Suggestions + criticisms invited.

Psychology 600-301
Proseminar in Psychological Methods, Spring Semester 2004

Reaction-Time Experimentation

Saul Sternberg (saul@psych.upenn.edu)
Revised, as of March 20, 2010

"The study of the time relations of mental phenomena is important from several points of view: it serves as an index of mental complexity, giving the sanction of objective demonstration to the results of subjective observation; it indicates a mode of analysis of the simpler mental acts, as well as the relation of these laboratory products to the processes of daily life; it demonstrates the close inter-relation of psychological with physiological facts, an analysis of the former being indispensable to the right comprehension of the latter; it suggests means of lightening and shortening mental operations, and thus offers a mode of improving educational methods; and it promises in various directions to deepen and widen our knowledge of those processes by the complication and elaboration of which our mental life is so wonderfully built up."
— Joseph Jastrow in *The Time-Relations of Mental Phenomena.* (1890)

Today's Goal: To provide acquaintance with some of the issues in designing, conducting, analyzing, interpreting, and evaluating reaction-time (RT) experiments. These issues are best considered in relation to particular substantive questions and interpretations, but time limitations prevent this. One reason for choice of Reading 1 (Keele): In devising experimental procedures, one needs to know what factors influence RT, to avoid confounding them with factors of interest, and to get low-variance data. Important omissions in Keele: sequential effects; aspects of RT distributions other than their means.

Warnings: The ideas to be presented reflect a personal and possibly idiosyncratic view about what sorts of questions are interesting and about how to go about answering them. Also, some of the recommendations have the status of "laboratory lore" — practices that I use and like but that haven't been systematically compared to alternatives, and may not be discussed or even mentioned in the literature. Finally, a good deal of useful information has been gleaned using simple, crude, and informal methods, which deviate considerably from the practices I recommend; please don't let the considerations below deter you from putting your hand in.

1. Why Reaction Time?

Permits studying the system when it is functioning well. (Contrast to traditional memory experiments, e.g., where system is revealed only by its failures when overloaded or otherwise taxed.)

Even when the responses are not fully determined by the stimuli, the time taken to initiate a response may be a more sensitive indicator of the underlying process than which response is chosen.[1]

Good at revealing the temporal organization of mental processes. (E.g., serial vs parallel organization; exhaustive vs self-terminating operations.)

Orderliness of the data, often found, suggests that they are telling us something important — that they may reflect in a straightforward way the underlying processes by which they are generated. [The present handout was accompanied by a collection of figures and their captions showing sixteen sets of pretty RT data, which included instances of additivity and linearity (a special kind of additivity) — that is, instances of the invariance of effect sizes.]

When is RT itself of interest? Seldom, in science. Sometimes in applications. E.g., time to press brake pedal; e.g., forensics (time to pull trigger in the police shooting of a Native Alaskan). What IS of interest? How experimental variables (factors) *change* RT: The *effects* of the factors and how these effects combine.

---

1. Pisoni & Tash (1974) provide a nice example: While the distribution of responses is consistent with "categorical perception" of stop consonants, the RTs of these responses reveal within-category discrimination. ("Same" responses are slower when stimuli differ than when they don't.)

A fundamental concept in thinking about RT data: selectivity of an effect. We are interested in a particular mental process (e.g., how a person makes a decision about a letter of a random size that is presented at a random orientation). The RT is the duration of some set of mental processes, including the one of interest. (It is a *composite measure*.) One task of the psychologist is to disentangle the subprocess of interest from the others. To study the subprocess of interest we would like to find one or more factors that influence only that subprocess, and not the others. If such *selective influence* obtains, then the effect of the factor is selective, and informs us about the subprocess of interest. Selectivity is sometimes assumed with little or no justification. Example: The effect of flash intensity on the RT has been assumed to reflect only its effect on the time to detect the flash (not on processes between detection and response) and used to study visual detection latency.

Analogy to signal-detection theory (SDT): if we are interested in a sensory process, then we vary the level of some factor and examine its effects on the pattern of errors. To correctly interpret such effects we have to acknowledge that these patterns are influenced not only by the sensory process of interest but also by decision processes. One approach is to find a measure that reflects only the sensory process (such as d', given certain assumptions). d' is then a selective measure, and the effects on it are selective effects. Thus, SDT is a method for decomposing the mental process in certain psychophysical experiments into sensory and decision subprocesses. Similarly, the way in which effects of factors combine in influencing RT can be used to make inferences about the organization of the processes that generate the RTs — the "mental architecture" — and thus draw conclusions about the effects of factors on particular subprocesses.

