, f_{b2}$, and p for a serial model so that each stage of processing gives equivalent intercompletion time predictions.

As above, for a second stage processing, simply set

$$f_{b2}(t_{b2}|t_{a1}) = g_{b2}(t_{b2}|t_{a1})$$

and

$$f_{a2}(t_{a2}|t_{b1}) = g_{a2}(t_{a2}|t_{b1}).$$

Now we focus on f_{a1} and p. For equivalence, it is sufficient that

$$pf_{a1}(t) = g_{a1}(t)\overline{G_{b1}}(t) \qquad (\star_1)$$

Integrating with respect to t,

$$p = \int_0^\infty g_{a1}(t)\overline{G_{b1}}(t)dt.$$

By dividing by p in the equation $\star_1$,

$$f_{a1}(t) = \frac{g_{a1}(t)\overline{G_{b1}}(t)}{p} = \frac{g_{a1}(t)\overline{G_{b1}}(t)}{\int_0^\infty g_{a1}(t)\overline{G_{b1}}(t)dt}.$$

The remaining density, $f_{b1}(t)$, can be solved in the same way as above, using the equation

$$(1-p)f_{b1}(t) = g_{b1}(t)\overline{G_{a1}}(t) \qquad (\star_2)$$

and the fact that

$$1-p = 1 - \int_0^\infty g_{a1}(t)\overline{G_{b1}}(t)dt = \int_0^\infty g_{b1}(t)\overline{G_{a1}}(t)dt.$$

Thus, the serial model that mimics the parallel is given by:

$$p = \int_0^\infty g_{a1}(t)\overline{G_{b1}}(t)dt$$

$$f_{a1}(t) = \frac{g_{a1}(t)\overline{G_{b1}}(t)}{\int_0^\infty g_{a1}(t)\overline{G_{b1}}(t)dt}$$

$$f_{b1}(t) = \frac{g_{b1}(t)\overline{G_{a1}}(t)}{\int_0^\infty g_{b1}(t)\overline{G_{a1}}(t)dt}$$

$$f_{b2}(t_{b2}|t_{a1}) = g_{b2}(t_{b2}|t_{a1})$$

and

$$f_{b2}(t_{b2}|t_{a1}) = g_{b2}(t_{b2}|t_{a1}).$$

A Simple but Convincing Example

Assume the exponential distribution for each position and channel in both a serial and a parallel model. Assume across-stage independence, too. We follow Townsend and Ashby's (1983) convention and use different parameter notations for serial (u)

and parallel (v) models. We will derive parameter mappings [$p = f(v_{a1}, v_{a2}, v_{b1}, v_{b2})$] that leave the serial and parallel equations equivalent.

Then the density function for the serial model is

$$f_{a1,b2}(t_{a1}, t_{b2}; <a,b>)$$
$$= p u_{a1} \exp(-u_{a1} t_{a1}) u_{b2} \exp(-u_{b2} t_{b2})$$

and the distribution for the parallel model is

$$g_{a1,b2}(t_{a1}, t_{b2}; <a,b>)$$
$$= v_{a1} \exp[-(v_{a1} + v_{b1}) t_{a1}] v_{b2} \exp(-v_{b2} t_{b2}).$$

Step 1: Set $u_{b2} = v_{b2}$ and $u_{b2} = v_{b2}$. Second stage equivalence achieved.

Step 2: Suppose the order is $<a,b>$ for serial processing. Then computing conditional means for the first stage processing,

$$E^s(\mathrm{T}_1 \mid <a,b>) = \frac{1}{u_{a1}},$$

whereas for $<b,a>$,

$$E^s(\mathrm{T}_1 \mid <b,a>) = \frac{1}{u_{b1}}.$$

But, for the parallel process the mean for the first stage of processing is

$$E^p(\mathrm{T}_1 \mid <a,b>) = \frac{1}{v_{a1} + v_{b1}} = E^p(\mathrm{T}_1 \mid <b,a>).$$

So, for equivalence

$$u_{a1} = u_{b1}.$$

Step 3: Now turn to

$$p = [\text{Probability that } a \text{ is first in a serial model}]$$
$$= P^s(<a,b>).$$

Recall that

$$P^p(<a,b>) = \int_0^\infty g_{a1}(t)\overline{G}_{b1}(t)dt$$
$$= \int_0^\infty v_{a1} \exp[-(v_{a1} + v_{b1})t]dt$$
$$= \frac{v_{a1}}{v_{a1} + v_{b1}}.$$

So, set

$$p = \frac{v_{a1}}{v_{a1} + v_{b1}}.$$

In sum, we have guided ourselves to the following propositions.

Proposition 2. *Given an across-stage independent and exponential serial model such that $u_{a1} \neq u_{b1}$, there is no exponential, within-stage independent, and across-stage independent parallel model that is equivalent to it.*

Although a logically sound statement, it may carry too many assumptions on the processing densities to serve for practical application. Below is a more general theorem.

Proposition 3. *Given a serial model, then* ***if*** *there exists a within-stage independent parallel model that is equivalent to it, it can be found by setting*

$$\overline{G}_{a1}(t) = \exp\left[-\int_0^t \frac{p f_{a1}(t')}{p\overline{F}_{a1}(t') + (1-p)\overline{F}_{b1}(t')} dt'\right]$$

$$\overline{G}_{b1}(t) = \exp\left[-\int_0^t \frac{(1-p) f_{b1}(t')}{p\overline{F}_{a1}(t') + (1-p)\overline{F}_{b1}(t')} dt'\right]$$

$$g_{b2}(t_{b2} \mid t_{a1}) = f_{b2}(t_{b2} \mid t_{a1})$$

and

$$g_{a2}(t_{a2} \mid t_{b1}) = f_{a2}(t_{a2} \mid t_{b1}).$$

Conclusions

In the listed two examples, one can see how making assumptions about the model can yield overlapping predictions in response times and the relationship with number of channels (or items). These conclusions obviously sound a warning siren with regard to drawing hasty inferences from the traditional logic concerning behavior of architectures. It would appear that the best that one could achieve would be to posit several classes of models, for example both parallel and serial architectures, which may explain the data together. However, subsequent sections will reveal how our metatheoretical approach can lead to experimental designs that assay such characteristics at a more fundamental level. For even more generality in model mimicry, see Townsend and Ashby (1983, Chapter 14).

Theoretical Distinctions

We now turn our attention toward theoretical conditions and measures under which models are *not* equivalent. Careful probing of these conditions will guide us to experimental designs that can overcome the parallel-serial dilemma and reliably distinguish between information-processing systems in a broad range of empirical settings. Because questions of processing characteristics have been motivated largely by the domain of visual and memory search, they provide the most natural examples. However, the field has recently moved toward more general applicability, and this section appropriately culminates in a discussion of the Double Factorial Paradigm, which allows the experimenter to evaluate architecture, stopping rule, capacity, and stochastic independence in a single block of trials and which can be applied in a multitude of perceptual and cognitive tasks.

Architecture Distinctions based on Generality

SERIAL SYSTEMS GENERALITY BASED ON DEGENERATE MIMICKING

First, we will tie up some loose ends from the previous section by using the results of Proposition 3 to point out an example of a serial model that is unable to be mimicked by a parallel model. When considered alongside the fact that any parallel model has an equivalent serial model (Proposition 2), we arrive at our first fundamental distinction, which we will later refine: *serial models can be more general than parallel models.* Later, we will discover ways in which parallel models can be more general than serial models.

Suppose we start with a serial model and want to check whether it can be mimicked by a parallel model. Recall that the four equations in Proposition 3 determine the form of that parallel model if it exists. Now, if either $\overline{G}_{a1}(t)$ or $\overline{G}_{b1}(t)$ fails to be a true survivor function (for example, if $\lim_{t\to\infty}\overline{G}(t) > 0$), then the corresponding parallel model is deficient, and the serial model cannot be mimicked. A serial exponential model in which the first-stage rates are not equal ($u_{a1} \neq u_{b1}$) is perhaps the simplest case in which the impossibility of mimicry can occur (Townsend, 1972). Following this line of reasoning, Townsend (1976a) derived a set of necessary and sufficient conditions for the existence of well-defined survivor functions.

Given that every quantity in these conditions is in principle derivable from reaction time data, a testable mechanism is provided for rejecting the possibility of parallel mimicry and confirming a true serial architecture. Ross and Anderson (1981) were among the first to apply these conditions to empirical data, testing an assumption of Anderson's Adaptive Character of Thought' (ACT) model that the spread of activation in memory search is parallel and independent. In the process of applying Townsend's conditions, they encountered and addressed several obstacles. The authors had to consider extending the theoretical results to account for the possibility of having a mixture or convolution of different reaction time densities (e.g., in hybrid models) and also find techniques to analyze the tail behavior of empirical distributions. They overcame these obstacles and their data indicated that it could have been produced by a parallel system of the type envisioned by the ACT model. As we have noted, however, the data could not rule out serial processing, it just failed to reject parallel processing.

PARALLEL SYSTEMS GENERALITY BASED ON PARTIAL PROCESSING

So far, we have considered processing models, even dynamic models, as functions of mapping stimuli onto RT and response probabilities. At this point in our development, we must enrich these models with the concept of a *state space.* Simply put, the state space of a dynamic system is the set of values that the system may obtain. A variety of models, including sequential sampling and random-walk models, traditional counting models, and multi-stage activation models (Ratcliff and Smith 2004), suppose that as some cognitive process unfolds dynamically in time, evidence is gathered from each element a_i in the stimulus space, and a threshold c_i must be reached before the element a_i has been completely processed. The state space, therefore, constrains the possible values that the amount of evidence can take at any given time. It turns out that probing the continuum of accumulated evidence reflected in the state space can be very informative in distinguishing between architectures. In fact, it reverses the mimicry-based serial systems generality proven earlier, such that we can use it to reject serial systems and legitimately confirm parallelism.