The method of additive factors (AFM) is one way to make such inferences. This approach to dividing complex mental processes into subprocesses depends on the observation that if a process contains subprocesses arranged in stages so that the RT is the sum of stage durations, and if two factors F and G (experimental variables) influence different stages ("selective influence") and influence no stages in common, then their effects on mean reaction time should be additive. That is, the effect of (changing the level of) F on mean RT should be invariant as the level of G is changed, and vice versa. Conversely, if G modulates the effect of F rather than leaving it invariant, then F and G must influence at least one stage in common. Suppose, then, that we have a process in which behavioral experiments have revealed two (or more) factors with (mutually) additive effects. One interpretation is that the process contains corresponding subprocesses arranged sequentially, in stages, with each of the factors influencing a different one of the subprocesses selectively. (Given stronger assumptions, selective influence plus stages implies properties of other features of the RT distributions in addition to their means.)[2]

[Exercise: Suppose a process consists of two subprocesses, **A** and **B** that operate in parallel, such that the response occurs when *both* **A** and **B** are completed. Suppose that factor *F* influences only the duration of process **A**, and factor *G* influences only the duration of process **B**. How will the effects of factors *F* and *G* on the mean RT combine? Hint: Assume two levels of each factor and consider the four resulting conditions. Extension: Suppose the response occurs when the *first* of **A** and **B** is completed. What would one conclude from the AFM in these cases?]

1.1 The problem of errors

One of the difficult issues associated with the interpretation of RT data arises from the occurrence of errors. Insofar as subjects are trading accuracy for speed, and may be doing so to different extents in different conditions, any straightforward interpretation of the RTs alone becomes difficult. Furthermore, the trading relation is likely to be sufficiently complicated so that "correcting" the observed RTs in different

---

2. Additive effects on $\overline{RT}$ have been of sufficient interest so that alternatives to stage models have been considered as explanations. Since the AFM was first introduced it has been discovered that under some conditions, other models, quite different in spirit from stage models, can also generate such additive effects. However, in all these cases, the prediction of means additivity depends on the existence of distinct processes ("modules") plus selective influence; hence, from the viewpoint of discovering modules (but not of how these modules are organized), the existence of these alternative possibilities doesn't weaken the argument from the additivity of factor effects on $\overline{RT}$ to the existence of modules. Their discovery, however, weakens the inference that these modules are organized as stages. Additional aspects of the *RT* data can sometimes help distinguish among stage models and alternatives. Other approaches to such model selection include techniques such as speed-accuracy decomposition and concomitant electrophysiological or behavioral measurements.

conditions for their associated error rates is likely to be impossible. For example, given that the time from stimulus to response is occupied by more than one process, there can be more than one tradeoff. (See, e.g., Osman et al., 2000, and Luce, 1986, Section 6.5.) And while there exist models (see references in Reading 2) which, if correct, "explain" both errors and RTs in terms of a single underlying mechanism, such models are controversial, complex, and likely to be valid only under limited conditions. (Work with such models usually requires relatively high error rates.) I believe that speed-accuracy trading can indeed occur, but that under "standard" RT instructions it usually doesn't. Instead, subjects respond when the process they are using is complete. My evidence? Mostly the orderliness of data collected under "standard" conditions. Informally, the invariance of mean RT under changes in error rate. (See Reading 2.)

2. Method: General Goals

One goal should be to reduce variability and drift of the RT. A second goal should be to eliminate systematic differences across conditions in fatigue, practice, motivation, and any other factors not explicitly manipulated. A third goal should be to get subjects to perform "optimally". (But "optimal" is not well defined in this context.) The ubiquity of individual differences argues for within-subject comparisons wherever possible. The ubiquity of decelerating practice effects argues for checking trends, trying to achieve some degree of stability before collecting the data of principal interest, and balancing over such effects.[3] To minimize variability calls for minimizing variation in alertness/fatigue and motivation.

3. Procedure

3.1 Response measurement

3.1.1 Manual responses

For two-choice tasks I have avoided finger responses, and have typically used two levers, one pulled by each hand. Between trials the hand can rest, fingers bent, fingertips touching the horizontal surface. The response — pulling the lever — involves the flexing of all four fingers. I don't like arrangements that require fingers to be poised over keys (fatiguing, I think, and conducive to variation in resting posture that might influence RT). I especially don't like arrangements where the same finger can make either response, starting from a "home" position, sometimes a key. This adds time, and invites differential preparation expressed in the resting state.

For more than two alternatives, I like a set-up where the palm rests on a curved surface, each finger-tip touching a short vertically oriented lever. Again, the aim is a posture that is not fatiguing, in which the resting effector is touching the manipulandum.

All such arrangements are problematic if the display consists of multiple items distributed over space, because of spatial "stimulus-response compatibility" effects. Thus, suppose visual search for the presence of a target in a horizontal array that contains two items, a left item and a right item. Suppose the target item is present, and on the right. If the right hand is assigned to "present" (yes), the left hand to "absent" (no), $RT_{yes} - RT_{no}$ will be less than if the left hand says "yes", the right hand "no".[4] Similarly, in a visual search task, if "present" is signalled by a right-hand response, the response will be faster for targets more toward the right of the display. In general, one needs to think about the compatibility of the required response with the outcome of the required mental operations.

3.1.2 Vocal responses

Although it may seem implausible, vocal responses (e.g., "yes", "no") can (especially under the above circumstances) be better than manual ones. I once compared them in an informal experiment in the above

---

3. The possibility of balancing of conditions over levels of some nuisance factor (such as practice) depends on additivity of the effects of conditions and the nuisance factor. Such additivity is often assumed without justification.

4. Such effects may be reduced (but not necessarily eliminated) by positioning both response alternatives in the sagittal plane.

task, and found not only shorter RTs for the vocal response, but markedly lower error rates as well. As expected, there were spatial S-R compatibility effects with the manual responses. Some care has to be used in measuring the RT for vocal responses. The starting sounds of different words differ in both energy and frequency range. (This calls for balancing, either actual or statistical.) And I like to keep the voice level high relative to its threshold. (I believe that if the voice level is sufficiently low so that it barely exceeds the threshold, the latency measurement will be more variable.) It must also be kept in mind that the amount of sound energy at the start (and end) of a word differs dramatically across words with different sounds, with (low-frequency) vowels being far more energetic than (high-frequency) fricatives or sibilants, for example. So, although not essential, I like to separate the speech signal into high- and low-frequency bands, with lower thresholds for the former, and to require the peak energy in each case to be high relative to the threshold. Also, even with this kind of arrangement, different words take different mean times to trigger the voice detector, so experiments must be designed that don't depend on comparing the times for different responses. (A good idea for non-vocal responses as well.) It is important to check your voice detector (whether hardware or software) with an oscilloscope.

Another consideration: merely opening the lips can produce a "pop" that triggers both low and high thresholds. (Subjects can be trained to reduce pop frequency by separating their lips slightly during the foreperiod.) But pops are distinguished from speech by their brevity. This is one reason for measuring the duration as well as the latency of vocal responses. Another reason is that duration information can sometimes reveal badly produced responses, or responses that start with one word and end with another, or responses that start with "uh". Ideally, however, these are caught by an experimenter. Using vocal responses does tend to require the expense of either an experimenter constantly present or a good artificial speech recognizer. But I feel that the presence of (and encouragement by) an experimenter is highly desirable for other reasons.

To reduce variability in speech latency measurement, it is desirable to maintain a constant distance between talker and microphone. A good way to do this is to use a boom mike, positioned at a standardized distance from the mouth, and out of the air stream. And in case it isn't obvious to the reader, sensitivity to low-energy sounds is aided by minimizing background noise.

3.1.3 More on manual responses

In the most recent version of the levers of the sort I have described above, each lever operates two switches, one early and one late during its travel, akin to low and high thresholds. Here one could use the early switch to register the RT, but (a) require the late switch to register, and perhaps (b) record the second time as well, using the time difference as a way of discovering responses that were executed unusually. (See Section 6.4 below.) For some purposes (e.g., avoiding the electrical artifacts generated by speech-related muscles, or operating in noisy environments such as MRI scanners), handwritten responses seem preferable to spoken ones. The technology is available to determine the times of the beginning and end of a written production and to decide whether it is correct or incorrect or abortive. By rewarding the subject for finishing fast, and penalizing anticipations, subjects can easily be trained to limit the contact between writing implement and surface to what is needed for production of the response. Duration measurements are useful for training purposes and to catch unusual response executions. But I know of very few studies that use such measurements. I have used a crude version of this method, and replicated effects previously found with vocal responses, when the responses were handwritten digits rather than their spoken names.

3.2 Stimulus design

Given a set of factors of principal interest, try to avoid confounding other aspects of the stimulus with them. Such confounding is sometimes unavoidable. For example, suppose the number of elements, *n*, in a visual display is a factor of principal interest. When this is varied, then either the size or density of the display must vary with it. One approach is then to deliberately vary the contaminating factor over a suitable range orthogonally with *n*, so as to measure its effect separately. (This might have to be done in a side experiment.) With luck, its effect might be negligible. Otherwise one might be able to "correct" for its effect statistically. But to do the latter one would need at least a primitive model that describes how the effect of size (or density) combines with the effect of *n*.

With large displays, making the reaction stimulus (RS) brief (< 200*ms*) can avoid contamination from eye movements, if that is desirable. With small displays, very brief stimuli (e.g., ≤ 50*ms*) may be advantageous: they encourage appropriate attention and fixation, discourage blinking at the wrong time, and increase the chance that a blink will lead to an error rather than inflate the RTs of correct responses. See Johns et al. (2009) on effects on RT of blinking and eye movements. Another way to reduce blinking when the stimulus is presented is to encourage it during the foreperiod.

3.3 Control of temporal expectancy

I believe it is best for the subject to know as precisely as possible when a stimulus that requires a response will be presented. Thus, in choice-reaction tasks I think two successive auditory warnings are good, in a count-down arrangement, with the interval between W_1 and W_2 the same as the interval between W_2 and *RS*, the reaction signal. A good interval is 0.7 sec. I believe that when concentration is required only at predictable times, with a chance to relax between them, even sleep-deprived undergraduates can produce good data. Similarly, in a simple-reaction task I prefer catch trials (omission of the *RS* on some fraction of trials) rather than a variable foreperiod, as the preferred way to minimize anticipations, because it reduces variation in the level of preparation.[5]

3.4 Other control of preparation

Subjects should be made as comfortable as possible and be isolated from extraneous stimuli. In some cases it may be good to have the subject initiate each trial, but I have seldom done this: it gives the subject another task, and may become automatic and hence not serve its ostensive purpose of ensuring that the subject is (adequately) prepared. Trial blocks should be relatively short (say, 20 trials), with intertrial intervals of two or three seconds and enforced rests between blocks. (A reaction-time experiment should not be a vigilance task.) Sessions should probably be no longer than an hour.