To formalize this reasoning, we denote the amount of information that has been sampled from the element a_i at some point in time by $\gamma_i(t)$. This function can be either discrete or continuous, as required by the phenomena being modeled. In a traditional discrete-stage model, $\gamma_i(t)$ could represent, say, the number of "features" from the feature set that have been sampled from an alphanumeric character (see Figure 4.6 for an example from visual search). Here, the state space is *finite*, since each character has a finite number of features that can be processed. A Poisson process, on the other hand, operates over a countably infinite state space, such that the possible amount of evidence gathered from each item can be put in correspondence with the integers. Finally, a continuous (uncountably infinite) scale for $\gamma_i(t)$ is familiar from connectionist models of cognition,

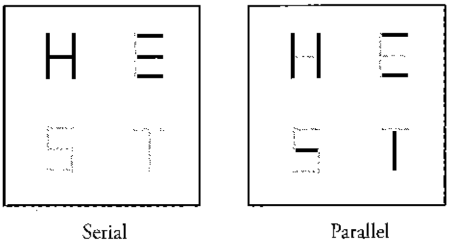

Fig. 4.6 Illustration of the underlying state space in visual search. The two panels represent stimulus displays for which a participant is instructed to "find the S." Individual features of each letter are grey and dotted if they have not yet been processed. If processing is cut off at some point—for instance, if the display terminates—the participant may be left with some letters in a partial state of processing. In the left panel, the letter "E" is in a partial state. Notice that the serial processor must treat the letters one at a time, so there is at most one stimulus in a partial state. In the right panel, on the other hand, all four of the letters are in partial states of processing because the parallel processor has no such restriction.

such as McClelland and Rumelhart's (1981; see also McClelland, Ramelhart, & the PDP Research Group, 1986) interactive activation model, where it is commonly referred to as the *activation value* of a node. This is also generally the case in diffusion processes.

If the state space contains more than one value, then we can, in principle, consider partial processing of elements (in the discrete, feature-based case, this corresponds to some but not all of the features in an object been processed at the end of the trial; see Figure 4.6). This helps us distinguish between architectures in the following way. A serial system can only sample features from a single element a_i at once, and only moves on to begin sampling from another element after completing the first. Thus, there can be at most one element in a partial state of completion (i.e., $0 < \gamma_i(t) < C_i$) in a serial system, whereas a parallel system can have arbitrarily many elements in such a state. One way to distinguish a parallel processor from a serial processor in an empirical setting, therefore, is to observe the underlying state space while more than one item could be in a partial state of completion.

Townsend and Evans (1983) developed a full-report experiment based on this premise. They collected second guesses from participants and examined the pattern of accuracy. Each underlying state (e.g., "item 1 totally processed; item 2 partially processed") maps onto some accuracy pattern (e.g., "item 1 correct and item 2 incorrect on first guess; item 2 correct on second guess"). However, serial models cannot produce underlying states with more than one item partially processed, so the two models predict observably different distributions of accuracies. The authors tested the hypothesis that processing was parallel against a null hypothesis where it was serial. In the statistical analysis, the predictions were passed through two progressively stricter "sieves" and the serial null hypothesis was unable to be rejected. This work was later expanded by Van Zandt (1988) to demonstrate patterns of individual differences in parallel and serial processing. One individual may perform a task in a serial mode, whereas a different individual experiencing identical experimental conditions may perform in a parallel mode.

SERIAL SYSTEMS GENERALITY BASED ON ORDER

The next fundamental distinction is the way in which *order of completion* is selected (Townsend and Ashby 1983, Chapters 3 and 15). As a concrete context for discussion, consider a typical visual search task in which a participant is instructed to decide whether a particular target is present in a display with distractors (e.g., the letter H in an array of other letters). Suppose that this search were carried out serially. It is plausible, then, that the participants could causally direct their attention, choosing at each stage which item will be examined next. They could even decide on some search strategy a priori: "Start with the left-hand stimulus on the top of the display and scan from left to right, top to bottom." In any case, it is apparent that the order of processing is not affected at all by the rates of completion of the various items.

If this task were carried out in parallel, on the other hand, order of completion is entirely determined by the relative rates of different items. If item *a* can be processed faster than item *b*, then the two items will be completed in the order <*a, b*> more frequently than in the order <*b, a*> simply by the stochastic nature of processing, and no a priori decision is able to affect this ordering.

PARALLEL SYSTEMS GENERALITY BASED ON THE IDENTITY OF ITEMS OR CHANNELS

To summarize our conclusions thus far, we have seen that the rate of processing at each stage in both serial and parallel models can depend upon the identity and order of *previously* processed items, but only in serial models can the rates potentially depend upon a predetermined order of *future* items. Parallel models are capable of a different kind of flexibility, however—dependency on the *identity* of items that have started, but not yet finished completion (Townsend 1976b). Consider a visual search with three items, *a*, *b*, and *c*. All items are being processed simultaneously, so if item *c* is particularly inscrutable and takes more effort to process, its presence in the display might slow down the completion time of items *a* and *b*, even if *a* or *b* end up finishing first. This is behavior that cannot be mimicked by a serial model. Note also that not all parallel models necessarily have this property; it is an additional *degree of freedom* we can draw upon when constructing a parallel model to fit observed data. If *a* is the target in a self-terminating UCIP model, for example, processing times would be independent of the identity of *c* by definition. However, a limited capacity parallel model predicts the desired behavior.

PARALLEL-SERIAL TESTING PARADIGM

These simple observations about generality based on order and identity are the foundation of the early *Parallel-Serial Tester* (PST) paradigm (Snodgrass and Townsend 1980; Townsend 1976b; Townsend and Ashby 1983, chapter 13; Townsend and Snodgrass 1974). The PST paradigm is built on three separate conditions of a simple matching experiment, in which a participant must search through a list of two items for targets. In Condition 1, the participant gives Response 1 (R1) if the target item appears on the left of the list and Response 2 (R2) if the target item appears on the right. Condition 2 is a simple AND task, where the target must appear in both positions for response R1, and Condition 3 is a simple OR task, where the target may appear in *either* position for response R1.

Condition 1 is used to get a baseline measure of order effects, which is compared with the cases of Conditions 2 and 3. The processing time of an item under serial processing cannot depend upon the identity of other uncompleted items, so each intercompletion time that we measure empirically must be the processing time of a single item. Intercompletion times under parallel processing face no such constraint. Although the mathematical details and precise predictions for both models are fleshed out in Townsend and Ashby (1983, Chapter 13; see also Townsend 1976b), the basic result for serial systems forces the sum of mean reaction times in the two possible orders of Condition 1 to equal the sum of mean reaction times in the redundant conditions. If the two sums violate this equality, we have empirical evidence for parallel processing.

This paradigm was used successfully by Neufeld and McCarty (1994) to investigate the effect of stress (e.g., periodic high intensity sound) on performance in the three conditions described above, with letters Q, R, T, and Z as stimuli. Contrary to expectations, they found that the presence of a stressor made the system more likely to operate in parallel. In his dissertation, Vollick (1994) applied PST to the clinical setting. It was suggested that the cognitive impairments found in paranoid schizophrenics do not stem from architectural issues, as they process stimuli in parallel the same as healthy individuals, but rather from inefficient deployment of their processing capacity.

Stopping Rule Distinctions based on Set-Size Functions

We take a short break from the serial-parallel dilemma to consider ways to distinguish between exhaustive and self-terminating stopping rules. Since Sternberg's classic work (see Section 2), it has been common to test the stopping rule by examining slope difference in response times to different set-sizes. Consider again the class of standard serial models, where the processing random variable is the same across all experimental variables such as processing position, location in a display, identity and so on. As we recounted, if processing is exhaustive in such models, one predicts that the slope of the lines would be equal, regardless of whether the target is present, every item must be checked. In addition, it is predicted that no display position effects on RT will be found in the data. If the participant is able to terminate the process as soon as the target is found, on the other hand, one predicts the mean RT on positive trials to be lower on average than on negative trials.

Townsend and Ashby (1983, Chapter 7) pointed out that if processing times can vary with display position, processing position, or identity (e.g., regardless of whether an item matches the probe), then exhaustive models could actually violate the above predictions. However, subsequent theoretical effort discovered that exhaustive systems are nonetheless extremely limited in how far they can deviate from those standard serial model predictions. A series of "impossibility" theorems (Townsend & Colonius,1997; Townsend & Van Zandt, 1990) have shown that large classes of exhaustive models *of any architecture* are incapable of producing significant slope differences. Further, they also showed that the ability of such models to evince strong display position effects is severely delimited. Commensurate with the other results of this section, these results offer us mechanisms to reject the possibility of exhaustive processing and confirm self-termination. Van Zandt and Townsend (1993) give a broad review of empirical applications of this test, ultimately concluding that participants employ a self-terminating stopping rule whenever they can properly do so in an overwhelming majority of experiments.

Distinctions based on Reaction Time Distributions

The search for an empirically tractable way to distinguish between underlying cognitive processing systems reached a milestone with the development of an experimental protocol called the Double Factorial Paradigm (DFP; Eidels, Townsend, & Algom 2010; Townsend, & Nozawa, 1995; Wenger & Townsend, 2001), which yields a singularly powerful test not only of the serial-parallel distinction but also of stopping rule and capacity (as well as stochastic independence, though less directly). This paradigm rests within the theoretical framework of system factorial technology, an extensive generalization of Sternberg's additive factors methodology (Ashby & Townsend, 1980; Townsend, 1984; Townsend & Ashby, 1983; see also Dzhafarov 2003; Kujala & Dzhafarov, 2008; Schweickert, 1978; Schweickert, Fisher, & Sung 2012; Schweickert & Townsend, 1989).

DOUBLE FACTORIAL PARADIGM

The DFP entails two concurrent manipulations, each creating a factorial design (hence the *double* factorial paradigm). The first manipulation varies the number of presented targets (workload) in the visual search task posed. This present-absent manipulation is ideal for probing the capacity of the system. The second manipulation (note that the designation "first" and "second" does not imply logical or temporal order) pertains to the salience of the stimulus features. The salience manipulation is ideal for probing the serial-parallel distinction along with further aspects of processing at both the mean and distribution levels.

Consider a Stroop display in which the words RED and GREEN are each presented in red or green ink, and a trial consists of a single word displayed in a single color (Eidels et al., 2010; Melara & Algom, 2003; Stroop 1935). Suppose further that (any kind of) "redness" is defined as the target, so that RED in red ink comprises a redundant target display, RED in green ink and GREEN in red ink comprise one-target displays, and GREEN in green ink is a no-target display. Note that the display can contain two, one, or zero targets, so that the effect of redundant targets is tested, too. The presence-absence factorial design thus created (WORD: target-present [RED], target-absent [GREEN] crossed with ink color: target-present [red], target-absent [green]), depicted at the bottom of Figure 4.7, enables the use of the capacity coefficient (Townsend & Nozawa, 1995) as well as important RT inequalities (Grice, Canham, & Gwyane, 1984; Miller, 1982; Colonius & Vorberg, 1994; see, Luce, 1986; Townsend & Eidels, 2011).

To carry out the salience manipulation in our example, the target word RED can appear in a highly legible or in a poorly legible font and, similarly, the target red ink color can appear in a focal or in an off-focal wavelength. The goal is to *selectively* speed up or slow down the processing of the specific feature, that is, to manipulate one channel without affecting the other. This second factorial design (word-target salience [high, low] X color-target salience [high, low]), depicted at the top of Figure 4.7, turns out to be highly diagnostic with respect to serial-parallel distinctions when paired with the mathematical machinery of systems factorial technology framework described next.