3.5 Instructions, feedback, and payoffs

Instructions such as "minimize your reaction time without making too many errors" are inadequate, in my view. To deal with the possibility of speed/accuracy trading, I prefer to give the subject a score that explicitly reflects both time and errors and that is ultimately convertible into cash. For example, one point for each hundredth of a second in the mean RT for a block of trials, and 20 points for each error. The cost of an error should be at least high enough so that chance performance with zero RT is not rewarded. Cash rewards can be based on the relation among scores for different subjects ("You'll get an extra dollar if your score today places you among the best half of the subjects."), or on the relation among the amounts of improvement ("You'll get an extra dollar if the amount of improvement in your score relative to yesterday places you among the best half of the subjects."), or on absolute improvement ("You'll get an extra dime for each block in which your score is better than your average score in the same condition yesterday.") The last example is good if conditions are blocked; I think that it tends to equate the amount of pressure on the subject across conditions (highly desirable). There should also be a payoff that depends on performance over the whole session, however. (Otherwise there is a risk that the subject will "give up" during a block after making a couple of errors.)

The use of RT deadlines to motivate subjects is questionable, because it is difficult to control the deadlines so as to put the same pressure on the subject in all conditions and on trials of different difficulty within conditions; other objections include the possibility of distorting the RT distribution.

I think that subjects' interest and motivation can be maintained if they get performance feedback (and encouragement by the experimenter) after each trial block, of mean RT and number of errors, as well as the score based on these. I like to inform the subject immediately if there is an error, because of the importance of keeping accuracy high, But otherwise give no trial-to-trial feedback. Providing RT feedback from trial to trial takes extra time and distracts the subject. I have found that occasional face-to-face contact with the experimenter (in addition to periodic oral communication over an intercom) improves subjects'

---

5. If for some reason variable foreperiods are desirable, one technique is to use a non-uniform distribution ("non-aging foreperiods") in which the conditional probability of a signal in the next short interval, given no signal to that point, is constant.

performance.

3.6 Error therapy

Given a scoring system in which errors are heavily penalized, most subjects can adjust their performance so that accuracy is high. But occasionally I have encountered a subject who seems unable to make this adjustment, and for whom each error seems to beget others. Such subjects generally benefit from one or two trial blocks in which time pressure is entirely removed and accurate responding is the only goal.

3.7 Calibration

Don't trust the computer or the program.

4. Design Issues

Four truths:

We cannot avoid individual differences.

We cannot avoid the effects of practice; RT diminishes with practice.

We may not be able to avoid strategy variations (alternative mental processes) in accomplishing a task.
(But it is desirable to try to create procedures with the improved experimental control that will do so.)

Durations of mental processes are variable.

An important assumption: Subjects try to and succeed in optimizing. That is, they want to do well, and, through practice, arrive at an optimal strategy, where "optimal" (not well defined) is presumably controlled by an explicit scoring system. Thus, efforts should be made to increase the chances that this assumption is valid. (One way is perhaps by explicitly suggesting to subjects that they explore alternative strategies during unscored practice blocks.) Otherwise it is unclear what any differences in RT across conditions might mean.

5. Factorial Experiments (YES!!)

"No aphorism is more frequently repeated . . . than that we must ask Nature . . . ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed."

(From Sir Ronald A. Fisher (1926), in the paper that reported his invention of the factorial experiment.)

5.1 Fishing expeditions — extra factors (YES!!)

It is desirable to use factorial designs that include subsidiary factors as well as the main factor(s) of interest. Advantages: test the generality of the effects of the primary factors; possibility of new discovery. Cost in terms of requiring additional observations: little. Introducing additional factors might initially seem to require much larger experiments. And certainly, to get good precision for all the possible interactions would require an increase in the number of trials. However, if one's primary concern is with only a subset of all possible interactions, experiments need not grow substantially. Suppose one starts by considering an experiment with two factors of principal interest, F and G. Now divide the trials for each F-G combination in half, and use two levels of factor H between the two halves. This permits examining (a) the interaction between F and G (for the average H) and in addition, (b) the interaction between F and H (for the average G), as well as (c) the interaction between G and H (for the average F). Without manipulation of factor H (which did not increase the size of the experiment) one could examine only (a).

5.2 Reminder about the language of factorial experiments

The set of values of a factor used in an experiment are called its **levels**. (For example, the factor *gender* has the levels *male* and *female*; in a particular experiment the factor *size of memory set* might have the levels 1, 2, and 4.) In the simplest factorial experiment (the "complete factorial") with two factors, *F* and *G*,

there is a **condition** for each possible combination of factor levels. Thus, consider a pattern-recognition experiment in which one factor is *orientation* of the pattern, with the three levels 0^o, 90^o, and 180^o, and the other factor is *familiarity*, with the two levels low, L, and high, H. The complete factorial experiment would then have the six conditions (0^o, L), (0^o, H), (90^o, L), (90^o, H), (180^o, L), and (180^o, H). The **effect** of a factor F that has two levels, F_1 and F_2 in a reaction-time experiment is the difference between the $\overline{RT}$s at its two levels: $\overline{RT}(F_2) - \overline{RT}(F_1)$. Thus we would measure the effect of familiarity at orientation 0^o by calculating the difference between $\overline{RT}$s in two conditions: $\overline{RT}(0^o, L) - \overline{RT}(0^o, H)$. Because this is the effect of familiarity at one particular level of orientation, it is called a **simple effect**. The **main effect** is the mean of all the simple effects. It is convenient to number the levels of F — that is, assign values to the **index**, i, of F_i — so the main effect is positive; we can then refer to the change from F_1 to F_2 as an "increase" in the level of the factor. The idea of an *effect*, defined here for a factor with two levels (familiarity) can be generalized to a factor with $k > 2$ levels (such as orientation); It is a vector with $k-1$ elements.

5.2.1 Interacting and additive factors

If the simple effect of one factor (say, familiarity) varies with the level of another factor (say, orientation), then we say that the two factors *interact*: the effect of one factor modulates the effect of the other. If, instead, the (simple) effect of familiarity is *invariant* across levels of the orientation factor, their interaction is zero, and familiarity and orientation have *additive* effects. That is, the effect of changing the levels of both factors is given by the *sum* of the effects of changing each of them separately. See notes on the numeral-naming experiment (Appendix 1) for more discussion of the meaning of interaction.

Suppose you are interested in the effect of the difference in familiarity between two particular levels. If factors such familiarity and orientation interact, the simple effects of familiarity (and hence the main effect) depend on the levels of orientation that happen to have been used in the experiment. This presents problems for generalizing from the size of the main effect.

5.3 Factors: How many levels?

In many psychological experiments, either factors are selected for which only two levels can be defined, or factors for which multiple levels could be defined are examined at only two. The use of multiple levels has several advantages. First, it provides another test of generality: For example, given many hypotheses, selectivity of a factor's effect (or interaction of that factor with another) should obtain for all pairs of levels, and not be an accident of a particular pair of levels. Second, it helps determine whether a slight failure of invariance — possibly non-significant — reflects a systematic trend or random variation. And third, it helps to test whether the factor is *unitary* over its range, such that each change in level influences the same processes and leaves the same other processes unaltered. (On the other hand, if factor levels are blocked, multiple levels increase the challenge of balancing them across the levels of other factors.)

5.4 Balancing, actual and statistical

Suppose that factor F is of principal interest, but other factors, such as G and H are also varied, either deliberately (e.g., which of a set of displayed items is probed) or unavoidably (e.g., level of training). The usual practice is to balance levels of G (for example) over levels of F: that is, to study each level of F at each level of G with equal sample sizes. The usual assumption is that by averaging over levels of G one gets a "true" measure at each level of F. It is important to recognize that this requires the assumption that the effects of F and G are additive on the measure that is used — here, $\overline{RT}$. Even if the experiment is designed with equal numbers of trials at each combination of levels, the occurrence of error trials and their removal from the data will violate this equality. Thus one should use *statistical balancing*: get means for each combination of levels, and then get the equally weighted mean of those means. Sometimes, even disregarding the problem created by the removal of error trials, balancing is not possible. For example, I have done experiments where horizontal arrays of different numbers of items were displayed in contiguous locations among a specified set of locations, and one location was probed. Factors that (might) have effects include array size (number of items, which might be the factor of principal interest), serial position of the probed item in the array from left to right, and absolute position of the probed item in the visual field. It isn't possible to balance both serial and absolute position: given a fixed set of allowed locations in the

visual field, some combinations of absolute and serial position are impossible. In this case, statistical balancing can be accomplished by using linear regression to provide estimates of the effects of serial and absolute position, and using these estimates to remove these effects. (The assumption of additivity that is required in using linear regression for this purpose is the same assumption used in conventional balancing, and should be tested, to the extent possible.)

Counterbalancing of levels of a factor that has an effect (e.g., the ordering of conditions across subjects) can cause erroneously inflated error terms. To avoid this one must either correct for the effects of such a factor, or include it as a factor in any anova.

5.5 Re-running of error trials

The occurrence of errors for certain trial types can reduce sample size below acceptable levels for those trial types, and can destroy the balance in a balanced design. Experimenters sometimes cope with this by re-running those trials, for example at the end of the block. But doing so means that these trial types are not balanced with respect to position in the trial sequence, and also may lead to the clustering of the more difficult trials. My preference is to anticipate the possibility of errors by increasing the designed number of the more difficult trials, and to use statistical balancing.

5.6 Within-subject versus between-subject designs

The ubiquity of individual differences in $\overline{RT}$ argues strongly for within-subject designs. But the possibility of asymmetric transfer effects from one condition to another emphasized by Poulton (1982), and the biases they may generate, argues at least for caution in interpreting the results from within-subject experiments, and for the value of checking for such effects.[6]

5.7 Should conditions be blocked or randomized?

Balancing of fatigue and practice effects over conditions is of course easier when conditions are randomized from trial to trial. And it can be argued that condition-specific strategies or other adjustments are less likely: the subject is more likely to be in the same state across conditions. On the other hand, if one is interested in condition-specific optimality and ultimate performance limitations, blocked conditions may be better. Ideally one would like to compare conditions in both ways, and hope for similar results.[7] Of course, some factors can be blocked while others are randomized.