MEAN INTERACTION CONTRAST

Consider the subset of trials in which both targets are presented. Hypothetical reaction time results for the four salience combinations are presented in Figure 4.8. Panel B depicts an additive outcome, implying no interaction across different levels of salience for the two channels. Panels A and C depict the two different species of interactions that might arise: A is overadditive, implying that processing is slow only when *both* channels are slow, and B is the under additive species, implying that processing is fast only when *both* channels are fast. Each factorial plot can be summarized by a simple statistic: Double difference (the difference of the pair of differences between the two values defining each line), or Mean Interaction Contrast (MIC),

$$(\overline{RT}_{LL} - \overline{RT}_{HL}) - (\overline{RT}_{LH} - \overline{RH}_{HH}),$$

where $\overline{RT}$ is the mean RT and L and H denote low and high salience conditions, respectively. Mean interactive contrast is zero for an additive outcome, negative for underadditive interaction,

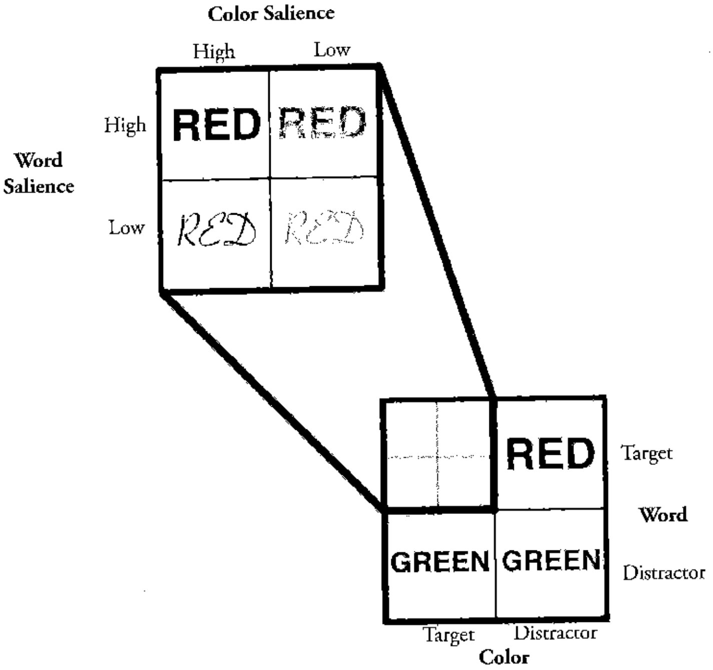

Fig. 4.7 Schematics of the double factorial paradigm (DFP) experiment.

and positive for overadditive interaction. These factorial plots at the level of the mean are already diagnostic with respect to the question of serial versus parallel processing. The diagnosis is qualified by the *stopping rule* in force, indicating whether processing is minimum time self-terminating or exhaustive (Townsend, 1974). We will now derive MIC predictions for four different models summarized in Figure 4.9: parallel self-terminating, serial self-terminating, parallel exhaustive, and serial exhaustive.

First, assume a parallel architecture with a self-terminating stopping rule. If either channel is fed a strong signal, then that channel completes processing quickly, implying that the overall response will be fast. If both channels are fed weak signals though (like a race of two weak and old horses against each other), even the faster one will take a long time. This gives rise to an overadditive interaction: MIC(t) < 0.

If we assume a serial architecture with the same stopping rule, the total processing time is just the average time taken at each stage, assuming different processing orders are equally likely. Thus, when both channels are fed strong signals, the response is fast, and when both channels are low intensity, the response is slow. On mixed intensity trials, the faster channel is processed first half the time and the slower channel is processed first in the remaining half; hence, the overall RT is the mean of these two completion times. This is consistent with an additive outcome: MIC(t) = 0.

In an exhaustive architecture, the system must await processing of both signals (which practically means awaiting completion of processing of the low-intensity signal) before a decision is made. This means that, in a parallel race, the response is fast only when both channels are high intensity, meaning an underadditive interaction: MIC(t) > 0. Applying the same logic to serial processing reveals that the architecture is still additive (although completion times differ from those attained with a self-terminating stopping rule). Put succinctly, MIC is always zero in serial processing, positive in parallel processing with a minimum time stopping rule, and negative in parallel processing within an exhaustive architecture.

SURVIVOR INTERACTION CONTRAST

These distinctions at the level of the mean are useful, but their extension to the distribution level provides further constraints and information. The four RT distributions (i.e., different combinations of target salience) can be sliced into small time bins and a factorial plot, similar to those in Figure 4.8, can be derived at each time bin. The resulting MICs

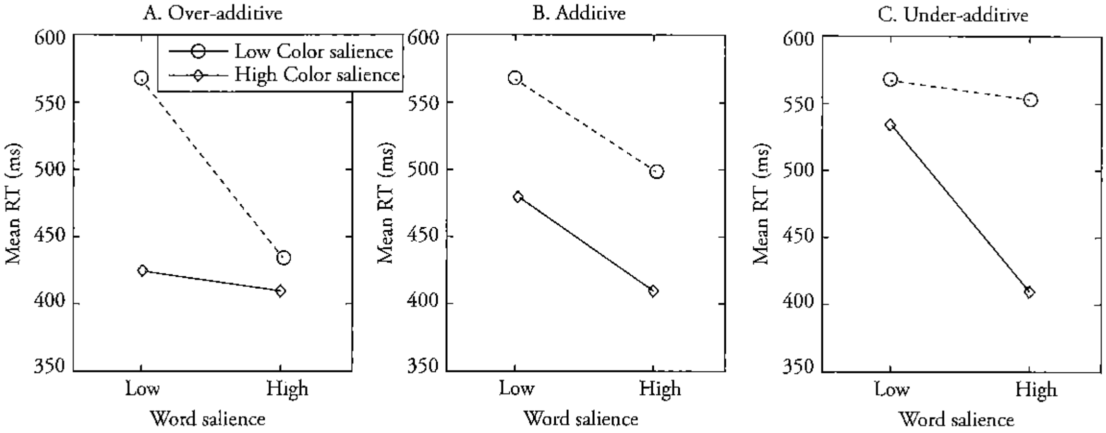

Fig. 4.8 Three different outcomes of the Mean Interaction Contrast in a Stroop experiment.

can then be plotted as a function of time, *t*, for the entire distribution. The mathematical predictions become simpler when applying the momentary values of the respective survivor functions, *S(t)*, (rather than the MICs or CDFs), and the resulting curve is dubbed the Survivor Interaction Contrast (SIC; Townsend and Nozawa 1995). The SIC permits a distinction between different species of parallel models (e.g., race versus co-activation) and Townsend and Nozawa (1995) derived fully diagnostic functions for various combinations of architecture and stopping rule (summarized in Figure 4.9).

SELECTIVE INFLUENCE

Analysis at the distribution level can also provide supporting evidence on *selective influence*, a critical stipulation for derivations based on the factorial manipulations. Selective influence means that a given factor or manipulation affects only the intended process or channel. A distinct ordering of the survivor functions (for the four RT distributions) is predicted with the following condition in force:

$$S(\text{LOW},\text{LOW}) > S(\text{LOW},\text{HIGH}),$$
$$S(\text{HIGH},\text{LOW}) > S(\text{HIGH},\text{HIGH})$$

at all time *t*. Recall that the first factor is the salience of the word and the second factor is the salience of the color. Violation of this ordering compromises interpretations based on the interaction contrasts. One should note though that the presence of the predicted ordering does not prove the assumption of selective influence. It is a necessary but not sufficient condition (i.e., the same ordering could have been obtained even if selective influence were violated).

WORKLOAD CAPACITY

Finally, consider briefly the other leg of the DFP, the factorial manipulation on the number of targets presented. This design is well suited to probe the sensitivity of the system to changes in workload. Recall the definition of the capacity coefficient that we gave in the section, Basic Issues Expressed Quantitatively, keeping in mind that the log survivor function is identical to the integrated hazard function, H(*t*). The numerator in our Stroop task example comprises the redundant-target trials in which the word RED is written in red ink. In the denominator, we have the same functions estimated from trials in which each channel appears in isolation. So:

$$C_{OR}(t) = \frac{H(\text{RED in red})(t)}{H(\text{RED in green})(t) + H(\text{GREEN in red})(t)}$$

Conceptually, this can be thought of as measuring the processing relationship between the "whole" in the numerator and the "sum of its parts" in the denominator.

We can think of the capacity coefficient as measuring the workload capacity relative to the benchmark UCIP model. Reiterating the predictions from the section Basic Issues Expressed Quantitatively, when channels are independent and parallel (as in the standard UCIP model), then the ratio is 1 and the system is unlimited capacity. When presenting two (multiple) targets concurrently slows the rate of processing, $C(t) < 1$, the system is limited capacity, which is less efficient than UCIP. Furthermore, if the system is limited to the extent

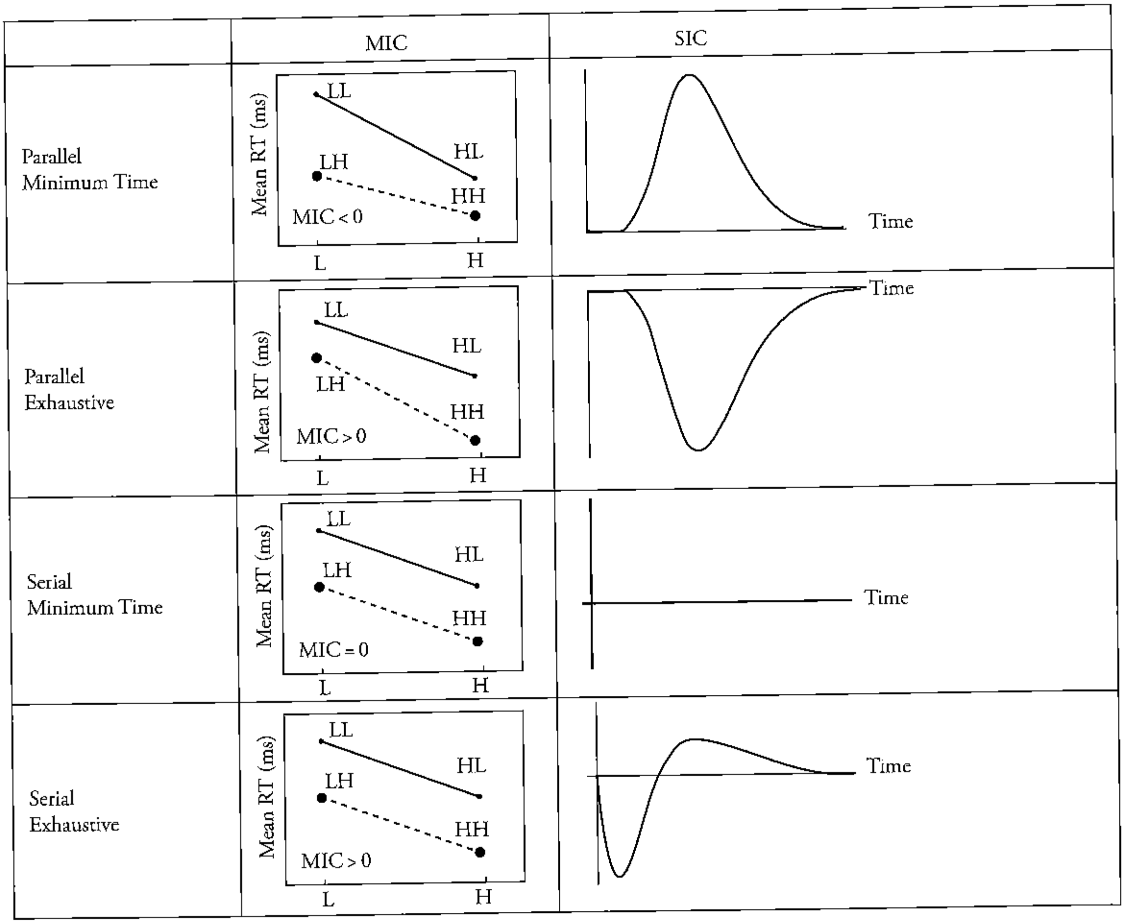

Fig. 4.9 Each combination of architecture and stopping rule has an MIC and SIC signature, as described in the text. See Townsend and Nozawa (1995) for derivations.

that performance with double targets is worse than with the faster of the two individual targets, or in terms of the capacity coefficient, then

$$C_{OR}(t) \leq \frac{\log\{MIN[S_A(t), S_B(t)]\}}{\log[S_A(t) \cdot S_B(t)]}$$

and the Grice inequality is violated (Townsend & Eidels, 2011). In this case, the system is of severely limited capacity.

Conversely, if the presence of two targets speeds up the ability of the system to process each target, $C(t) > 1$, the system is supercapacity, which is more efficient than UCIP. Note that, when (target) signals are presented simultaneously to two channels, detection is usually faster than when a single signal is presented in one channel. This redundant target effect can derive from mere statistical facilitation (the minimum of two random variables has a smaller mean even than that of the faster of the individual random variables alone). However, if the channels interact, a larger redundant target effect can be expected. If the former is the case, the Miller- or race-model inequality (see the section Basic Issues Expressed Quantitatively) must hold:

$$F(\text{RED in red})(t) \leq F(\text{RED in green})(t) + F(\text{GREEN in red})(t),$$

where F denotes the RT CDFs for the respective redundant target and the two single target conditions. When this race-model inequality is violated, traditional thinking has been that all parallel race models are thereby falsified (also observe that satisfying it does not mean that the model is necessarily parallel). However, it is straightforward to construct parallel race models exhibiting super workload capacity that readily violate Miller's race model bound (for examples, see Townsend & Wenger, 2004a). Such models can be created through mutual channel facilitation. It has been suggested that configural perception (e.g., words, faces, other Gestalts) may be explained by such parallel channel facilitation (e.g., Eidels, Townsend, & Pomerantz, 2008; Wenger & Townsend, 2001; 2006; but see Eidels et al., 2010 and Eidels, 2012).

In their original development, Townsend and Nozawa (1995) showed that the Miller and Grice inequalities can be recast as statements about the capacity of the system. When both hold, the system capacity falls between those extremes and, thus, could be of moderately limited capacity, unlimited capacity, or moderately supercapacity. If Grice inequality is violated, the system is of severely limited capacity. If Miller's inequality is violated, the system is of supercapacity (at that t). In a related development, Colonius (1990) has shown earlier that if the marginal probabilities F(Red in red)(t) and F(Green in red)(t) are invariant from the single target to the redundant target conditions, then the Miller and Grice inequalities correspond to maximum negative and positive dependence between parallel channels (see also Van Zandt, 2002). Subsequently, Townsend and Wenger (2004) showed that, interestingly, actual dynamic parallel systems whose channels interact by assisting one another (e.g., an increase in information in one channel leads to an increase in other channels) typically don't produce invariant probabilities and do produce supercapacity and maybe violate Miller's inequality (in apparent, but not real, contradiction of Colonius' mathematically impeccable theorems). Conversely, typical dynamic parallel systems with mutually inhibiting channels evidence negative correlations, again cause failure of marginal invariance, and effect strong slow-downs of processing and possible violations of the Grice inequality.

Cognitive-Psychological Complementarity

How well or poorly have these developments been incorporated into mainstream cognitive psychology? Without attempting an exhaustive analysis of the vast volume of research in current cognitive science, it is fair to say that the influence on substantive theorizing and experimentation of the mathematical developments has been uneven at best. Although early work on parallel and serial processing by Estes, Murdock, Schweickert, Dzhafarov, Egeth, Bernstein, Biederman, and Townsend (among others) has been well accepted in mainstream cognitive psychology of the time (early 1970s) and had a sobering influence on the field, this work has been subsequently ignored and superseded by the following groundless logic:

> If, in a search task, the mean RT is linearly related to the number of stimuli and this function has a positive slope, then serial processing is implied.

Theories based on this untenable statement engulfed the field to the extent that mathematical proofs and violations have been completely ignored. It is only during the past decade that cognitive psychology has finally overcome this detour from logic and mathematical rigor.

Ignoring Parallel-Serial Mimicry: The Case of a Linear RT-Set Size Function with a Positive Slope

Treisman's celebrated work on feature integration theory (e.g., Treisman & Gelade, 1980; Treisman & Gormican, 1988; Treisman & Schmidt, 1982) can serve as a convenient point of departure. This work suggests that when searching for a target that differs from nontargets in terms of a single conspicuous feature (e.g., color, orientation, or shape), the number of elements in the display matters little (*feature search*). However, when the target is defined in terms of a conjunction of features (such as a red vertical line among red tilted lines and green vertical lines), search time increases linearly with the number of elements in the display (*conjunction search*). In the theory, the main diagnostic tool to tell the two types of search apart is the slope of the respective RT-set size functions. The steep slopes obtained with conjunction targets are interpreted to implicate serial search, whereas the much shallower slopes with the one-dimension feature targets are interpreted to implicate parallel search.

Feature integration theory has had a tremendous impact on attention research – as of the printing of this chapter, those three articles alone combine for close to 10,000 citations in the literature. And, this theory is mentioned as a major accomplishment in Treisman's achievement of the highest scientific honor the United States can offer, the coveted National Medal of Science. The associated burgeoning literature helped to uncover valuable aspects of the cognitive processes engaged when people search for a target (whether in a cluttered computer screen or a crowded airport terminal). A less salutary outcome of this trend has been neglect of the possibility of mimicry. Many investigators have ignored proof that a putatively serial (mean)RT function can be mimicked by a parallel one and vice versa. For a trivial yet telling example, Townsend (1971a) was referenced by Treisman and Gelade (1980) but the citation concerned the makeup of the stimuli, not the assay of parallel versus serial processing! Generations of cognitive psychologists appear to have been rendered oblivious to the developments in mathematical psychology on the importance and (im)possibility of distinguishing between parallel and serial processing based on straight line mean RT functions (cf. Townsend, 1971a; 1990a; Townsend & Wenger, 2004b).

The difference between feature search and conjunction search impacted ensuing research to the extent that quite severe violations of the original pattern of results and conclusions, undermining basic tenets of the theory, were largely overlooked. Consider the study by Pashler (1987) for an instructive example. In relatively small displays of up to eight items, Pashler found the same slope for target-present and target-absent trials. This result is consistent with a serial exhaustive search rather than with the serial self-terminating search suggested by the theory. In a further experiment, Pashler added a second target on some portion of the trials and found a redundancy gain, the mean RT was faster when the display included two targets than when there was a single target. Redundancy gain is incompatible with both exhaustive and self-terminating varieties of serial models (cf. Egeth & Mordkoff, 1991; see also, Egeth, Virzi, & Garbart, 1984).

These findings disconfirm the theory themselves, but the fact remains that they (along with a fair number of similar results) generated little traction at the time. The bright side to the story though is the increased use of the redundant target heuristics. From a modest start in the late 1960s (e.g., Bernstein, Blaken, Randolph, & Hughes, 1968; Egeth, 1966), this type of experimental design evolved into a major tool not only for investigations into visual search but for uncovering aspects of elementary cognitive processes in general. Nevertheless, outside of the work of a few investigators such as H. Egeth and colleagues and our own research, the visual search literature and that focusing on redundancy have unfortunately been largely nonintersecting.

Egeth and Mordkoff (1991) used the redundant target design in tandem with Miller's race model inequality as a further means of theoretical resolution. They concluded that the large violations of the inequality found were incompatible with any species of serial processing (and with certain varieties of parallel processing).

In another interesting interrogation, Pashler and Badgio (1985) included trials in which all items were visually degraded and found the effects of set size and degradation to be additive. The additive pattern clearly refutes models of serial identification. A conceptually similar study (indeed, complementary to that by Pashler & Badgio, 1985; cf. Pashler, 1998) was conducted by Egeth and Dagenbach (1991). In their study, the observers searched two-element displays in which each item could be visually degraded independently of the other element. The authors found a subadditive pattern, confirming again a parallel process of letter identification.

Townsend and Nozawa (1995) investigated the redundant target paradigm along with a range of RT inequalities in a more general context, developing measures for the identification of different cognitive architectures. Survivor function interaction contrasts and processing capacity play key roles in this effort. In particular, Townsend and Nozawa showed that Miller's inequality (among other RT inequalities) is actually a statement about the capacity of the process under test. What these developments demonstrate is the futility of drawing strong conclusions based on any simple RT (detection) function, if for no other reason than the brutal reality that many if not all such functions can be mathematically mimicked (e.g., Dzhafarov, 1993, 1997). A broader angle of attack is needed, one guided by a system of theorems and associated tools (the SFT proved serviceable in that role) within the framework of which absolute or mean RTs or density functions serve as points of departure. In this respect, Townsend and Wenger (2004a) generalized the earlier results to include conjunctive, rather than only disjunctive, decisions and illustrated their findings within the large class of interacting channels, parallel, linear stochastic systems.