5.8 Sequential structure of trial series

It is tempting to use pure randomization to determine the order in which different trial types are presented in the trial sequences of an experiment. (For testing of some models this is necessary, to guarantee the absence of any sequential constraint.) This wouldn't even control the relative frequency of the set of trial types (which one might want to equalize, for example). So one might add the constraint of specifying the number of occurrences of each trial type, and then use random permutation to determine the order. (Note, however, that in principle, this alone permits the subject to predict the next trial type with greater than chance accuracy, based on the trials that have already occurred.)

To avoid accidental confounding of practice with trial type, it is good to apply such a constraint in units smaller than the whole experiment or even the whole session. I call such units "micro-replications", and try to have at least two or three in each session. (This helps to balance practice effects across trial types, and also facilitates the estimation of such effects.) To reduce the subject's ability to predict, I try to arrange block boundaries not to coincide with micro-replication boundaries, and, in addition, I include one to three "dummy" trials at the start of each block that are not part of the micro-replication and are also there for

---

6. To the extent that individual differences are expressible as additive constants, one way to reduce the size of the samples required for between-subject designs to compare the conditions of primary interest is to train all subjects under a condition or in a task believed to be neutral relative to them, and use the performance in that task as a covariate — i.e., to provide estimates of the additive constants.

7. Poulton (1982) argued that his concerns about biases in within-subject designs, due to asymmetric transfer across conditions, apply especially when conditions are random from trial to trial.

warm-up purposes, and to accustom the subject to new conditions.

Subjects may try to predict the next trial type even when sequences are perfectly random. (For example, they may engage in the "gambler's fallacy", acting after a "long" run of one kind of trial as if there is likely to be a change.) This is a problem because variation in subjects' expectations from trial to trial add variability. (It may help to inform the subject that there is no pattern in the trial sequences, and that prediction isn't possible.) It requires only looking at a sequence of random numbers to appreciate the fact that random sequences often appear non-random. This is especially a problem in experiments with only two alternative responses. For such experiments I usually construct trial sequences in which run lengths are constrained. One approach is to compute the expected run-length distribution for a purely random sequence and use a distribution that is similar, but excludes runs exceeding some length, such as three or four. For an experiment with two trial types, sequence preparation then involves alternation between randomly selected runs from the distribution of the two types. Another approach is to apply the run-length constraint by generating many sequences and rejecting those in which it is violated; this has to be done with care, however, to avoid inadvertently imposing other constraints that are unwelcome.[8] Another kind of constraint that may be helpful in reducing variability, if there are multiple stimuli and responses, is to avoid immediate repetition of stimuli and/or responses from trial to trial.

5.9 Reduction of practice effects by pretraining

The initial performance of untrained subjects tends to be very slow and highly variable. There is much to be said for training subjects in each of the conditions of an experiment before the experiment proper begins. While practice effects never end, they decelerate. Thus, while we need to balance over practice with trained as well as untrained subjects, the additivity assumptions required for balancing to work are more likely to be approximated when changes in performance during the experiment proper have been reduced by training.[9] For the sample experiment described in Appendix 1, in which the RT distributions as well as their means were of interest, it seemed especially important for the means to be stable and variances low. I therefore gave the subjects five hours of training in the eight conditions before the experiment proper.

I generally like to give subjects occasional rests, so I use trial blocks that contain no more than about 30 trials. Also, to increase the chance that subjects will be in approximately the same state on each trial that I analyze, I generally start blocks with one or two "warm-up" trials that are discarded.

6. Treatment of Data

6.1 Asymmetry of traditional statistical tests

Strong inferences (as in the AFM) can sometimes be made from the *invariance* of the effect of one factor as other factors are varied (the absence of interaction). Traditional hypothesis testing identifies such invariance as the null hypothesis, and induces researchers to draw strong conclusions (e.g., invariance) by default, as a failure to reject the null hypothesis — simply because the data are weak or noisy. If such tests are used it is important to evaluate their power; one way is to calculate confidence intervals around a measure of the size of an effect or interaction.

6.2 Use of focused rather than global hypothesis tests

Suppose a factorial experiment in which one or more of the factors are studied at more than two levels. Frequently the levels are ordered. (Examples: display sizes, retention intervals, list lengths, ages.) Suppose

---

8. In this situation, the correct response on a trial following the longest run is, in principal, perfectly predictable, so the data from such trials should be checked, and possibly rejected.