A very selective review of the major findings in the field of speeded visual search during the past three decades reveals that a wealth of stimulus properties (spatial distribution of the items, target-distractor similarity, stimulus discriminability or task difficulty, practice, common shape and/or semantic category, and even particular attributes such as form and color) have increasingly replaced the number of stimuli (set size) as the variable of interest. The dichotomy, efficient-inefficient search, has been gradually superseding the dichotomy, parallel-serial processing.

This course comprises a rather mixed bag. On the one hand, it reflects the growing recognition by cognitive scientists of the pertinent mathematical developments. In this respect, it took some 20 years for psychologists to finally conclude that models are needed that move "beyond Treisman's original proposal that conjunction search always operate serially" (Pashler, 1998, p. 143). On the other hand, the same course also reflects a tendency of moving away from the issue of parallel versus serial processing altogether.

This is unfortunate because, for all the difficulties involved, the issue *is* tractable and it is consequential for a wide range of cognitive processes. What we need is a naturally emerging integration of a given RT model and a certain cognitive theory. RT data, especially those embedded within a larger system, provide rich information about cognitive processing. Nevertheless, given the prevalence of mathematical and statistical equivalence, RTs, even when sustained by explicit models, will not always be diagnostic (cf. Van Zandt, 2002). It is at this juncture that the need for substantive theory becomes pellucid. Van Zandt (2002, p. 506) concludes that it is, therefore, "very important that RT analyses be conducted in the context of... mechanistic... explanations of the process under study."

Speeded visual search continues to fascinate investigators because it is such a ubiquitous human activity (from locating your baggage on the conveyor belt in the terminal to picking up your favorite Cheerios in the crowded aisle of the grocery store to finding your article in the list of those appearing in the journal). The theory proposed and periodically revised by Wolfe (1994; Cave & Wolfe, 1990), *Guided Search*, suggests that all kinds of searches (whether feature or conjunction) involve two consecutive stages. The first stage entails the simultaneous activation of all potential target features. Activity in the second stage is guided by the outcome of the first (i.e., the distribution of activations of the various features), testing serially combinations of activated features until one matches the target. The theory entails the notions of parallel and serial processing, but envisages situations in which either one can become gratuitous. Incorporating further flexible features, the theory is able to account reasonably well for a broad range of data.

Another influential approach implicates similarity as the major determinant of search (Duncan & Humphreys, 1989). A little noticed aspect of the original Treisman experiments is that each nontarget shares a feature with the conjunction target (hence is similar to the target), whereas in feature search each nontarget is different from the target. Duncan and Humphreys showed that search is easy for a distinctive target on the background of relatively uniform distractors but it is difficult on the background of highly diverse distractors. More recently, Ben-David and Algom (2009; see also Fific, Townsend, & Eidels, 2008) applied the machinery of SFT to uncover the influence of species of target and distractor similarity and sameness (physical, nominal, semantic) on various aspects of the architecture of the underlying process.

The additive factors method itself has been incorporated into mainstream cognitive research to the extent that, more often than not, Sternberg is no longer even referenced. Of the multitude of studies, the sustained program of research by Besner and his associates (see Risko et al. 2010, for a recent contribution) stands out for the methodic application of the additive factors method to probe reading processes. For example, in the study by Borowsky and Besner (1993), context or meaning was found to interact with word frequency, on the one hand, and with stimulus quality, on the other hand, yet the latter two factors were additive. The pattern of joint effects was accommodated by a multistage activation model. Nonetheless, it might be well worth to employ the kinds of strategies outlined herein to falsify or accommodate the various types of models in a nonparametric fashion.

When discussing models, we should address (but space does not allow us to truly address) the issue of the degree to which processing across different stages is discrete or in cascade. That is, we conceive of processing on different items or subsystems as occurring in a sequential manner but which may overlap in time. These are often referred to as continuous flow systems. Taylor (1976) was one of the first to proceed to a quantitative analysis of such models but others soon followed (e.g., McClelland, 1979; Miller, 1988). Let us just note that McClelland (1979) sanctioned the use of additive factor methodology to identify separable stages of processing, and that, separately, Ashby and Townsend (1980), Ashby (1982) and Roberts and Sternberg (1993), too, have demonstrated that purely cascaded models can produce additive effects on the mean RTs (provided certain boundary conditions are respected; see O'Malley & Besner, 2008). Schweickert and Mounts (1998) studied and made predictions from a quite general class of continuous flow systems. The issues are quite complicated and the interested reader should consult Logan (2002) on the broad distinction between discrete and continuous processing. More general and robust metatheoretical effort is required to experimentally and effectively segregate such systems from ordinary parallel and serial systems.

Extending the Metatheory to Encompass Accuracy

The great bulk of the theory enveloped in this chapter has featured response times. However, certain (included) theory-driven experimental designs were based on accuracy, such as the paradigms utilizing second-guesses. Recall that the second-guess strategy exploited the fundamental ability of parallel systems to represent many objects (items, features, channels) in partial states of completion as opposed to strict serial systems being confined to a single object in a partial state of completion. Compared with factorial strategies, these latter techniques have been woefully underused but, as noted, have seen some renewed activity recently. There are some other directions that should be broached.

Extending Capacity Theory to Include Accuracy: Moving Beyond Simple Speed-Accuracy Tradeoff

The motivation to extend our capacity theory to include accuracy is twofold. Foremost, the more observable variables there are to constrain models and theories, the better. Additionally, measures that can simultaneously gauge and combine speed and accuracy can address important questions that neither alone is able to do. A specific manifestation of the relation of speed to accuracy arose in the 1960s and was called the *speed-accuracy tradeoff* (e.g., Pew, 1969; Pachella 1974; Swanson & Briggs, 1969; Yellott, 1971). The idea here is that one must be wary when observing say, a speed-up in one's data and drawing perceptual or cognitive conclusions. The reason is that the error rate may have increased, perhaps reflecting an alteration of a decision criterion rather than an improvement in cognitive efficiency.

It has been obligatory in cognitive psychology ever since, when either response times or accuracy changes, to check out how the other is varying. For instance, if the experimenter increases workload in a task and errors increase, she/he makes sure that response times stay the same or increase. It is then concluded that there is no speed-accuracy trade-off. This inference is unwarranted. Consider the following possibility: The workload is harder and errors increase, but the participants have nonetheless also increased their decision criterion a modest amount, but not enough to offset the increase in errors. So, there has indeed been a speed-accuracy trade-off in the sense that even higher inaccuracy would have occurred had not the participants altered their criterion. This kind of subtlety requires a quantitative approach to be adjudicated.

Our tactic has been to extend the response-time based workload-capacity function developed earlier (see Basic Issues Expressed Quantitatively and Theoretical Distinctions sections) to include accuracy (Townsend & Altieri, 2012). Detail is ruled out in this chapter, but the basic trick is to work out the predictions for the standard parallel class of models that are themselves enlarged to generate errors. In addition, for most of the speed-accuracy combinations, the value-loaded term *capacity* is inappropriate. For instance, is it higher capacity and, therefore, "better" to be fast and inaccurate or slow and accurate? For such reasons, it was necessary to introduce value-free terminology, in this case, the term *assessment function* called A*(t)*. Then the assessment functions are assembled, as was the traditional statistical function, in a distribution-free and nonparametric manner. A simplified, symbolic formula is

$$A(t) = \frac{P_{\text{obs}}\,(\text{speed is fast and error occurs})}{P_{\text{par}}(\text{speed is fast and error occurs})}$$

where obs=observed from data and par=theoretical prediction on the basis of the standard parallel model.

Furthermore, either the numerator or denominator can be decomposed into separate accuracy and conditional response time elements. Thus, Pobs(speed is fast and error occurs) = Pobs(speed is fast and error occurs) Pobs(error occurs). Then comparison of the observed and predicted quantities can aid in a number of useful theoretical inferences. For instance, initial analysis of an AND condition carried out by Ami Eidels (personal communication; see Townsend & Altieri, 2012) shows that the above $A(t) > 1$ indicating that the observed joint event of error-plus-speed in terms of response times and inaccuracy greatly exceeded that expected from the standard unlimited capacity independent parallel model. Furthermore, error-plus-slow tended to move in the other direction; that is, toward the $A(t) = 1$ line. Next, the correct-plus-speed $A(t) < 1$ by a massive degree, whereas that for correct-plus-slow was a bit higher but still quite low. All this is predicted by parallel models, which are limited capacity.

We next decomposed the statistics as just described and discovered that inaccuracy was considerably greater from that predicted by the standard model. This aspect then contributed to this specific $A(t) > 1$. However, it also transpired that, given that an error occurred, the likelihood of a fast response was also greater than that expected from the standard prediction. In fact, it appeared that overall, speed was greater than expected whether conditioned on correct or incorrect performance although most impressive when incorrect.

Some tentative interpretations of these results, as well as for an OR condition, were offered in Townsend and Altieri (2012). However, it is simply the case that we know very little about how even parameterized models will reflect limited capacity effects when both accuracy as well as response times are analyzed. For instance, suppose, as is usually assumed, that changes in difficulty, when accomplished within trial blocks, indicate only changes in correct and incorrect processing speeds, *not* any change in decision criteria. The truth is, we don't know to what extent an increase in errors, as in the AND data above, exhibit lower or higher (as earlier) speeds than would be predicted by the standard parallel model, all on the bases of only changes in processing rates.

We anticipate that analysis of traditional models of response times and accuracy, such as parallel diffusions and races as well as serial models plus examination of many data sets, will show the way to a considerably enhanced understanding of speed and accuracy. As a single example of this type of progress, we mention a recent expansion of the accuracy-oriented general recognition theory (Ashby & Townsend, 1986) to stochastic parallel systems, which thereby includes response times as well and the patterns of confusion (Townsend, Houpt, & Silbert, 2012).

Model Mimicking in Psychological Science

Psychological, cognitive, and brain sciences are, outside the most ludicrous parody of behaviorism, black box enterprises in which hypotheses about inner workings must be made and tested. The brain sciences are included here due to the brain's amazing complexity. Even if we were without society's ethical and moral strictures (and a good thing we are not), the brain's machinery is so mysterious that even with behavioral and neuroscientific approaches together, we are a little like a bunch of educated squirrels watching people drive, then poking about under the automobile's hood at night and drawing profound squirrel-science inferences about auto design, functioning, and maintenance. Nonetheless, the resources of mathematical modeling, neuroscientific knowledge and techniques, and excellent behavioral and neuropsychological experimental designs offer the best we can hope for.

Before moving on and as stated in our introduction, all the warnings about pitfalls associated with various aspects of mathematical modeling pale in comparison with verbal theorizing. Electrical engineering and computer science have long possessed rigorous quantitative bodies of knowledge; we could call them metatheories, of how to infer the internal mechanisms and dynamics from observable behavior. One of the most elegant of these is that associated with deterministic finite-state automata. Of course, when the number of states becomes infinite or random aspects intrude, things get more complicated. Yet even these accommodate mathematically rigorous and applicable methodologies. Naturally, the obverse side of the coin of "degree of uniqueness" of a prospective description of an observed system is the "class of mimicking models" (using our terminology). Even within the class of finite-state automata, however much data is collected, there will always be an equivalence class of machines able to predict said data. If the average graduate student in psychology were a little better prepared in mathematics, at least one course in such a topic would provide a beneficial and sobering message regarding the challenges facing them in their careers.

The fact that even so diametrically opposed concepts as are embodied in parallel versus serial processing systems can readily be mathematically equivalent within common and popular paradigms should, along with the implicit forewarnings from other sciences, lead an incipient science like psychology to emphasize the study of such challenges in training their students and planning and carrying out their own research programs. Alas, that prescription does not seem likely to eventuate in the foreseeable future. However, mathematical psychology can strive to better train their own people in these matters, and conduct their own research accordingly.

It should be evident that the use of metatheory to experimentally segregate large classes of models (e.g., all parallel models) within a certain domain comes up short with regard to specifying a highly precise and detailed computational account. We have, therefore, proposed that researchers adopt a kind of successive *sieve* approach, where finer-grain models are probed at each step. Thus, after determining, say, serial processing in a memory task, then one might begin to assess certain particular process distributions. This approach is complementary to Platt's *strong inference* tactic, but not the same. Next, it is important to ponder the different echelons at which model and theory mimicking can take place, as we see in the following subsection.

Species of Mimicry

First, there is mathematical equivalence such as we have discovered over many decades (e.g., Townsend 1972, 1976a; Townsend & Ashby, 1983; Townsend & Wenger, 2004b; Williams, Eidels, & Townsend, 2014). Little had been accomplished in terms of model identification in psychology outside of seminal work in mathematical learning theory by Greeno and Steiner (1968). Their rigorous efforts explored identifiability issues in learning and memory theories based on Markov chains. However, within the realm of closed-form proofs of mathematical equivalence, the work of Dzhafarov should be mentioned. Consider the quite general class of all models based on a race of two or more parallel channels. The Grice models (e.g., Grice et al., 1984) are members of this class. They place all the variance in the decision bounds. Other models such as the Smith and Vickers (1988; see also Vickers, 1979) accumulator model or the race models of Townsend and Ashby (1983) or Smith and Van Zandt (2000) place the variance in the state space of the channel activations. Dzhafarov (1993) proved that in the absence of assumption of specific distributions, these classes are mathematically equivalent within the usual experimental designs.

As we have seen, with a little luck and lots of hard work, one may aspire to employ the very metatheory used to demonstrate model mimicking to aid in the design of experiments that test the model classes at deeper levels. Second, there is incomplete but exact mathematical mimicking to consider. For instance, a class of models might mimic a nonequivalent type of model at, say, the level of the first and/or second moments (i.e., mean and variance) but not be completely equivalent. A case in point is the prediction of mean response time additivity by standard serial models. As one could expect, this prediction can be made by a huge class of alternative models, as proven by Townsend (1990b). The constraints put on the mimicking model class are extraordinarily weak. Just including variances helps a lot but, of course, does not totally remediate the problem (e.g., Schneider & Shiffrin, 1977; Townsend & Ashby 1983, chapters 6, 7).

Third, there is mimicking by approximation. Though perhaps not so intriguing as mathematical equivalence, it is more widespread than the latter and at least as underappreciated by the average researcher. Examples of this type of mimicking in the present venue are the abililty of sequential, but not strictly serial, continuous flow dynamic models to predict approximate additivity of mean response times (e.g., McClelland 1979; see also Schweickert et al., 2012, chapter 6).

It is likely that the third type of mimicking is that which threatens the bulk of model testing in the literature. Ordinarily, two, sometimes more, parameterized models are compared in their ability to fit or predict numerical data. Now, if psychology possessed a level of precision of measurement even remotely close to that in, say, physics or many areas of chemistry, then this policy might be quite optimal. Why test entire classes of models, when you can move directly to the precise model, with all its exact formulas and estimate parameters and thus be done?

Unfortunately, the tolerance of psychological data is much too coarse for such a hope to be realized. Some help is afforded by way of comparison of models against one another, rather than simply fitting one's favorite model to the data. However, even here, the possibility exists that one or the other model will simply fail to fit the data, even as well as another model, due only to the specific quantitative formulation of the psychological precepts, rather than the fundamental characteristics of the latter. For instance, consider an investigator who has correctly pinpointed the proper architecture, stopping rule, and so on for a task but failed to employ the valid associated stochastic process. Thus, perhaps the architecture is standard parallel, with each channel being described by a race between the correct and incorrect alternatives, the race distributions being gamma. Our hapless investigator inappropriately has selected Weibull distributions to describe each channel's race. Meanwhile, a theoretical competitor might produce an incorrect specification of the architecture (e.g., serial), but employed a set of stochastic processes that adventitiously provide a superior fit.

Another vital aspect of model testing in cognitive science, has always been the issue of whether one model is merely more complex than another and thus can reach sectors of the data space that are unavailable to its competitors. Attempts to ameliorate this challenge have long depended on the assumption that if two models possess equal numbers of parameters, they must be of equal complexity. This is rarely the case, as has been recognized for some time. Indeed, we have shown that large classes of models of classification (e.g., identification, categorization) can be much more falsifiable than other classes, though they possess a huge number of parameters (Townsend & Landon, 1983). A special case of some interest is the well-known similarity choice model (Luce, 1963), which is much more flexible in its model fitting ability (often referred to as the champion model of human pattern recognition) than competitors, such as the overlap model (Townsend, 1971b) though the number of parameters is identical.

A deep and vital antidote, when it can be brought to bear, is the quantification of model flexibility-to-fit data, through cutting edge theories of complexity (Myung & Pitt, 2009).This approach is able to quantify a model's complexity and compensate for it in model comparisons. The main obstacle here is that so far, only models with certain types of pellucid specification are subject to this analysis, at least in realistic computational terms. Nonetheless, as computers' powers continue to augment, this strain of technology offer hope for the future of complex psychological science.

We have had little to say, beyond mere platitudes, concerning aspirations of merging principled quantitative modeling with cognitive and sensory neuroscience and especially neuroimaging. Some extremely credible senior psychological commentators are frankly skeptical of the contributive power of neuroscience and in particular, neuroimaging, to the cognitive sciences. We urge serious reflection on the observations of William Uttal (e.g., 2001) in this vein. Though we do not fully agree with his final inferences, we are convinced that his astute scrutiny can only serve to improve the science and its associated methodologies.

The past several decades have seen a vibrant growth of modeling extending from basic psychophysics to higher mental processes to complex social phenomena. It seems inevitable that such multifaceted models, even when found to predict or fit data in an accurate manner, will be vulnerable to a broad spectrum of competitive, mimicking alternative conceptions. We believe that fields endeavoring to explain or predict human behavior, including those based on neurophysiology, will progress faster by developing appropriate metatheories of mimicking, and how to best circumvent these experimentally.

Conclusions and the Future

- The resurgence in the 1950s and 1960s of research attempting to identify mental mechanisms and its continuation today prove that cognitive psychology can be cumulative as well as scientific.
- Mathematical models have made substantive ingress to cognitive psychology since 1950 and made many contributions to rigorous theory building and testing.
- Nonetheless, even in the realm of mathematical modeling, mimicry of one model or

Box 2 Model Mimicking in Psychology

From one vantage point, it might seem that model and theory mimicking reside in a fairly small and technical dominion of scientific psychology. Yet, in a broad sense mimicking is ubiquitous. Every time two theoretical explanations are in contention, it is because the data at hand are not decisively in favor of one over the other. That is, mimicking at one or more levels is occurring and it is up to the champions of either approach, or "innocent bystanders", to invent new observations to resolve the issue. Within the realm of verbalized theory, what often happens is that the two theoretical structures evolve, due to ministrations from their advocates in order to conform to the latest data. The end result is that the theories may end up still handling the larger corpus of data, each being significantly more complex (and thus more difficult to falsify), and yet remaining empirically indistinguishable. A famous case in point is the decades-long clash of the more behavioristic theory of Clark Hull versus the more cognitive theory of Edwin Tolman. Though fundamentally distinct in their theoretical foundations, their long-lasting struggle must be said to have eventually faded away inconclusively. Mathematical modeling and the explicit study of model mimicking seems a promising remedy for such ailments.

class of models by other models, poses a formidable challenge to science building in this complex arena.

- Metatheory is, in our approach, a theoretical and quantitative enterprise that attempts to formulate highly general mathematical characterizations of psychological notions in such a way as to point toward development of robust experimental methodologies for systems identification.
- The reviewed body of research on metatheory has led to redoubtable methodologies for assessing various strategic aspects of elementary cognition in a manner that is resistant to model mimicking. Most of the new theory and technology is founded on distribution and parameter-free theorems and methods.
- As of now, the metatheory and associated methodologies are largely segregated into those resting on RTs as observable variables and those relying on patterns of accuracy as observable variables.
- A major goal for the immediate future is to create a unified theory that merges both RTs as well as accuracy. First steps have been taken in that direction.
- Very little is known about the perils from model mimicking to incremental science in more complex spheres of cognitive science. It may be that such theoretical research will be indispensable to future methodologies, if cognitive science is to avoid devolvement into a maundering, largely inconclusive field.

Glossary

Across-stage independence: Assumes the independence of intercompletion times in serial models. It is defined as the property that the probability density function of two or more stages of processing is the product of the component single-stage density functions.

Capacity coefficient: $C_{OR}(t) = \frac{\log[S_{AB}(t)]}{\log[S_A(t)\cdot S_B(t)]}$, is a measure for processing efficiency as workload (number of signals to process) increases. $C(t)=1$ indicates unlimited capacity – performance is identical to that of a benchmark UCIP model (see later). $C(t) < 1$ and $C(t) > 1$ indicate limited and super-capacity, respectively. $C_{OR}(t)$ is appropriate for OR tasks, while a comparable index, $C_{AND}(t)$, exists for the AND case, with a different formula but similar interpretation.

Cumulative distribution function (cdf): $F(t) = p(T \leq t)$, gives the probability that the process of interest is finished before or at time t.

Deterministic process: Always yields a fixed result, such that the effect or phenomenon we observe has no variability.