9. Initial practice is especially important when data collection is expensive, as in fMRI, or when the experiment proper has to be brief because stimuli are limited in number and cannot be repeated. In the former case, subjects should ideally be trained outside the scanner, lying down in an appropriately noisy environment. In the latter case, subjects can be trained with stimuli that are similar to but not the same as the critical ones.

you want to decide whether such a factor interacts with others. The "alternative hypothesis" (to additivity) in the most common test (e.g., in an anova) is probably a *global* hypothesis — sensitive to any difference in the pattern produced by the levels of one factor as the levels of other factors are changed. But often what we expect if there is any interaction is a *monotone interaction*: For example, if increasing the level of G from G_1 to G_2 increases the $\overline{RT}$ difference from F_1 to F_2, then a further increase, from G_2 to G_3 should increase $\overline{RT}(F_2) - \overline{RT}(F_1)$ even further. In that case what we would like is a *focused* hypothesis test rather than an "omnibus test", more sensitive to a monotone interaction while being less sensitive to other kinds of deviation from additivity. One way to approach this is to consider the levels of F as a numeric variable rather than a categorical factor, in anova programs. For discussion of this idea and an explanation of how a monotone interaction can be expressed as a multiplicative one with an associated 1 *df* measure of the size of the interaction, see Sternberg (2001, Section 15.1).

6.3 Contamination by outliers

Inattention or other "glitches" can sometimes cause RTs to be prolonged, and bring them far from the bulk of the RT distribution. Less often, anticipations or other slips can generate unusually short RTs. These are "outliers". There are various methods (e.g., in the statistical literature) for identifying outliers; investigators are often tempted to use one of them, and adjust or discard the outliers. (Some investigators use robust alternatives to the sample mean, such as medians or trimmed means, to deal with this problem.) If one is interested in the population means under a set of experimental conditions, such methods can often introduce large and unknown biases in their estimates. Suppose, as is often the case, the population mean is of interest. The sample mean estimates it without bias, but possibly with great sensitivity to outliers. A robust measure such as the median or trimmed mean may have lower variability, but may have a bias that varies across conditions. (See, e.g., Ulrich & Miller, 1994.)

Some years ago the late John W. Tukey, one of the world's most famous statisticians and an expert in methods for identifying outliers, dropped into my office to discuss a report I had just produced in which I had made use of some statistics of RT distributions (the second, third, and fourth moments) that are extremely sensitive to contamination by outliers. I was delighted; here was my chance to get some advice about the latest statistical methods for dealing with such contamination. To my surprise, he was unwilling to discuss outlier removal, and instead focused on discussing why the outliers occurred, and what could be done to reduce their frequency. In short, his advice was that I should use better experimental technique, aimed at reducing variability, rather than a statistical band-aid.

6.4 Concomitant observations

One way to identify aberrant data is to use supplementary observations on each trial, in addition to the RT itself. The duration of the response is one example. In one session of an experiment with vocal responses, one of the subjects produced some unusually short RTs. Fortunately we were measuring the durations as well as the latencies of the responses, and noticed that the durations were extremely short. It emerged that on that day the subject had been wearing a dress with sequins, and her abrupt movements preceding her vocal responses on some trials caused short RTs to be recorded.

6.5 RTs on error trials

The correctness of the response can be regarded as a concomitant observation; under conditions of high accuracy it is reasonable to assume that if an error occurred, the process being studied was not carried out "normally". It should be kept in mind, however, that some of the responses likely to have been generated by the same abnormal process are correct, and that the contamination they produce can't be readily removed. One reason to keep accuracy high is to minimize such contamination.[10]

6.6 RTs after error trials

Performance on trials after errors tends to be slower. (A procedural device that may help to reduce this

---

10. Suppose, for example, that the 5% of errors in a two-choice experiment with equally frequent responses were produced by a random "guessing" process. Then a similar proportion of the correct responses are likely to have been generated by the same guessing process.

effect is to lengthen the intertrial intervals after errors.) One way to reduce contamination is to remove the data from such trials from the analysis.

6.7 Nonlinear transformations of RT

One example is speed of response (the reciprocal of RT). If effects on the non-transformed mean RTs are additive, the additivity will be destroyed by such transformations. In contemplating such transformations, it is good to keep in mind that even if the subprocesses of interest are not arranged serially, at least some of the input and output processes are likely to be. Whether the distortion of the additivity of their durations is important will depend on the kinds of inference to be made from the data.

6.8 RT distributions

LOOK at the distributions. If the mean changes, what about the RT distribution is changing? Examine how measures other than the mean vary with conditions. Examples of other measures are the minimum (Donders used this in the very early days); maximum; variance or other measures of spread; skewness.

7. RT Paradigms

Simple RT (Donders' "a-reaction")
(One stimulus, one response known in advance)
Useful for studying stimulus detection (contrasted with discrimination)
Useful for studying response execution (contrasted with selection).
Need for "catch" (no-signal) trials, or variable foreperiod.
Idea of non-aging foreperiod (where conditional signal probability doesn't increase with time).
(Preference: catch (non-signal) trials, with suitable penalty for responding.)

Choice RT: 1-1 mapping. (Donders' "b-reaction")
Number, N, of S-R alternatives.
Evidence that N has an effect on at least two processes (stimulus identification and response selection).
Probability effects; sequential effects.

Choice RT: Go/No-Go. (Donders' "c-reaction")
Issue of delayed responses (No-go can be corrected to Go).

Choice RT: Many-one mapping (categorizing).