Exponential distribution: A probability distribution that describes the time between events in a Poisson process. It is very useful in response time modelling, and has the form $f(t)=ve^{-vt}$, where v is the rate parameter. It also has the "memory-less" property, meaning that the likelihood of an event to occur in the next instance of time is independent of how much time had already passed.

Grice inequality: $F_{AB}(t) \geq MAX[F_A(t), F_B(t)]$. This inequality states that performance on double-target trials, $F_{AB}(t)$, should be faster than (or at least as fast as) that in the faster of the single-target channels. If this inequality is violated, the simultaneous processing of two target-signals is highly inefficient and the system is very limited capacity. For instance, if $F_A(t) = F_B(t)$ then $C_{OR}(t) < \frac{1}{2}$. The special case when $C_{OR}(t) = 1/2$ is referred to as fixed capacity.

Intercompletion time: The time required for a stage of processing to be completed. In a serial model, the intercompletion times are just the processing times.

Mean Interaction Contrast (MIC): A test statistic for the interaction between two factors with two levels each, which allows assessment of architecture and stopping rule from mean RTs. Calculated as the difference between differences of mean RTs in a factorial experiment, $\text{MIC} = (\overline{RT}_{LL} - \overline{RT}_{HL}) - (\overline{RT}_{LH} - \overline{RT}_{HH})$, where $\overline{RT}$ is the mean RT and L and H denote low and high salience conditions, respectively. Because all stopping rules for serial models predict that MIC = 0, they cannot be distinguished for serial models by MIC.

Probability density function (pdf): $f(t) = p(T = t)$, gives the likelihood that some process that takes random time T to complete will actually be finished at time t.

Race model (Miller's) inequality: $F_{AB}(t) \leq F_A(t) + F_B(t)$. This inequality states that the cumulative distribution function for double-target displays, $F_{AB}(t)$, cannot exceed the sum of the single-target cumulative distribution functions if processing is a race between parallel channels, with the added constraint that the marginal distributions for A and B do not change from when one of the two channels is

Notes

[1] Of course, in-between cases are often used, for instance, that of the highly popular single-target-among-distractors. In such a design, the processor may cease as soon as the target is located.

[2] If t_A, t_B are independent random variables then the previous statement will hold true. However, if we assign distinct random variables to the actual processing times in parallel and serial systems, then very broad questions can be asked and answered with regard to vital parallel-serial mimicking issues (Townsend & Ashby 1983, chapter 1).

[3] From Eq. 1 we can also derive the pdf, f(t), of an independent race model with two parallel channels (i.e. the probability that channel A or channel B finish at time t), which is $f_{AB}(t) = f_A(t) \cdot S_B(t) + S_A(t) \cdot f_B(t)$.

[4] Stochastic serial models are a bit more complex, since one needs to take into account the order by which process occur (e.g., channel A before B or vice versa). Full treatment is given in Townsend and Ashby (1983).

Glossary

presented to when both are presented. This stipulation is known as context invariance. When the upper bound implied in the inequality is violated, capacity must be super; that is, $C_{OR}(t) > 1$.

"Stage" of processing: The time from one item finishing processing to the next item finishing processing.

Stochastic independence: Two events are independent if the occurrence of one does not affect the probability of the other. This concept can be expressed in terms of the joint pdfs, $f_{AB}(t_A, t_B) = f_A(t_A) \cdot f_B(t_B)$, which means that the joint density of processes A and B both finishing at time t is equal to the product of the probability of A finishing at time t_A and the probability of B finishing at time t_B.

Stochastic process: The events cannot be characterized by fixed values and should be represented by a *random variable*. A random variable does not have a single, fixed value but rather takes a set of possible values, with their likelihood characterized by a probability distribution.

Survivor function: $S(t) = 1 - F(t) = p(T > t)$, This function is the complement of the cdf, and tells us the probability that the process of interest had not yet finished by time t.

Survivor Interaction Contrast [SIC(t)]: Same as MIC but calculated for survivor functions, $S(t)$, rather than mean RT at each time bin of t. $SIC(t) = [S_{LL}(t)\text{-} S_{HL}(t)] - [S_{LH}(t)\text{-} S_{HH}(t)]$. The SIC(t) functions predict distinctive curves for serial and parallel models for various stopping rules.

UCIP model: A processing model characterized by Unlimited Capacity and Independent Parallel processing channels.

Within-stage independence: The statistical independence of intercompletion times across two or more parallel channels in the same stage.

References

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. *Psychological Review, 89,* 599–607.

Ashby, F. G., & Townsend, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology, 21,* 93–123.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Atkinson, R. C., Holmgren, J. E., & Juola, J. F. (1969). Processing time as influenced by the number of elements in a visual display. *Perception&Psychophysics, 6,* 321–326.

Baddeley, A. D., & Ecob, J. R. (1973). Reaction time and short term memory: Implications of repetition effects for the high-speed exhaustive scan hypothesis. *Quarterly Journal of Experimental Psychology, 25,* 229–240.

Bechtel, W., & Richardson, R. C. (1998). Vitalism. In E. Craig (Ed.), *Encyclopedia of philosophy*. London, England: Routledge.

Ben-David, B. M., & Algom, D. (2009). Species of redundancy in visual target detection. *Journal of Experimental Psychology: Human Perception & Performance, 35,* 958–976.

Bernstein, I. H. (1970). Can we see and hear at the same time? Some recent studies of the intersensory facilitation of reaction time. *Acta Psychologica, 33,* 21–35.

Bernstein, I. H., Blake, R., Randolph, R., & Hughes, M. H. (1968). Effects of time and event uncertainty upon sequential information processing. *Perception & Psychophysics, 3,* 177–184.

Blaha, L. M., & Townsend, J. T. (2006, May). Parts to wholes: Configural learning fundamentally changes the visual information processing system. Vision Sciences Society Annual Meeting, Sarasota, Florida.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19,* 813–840.

Cave, K. R., & Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology, 22,* 225–271.

Colonius, H. (1990). Possibly dependent probability summation of reaction time. *Journal ofMathematical Psychology, 34,* 253–275.

Colonius, H., & Townsend, J. T. (1997). Activation-state representation of models for the redundant signals effect. In A. A. J. Marley (ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce*, Mahwah, NJ: Erlbaum.

Colonius, H., & Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology, 38,* 35–58.

Corcoran, D. W. J. (1971). *Pattern recognition.* Middlesex, PA: Penguin.

Diederich, A., & Colonius, H. (1991). A further test of the superposition model for the redundant-signals effect in bimodal detection. *Perception & Psychophysics, 50,* 83–86.

Donders, F. C. (1868). Die Schnelligkeit Psychischer Processe. Archiv fur Anatomie und Physiologie und Wissenschaflitche Medizin, 657–681.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458.

Dzhafarov, E. N. (1993). Grice representability of response time distribution families. *Psychometrika, 58,* 281–314.

Dzhafarov, E. N. (1997). Process representations and decompositions of response times. In A. A. J. Marley (Ed.), *Choice, decision, and measurement*: Essays in honor of R. Duncan Luce (pp. 255–278). Mahwah, NJ: Erlbaum.

Dzhafarov, E. N. (2003). Selective influence through conditional independence. *Psychometrika, 68,* 7–25.

Egeth, H. E. (1966). Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics, 1,* 245–252.

Egeth, H. E., & Dagenbach, D. (1991). Parallel versus serial processing in visual search: Further evidence from subadditive effects of visual quality. *Journal of Experimental Psychology: Human Perception & Performance, 17,* 551–560.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception* & *Performance, 10,* 32–39.

Egeth, H. E., & Mordkoff, J. T. (1991). Redundancy gain revisited: Evidence for parallel processing of separable dimensions. In J. Pomerantz and G. Lockhead (Eds.), *The perception of structure* (pp. 131–143). Washington, DC: APA.

Eidels, A. (2012). Independent race of colour and word can predict the Stroop effect. *Australian Journal of Psychology, 64,* 189–198.

Eidels, A., Townsend, J. T., & Algom, D. (2010). Comparing perception of Stroop stimuli in focused versus divided attention paradigms: Evidence for dramatic processing differences. *Cognition, 114,* 129–150.

Eidels, A., Houpt, J. W., Altieri, N., Pei, L., & Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology, 55,* 176–190.

Eidels, A., Townsend, J. T., & Pomerantz, J. R. (2008). Where similarity beats redundancy: The importance of context, higher order similarity, and response assignment. *Journal of Experimental Psychology: Human Perception & Performance, 34,* 1441–1463.

Estes, W. K., & Taylor, H. A. (1966). Visual detection in relation to display size and redundancy of critical elements. *Perception&Psychophysics, 1,* 9–16.

Fancher, R. E. (1990). *Pioneers of psychology.* New York, NY: Norton.

Fific, M., Townsend, J. T., & Eidels, A. (2008). Studying visual search using systems factorial methodology with target–distractor similarity as the factor. *Perception & Psychophysics, 70,* 583–603.

Greeno, J. G., & Steiner, T. E. (1968). Comments on "Markovian processes with identifiable statesÍGeneral considerations and applications to all-or-none learning. *Psychometrika, 33*(2), 169–172.

Grice, G. R., Canham, L., & Gwynne, J. W. (1984). Absence of a redundant-signals effect in a reaction time task with divided attention. *Attention, Perception, & Psychophysics, 36*(6), 565–570.

Hawking, S. W. (1988). *A brief history of time.* London, England: Bantam.

James, W. (1890). *Principles of psychology,* 2 vols. New York, NY: Dove.

Johnson, D. M. (1955). *The psychology of thought and judgment.* New York, NY: Harper.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kujala, J. V., & Dzhafarov, E. N. (2008). Testing for selectivity in the dependence of random variables on external factors. *Journal of Mathematical Psychology, 52*(2), 128–144.

Külpe, O. (1895). Grundriss der psychologie [Outline of psychology], (translated by E. B. Titchener). New York, NY: Macmilan.

Lacmann, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction.* Hillsdale: Erlbaum.

Laming, D. (1968). *Information theory of choice-reaction times.* London, England: Academic.

Logan, G. D. (2002). Parallel and serial processing. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology* (Vol 4 pp. 271–300). New York, NY: Wiley.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61–79.

Luce, R. D. (1986). *Response times.* New York, NY: Oxford University Press.

Marx, M. H., & Cronan-Hillix, W. A. (1987). *Systems and theories in psychology.* New York, NY: McGraw-Hill.

McClelleand, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88* (5), 375.

McClelland, J. L., Rumelhart, D. E., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. II).* Cambridge, MA: MIT Press.

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110*(3), 422–471.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Miller, J. O. (1978). Multidimensional same-different judgments: Evidence against independent comparisons of dimensions. *Journal of Experimental Psychology: Human Perception & Performance, 4,* 411–422.

Miller, J. O. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14,* 247–279.

Miller, J. O. (1988). Discrete and continuous models of information processing: Theoretical distinctions and empirical results. *Acta Psychologica, 67,* 191–257.

Murdock, B. B. (1971). Four channel effects in short-term memory. *Psychonomic Science, 24,* 197–198.

Murray, D. J. (1988). *A history of western psychology.* Enaglewood Cliffs, NJ: Prentice Hall.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116,* 499–518.

Nickerson, R. S. (1966). Response times with a memory-dependent decision task. *Journal of Experimental Psychology, 72*(5), 761–769. doi:10.1037/h0023788

Neufeld, R. W., & McCarty, T. S. (1994). A formal analysis of stressor and stress—proneness effects on simple information processing. *British Journal of Mathematical and Statistical Psychology, 47*(2), 193–226.

Nietzsche, F. (1873). Über Wahrheit und Lüge im auSSermoralischen Sinn (On Truth and Lies in an Extra-Moral Sense). In Friedrich Nietzsche (Ed.), *The birth of tragedy and other writings.* Cambridge, England: Cambridge University Press, 1999.

O'Malley, S., & Besner, D. (2008). Reading aloud: Qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34,* 1400–1411. Qualitative

Pachella, R. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human Information Processing: Tutorials in performance and cognition.* Hillsdale, NJ: Erlbaum.

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception & Psychophysics, 41,* 191–201.

Pashler, H. (1998). *The psychology of attention.* Cambridge, MA: MIT Press.

Pashler, H., & Badgio, P. C. (1985). Visual attention and stimulus identification. *Journal of Experimental Psychology: Human Perception & Performance, 11,* 105–121.

Pew, R. W. (1969). The speed-accuracy operating characteristic. *Attention and Performance II.* Amsterdam, Netherlands: North-Holland, 1969.

Rakover, S. (2007). *To Understand a Cat: Methodology and Philosophy.* Amsterdam, Netherlands: John Benjamins.

Ratcliff, R. A. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333.

Risko, E. F., Stolz, J. A., & Besner, D. (2010). Spatial attention modulates feature cross talk in visual word processing *Attention Perception & Psychophysics, 72,* 989–998.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer and S. Kornblum (Eds.) *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 611–653). Cambridge, MA: MIT Press.

Ross, B. H., & Anderson, J. R. (1981). A test of parallel versus serial processing applied to memory retrieval. *Journal of Mathematical Psychology, 24*(3), 183–223.

Schneider, W., & R. M. Shiffrin. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychological Review, 84,* 1–66

Schwarz, W. (1994). Diffusion, superposition, and the redundant-targets effect. *Journal of Mathematical Psychology, 38,* 504–520.

Schweickert, R. (1978). A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology, 18*(2), 105–139.

Schweickert, R. (1982). Scheduling decisions in critical path networks of mental processes. Paper presented in the meeting of the Society for Mathematical Society.

Schweickert, R., Fisher, D.L., & Sung, K. (2012). *Discovering cognitive architecture by selectively influencing mental processes.* Advanced Series on Mathematical Psychology. Singapore: World Scientific.

Schweickert, R., & Mounts, J. R. W. (1998). Additive effects of factors on reaction time and evoked potentials in continuous flow models. In C. Dowling, F. Roberts, and P. Theunes (Eds.) *Recent progress in mathematical psychology* (pp. 311–327). Mahwah, NJ: Erlbaum.

Schweickert, R., & Townsend, J. T. (1989). A trichotomy method: Interactions of factors prolonging sequential and concurrent mental processes in stochastic PERT networks. *Journal of Mathematical Psychology, 33,* 328–347.

Smith, G. A. (1977). Studies of compatibility and models of choice reaction time. In S. Dornic (Ed.), *Attention and performance VI,* Hillsdale, NJ: Erlbaum.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology, 53,* 293–315.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology, 32,* 135–168.

Snodgrass, J. G., Luce, R. D., & Galanter, E. (1967). Some experiments on simple and choice reaction time. *Journal of Experimental Psychology, 75,* 1–17.

Snodgrass, J. G., & Townsend, J. T. (1980). Comparing serial and parallel models: Theory and implementation. *Journal of Experimental Psychology: Human Perception & Performance,6,* 330–354.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74,* 1–29.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153,* 652–654.

Sternberg, S. (1969a). The discovery of processing stages: Extensions of Donders' method. In W. D. Koster (Ed.), *Attention and performance II. Acta Psychologica, 30,* 276–315.

Sternberg, S. (1969b). Memory scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57,* 421–457. (Reprinted in J. S. Antrobus (Ed.), *Cognition and Affect* (pp. 13–58). Boston: Little, Brown).

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science, 12,* 462–466.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643.

Swanson, J. M. & Briggs, J. E. (1969). Information processing as a function of speed vs. accuracy. *Journal of Experimental Psychology, 81,* 223–289.

Taylor, D. A. (1976). Stage analysis of reaction time. *Psychological Bulletin, 83,* 161–191.

Theios, J., Smith, P. G., Haviland, S. E., Traupmann, J., & Moy, M. C. (1973). Memory scanning as a serial self-terminating process. *Journal of Experimental Psychology, 97,* 323–336.

Titchener, E. B.(1905). *Experimental psychology.* New York, NY: Macmillan.

Townsend, J. T. (1969). Mock parallel and serial models and experimental detection of these. *Purdue centennial symposium on information processing.* Purdue, IN: Purdue University Press.

Townsend, J. T. (1971a). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics, 10,* 161–163.

Townsend, J. T. (1971b). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9,* 40–50.

Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical and Statistical Psychology, 25,* 168–199.

Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 133–168). Hillsdale, NJ: Erlbaum.

Townsend, J. T. (1976a). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology, 14,* 219–238.

Townsend, J. T. (1976b). A stochastic theory of matching processes. *Journal of Mathematical Psychology, 14,* 1–52.

Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology, 28,* 363–400.

Townsend, J. T. (1990a). Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science, 1,* 46–54.

Townsend, J. T. (1990b). A potpourri of ingredients for Horse (Race) Soup. (Technical Report #32). Cognitive Science Program. Bloomington, Bloomington, IN.

Townsend, J. T., & Altieri, N. (2012). An accuracy–response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review, 119*(3), 500–516.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan and F. Restle (Eds.), *Cognitive Theory* Vol. III (pp. 200–239). Hillsdale, NJ: Erlbaum.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge, England: Cambridge University.

Townsend, J. T., & Colonius, H. (1997). Parallel processing response times and experimental determination of the stopping rule. *Journal of Mathematical Psychology, 41,* 392–397.

Townsend, J.T., & Eidels, A., (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review, 18,* 659–681.

Townsend, J. T., & Evans, R. (1983). A systems approach to parallel-serial testability and visual feature processing. In H. G. Geissler (Ed.), *Modern Issues in Perception* (pp. 166–189). Berlin: VEB Deutscher Verlag der Wissenschaften.

Townsend, J. T., & Honey, C. J. (2007). Consequences of base time for redundant signals experiments. *Journal of Mathematical Psychology, 51,* 242–265.

Townsend, J. T., Houpt, J. & Silbert, N. H. (2012). General recognition theory extended to include response times: Predictions for a class of parallel systems. *Journal of Mathematical Psychology.* Available at: http://dx.doi.org/10.1016/j.jmp.2012.09.001

Townsend, J. T., & Landon, D. E. (1983). Mathematical models of recognition and confusion in psychology. *International Journal of Mathematical Social Sciences, 4,* 25–71.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology, 39,* 321–360.

Townsend, J. T., & Snodgrass, J. G. (1974). A serial vs. parallel testing paradigm when "same" and "different" comparison rates differ. Paper presented to *Psychonomic Society,* Boston, MA.

Townsend, J. T., & Van Zandt, T. (1990). New theoretical results on testing self-terminating vs. exhaustive processing in rapid search experiments. In H. G. Geissler, M. H. Müller, and W. Prinz (Eds), *Psychophysical explorations of mental structures.* Stuttgart: Hogrefe and Huber Publishers.

Townsend, J. T., & Wenger, M. J. (2004a). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111,* 1003–1035.

Townsend, J. T., & Wenger, M. J. (2004b). The serial parallel dilemma: A case study in a linkage of theory and method. *Psychological Bulletin& Review, 11,* 391–418.

Treisman, A. M., & Gelade, G. (1980). A Feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136.

Treisman, A. M., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95,* 15–48.

Treisman, A. M., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107–141.

Uttal, W. R. (2001). *The new phrenology.* Cambridge, MA: MIT Press.

van der Heiden, A. H. C. (1975). Some evidence for limited capacity parallel self-terminating process in simple visual search tasks. *Acta Psychologica, 39,* 21–41.

Van Zandt, T. (1988). *Testing Serial and Parallel Processing Hypotheses in Visual Whole Report Experiments.* (Master's thesis). Indiana University

Van Zandt, T. (2002). Analysis of response time distributions. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology,* (Vol 4, pp. 461–516). New York, NY: Wiley.

Van Zandt, T.,& Townsend, J. T. (1993). Self-terminating vs. exhaustive processes in rapid visual, and memory search: An evaluative review. *Perception & Psychophysics, 53*(5) 563–580.

Vickers, D. (1979). *Decision processes in visual perception.* New York, NY: Academic.

Vollick, D. N. (1994). *Stochastic models of encoding-latency means and variances in paranoid schizophrenia.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. NN93247). University of Western Ontario, Canada.

Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review, 17,* 479–485.

Welford, A. T. (1980). (Ed.), *Reaction Times.* London: Academic.

Wenger, M. J.,& Townsend, J. T. (2001). *Computational, Geometric, and Process Issues in Facial Cognition: Progress and Challenges.* Mahwah, NJ: Erlbaum.

Wenger, M. J.,& Townsend, J. T. (2006). On the costs and benefits of faces and words: Process characteristics of feature search in highly meaningful stimuli. *Journal of Experimental Psychology: Human Perception & Performance, 32,* 755–779.

Williams, P., Eidels, A., & Townsend, J. T. (2014). The resurrection of Tweedledum and Tweedledee: Bimodality cannot distinguish serial and parallel processes. *Psychonomic Bulletin & Review.* In press.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1,* 202–238.

Woodworth, R. S. (1938). *Experimental Psychology.* New York: Holt.

Wundt, W. (1892). Die geschwindigkeit des gedankens (The velocity of thought).*Die Gartenlaube, 26,* 263–265.

Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy trade-off in choice reaction time. *Journal of Mathematical Psychology, 8,* 159–199.