8. Paradigm Transfer

Use of experiments from the cognitive lab with patients or with brain measurements, where procedures may have to be adjusted. One issue: Can the adjusted procedures, applied to normals, produce data that match the "standard" results? I have frequently noticed that the standard results are not replicated, and that important details of procedure have been overlooked; even control data (from normals) often seem highly variable; mean RTs are often very long; no description of efforts to motivate subjects to perform well; inadequate practice/training; no examination of practice effects.

9. Data Retention

Data are precious; storage is (now) remarkably cheap. RT data especially are very rich. You (or someone else) may have a question that a retained data set can answer. Save the data, including the trial by trial sequences. And (less obvious) save a complete description of the experiment, including information about which data come from which conditions. Even your memory is fallible.

References (with some updates, 2015)

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503-513.

Johns, M., Crowley, K., Chapman, R., Tucker, A., & Hocking, C. (2009). The effect of blinks and saccadic eye movements on visual reaction times. *Attention, perception, & psychophysics, 71,* 783-788.

Osman, A., Lou, L., Muller-Gethmann, H., Rinkenauer, G., Mattes, S., & Ulrich, R. (2000) Mechanisms of speed-accuracy tradeoff: evidence from covert motor processes. *Biological Psychology, 51,* 173-199.

Pisoni, D. B. & Tash, J. (1974) Reaction times to comparisons within and across phonetic boundaries. *Perception & Psychophysics, 15,* 285-290.

Poulton, E. C. (1982) Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin, 91,* 673-690.

Sternberg, S. (2004) Some Pretty RT Data. http://www.psych.upenn.edu/˜saul/pretty.pdf

Sternberg, S. (2013) More Pretty RT Data. http://www.psych.upenn.edu/˜saul/more.pretty.data.html

Sternberg, S. (2013) The meaning of additive reaction-time effects: Some misconceptions. *Frontiers in Psychology, 4:* 744. doi:10.3389/fpsyg.2013.00744

Sternberg, S. & Backus, B. (In Press, 2015). Sequential processes and the shapes of reaction-time distributions. *Psychological Review*. http://www.psych.upenn.edu/˜saul/ms.web.pdf

Ulrich, R. & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123,* 34-80.

Acknowledgement

I thank Allen Osman for comments and suggeestions.

Appendix 1: Sample Experiment: A $2 \times 2 \times 2$ Experiment on Numeral Naming

The process of naming a numeral (or a letter, or a word) is rapid (about 350 ms from display to response, for numerals), and seems automatic, effortless, and unavailable to consciousness. A mystery that has never been adequately explained is the influence on the *RT* not only of the item presented, but also of the set of items that *might* have been presented: if the item is drawn from a smaller set of possibilities, the corresponding response can be produced more rapidly. These properties make the problem of analyzing the naming process into components especially intriguing. In an experiment I did to look into this I varied the difficulty of the task in three different ways, manipulating the levels of three factors. My goal was to determine how their effects would combine in influencing *RT*, and thereby learn about the structure of the process. The resulting mean RTs from five practiced subjects are given in Table 1; the main effects and interactions derived from the values in Table 1 are given in Table 2.

On each trial the subject saw a one-digit numeral and responded by speaking the name of a numeral as fast as possible consistent with high accuracy. In some series of trials the correct response was the name of the displayed numeral. (If a "3" was displayed, the subject would speak the word "three".) In other series the correct response was the name of the next larger numeral. (If a "3" was displayed, the subject would speak the word "four"). The factor being varied is the *mapping* of stimuli onto responses, in particular, the familiarity of this mapping. We can call the two levels of mapping familiarity (*MF*) in the experiment "familiar" and "unfamiliar".

The second factor that I varied was the "quality" of the stimulus, *SQ*. The numeral was either easy to read — that is, *intact* — or it was *degraded* by a superimposed pattern. Not surprisingly, degrading the stimulus increases $\overline{RT}$. (The trick is to adjust the extent of degradation to maximally slow the response, subject to keeping accuracy high (e.g. ≥ 95%.)

One important question is whether the *effect* of *SQ* — the change in $\overline{RT}$ as the level of the stimulus quality is changed from intact to degraded — is *invariant* over the two levels of the mapping-familiarity factor. Note that the invariance of the effect of one factor across the levels of another is a symmetric relation: if the *SQ* effect is invariant over the levels of *MF*, then the effect of *MF* — from high-familiarity to low — would be invariant over the levels of *SQ*. Given such invariance, the two factors have *additive* effects: the effect of changing the levels of both factors is the sum of the effects of changing each of them separately. The alternative is that *MF* modulates the effect of *SQ*. Again, this relation is symmetric: if *MF* modulates the effect of *SQ*, then *SQ* modulates the effect of *MF*, and we say that the two factors *interact*.

When the factors have only two levels, as in this experiment, the question whether they interact or have additive effects can be decided by examining a single number — the "interaction contrast". Suppose that the levels of *SQ* are indexed by $i = 1, 2$ and the levels of *MF* by $j = 1, 2$. Then, if those were the only two factors, the experiment would produce four mean RTs (for each subject) that we could call $T_{ij} = T_{11}, T_{12}, T_{21}, T_{22}$. The interaction contrast could then be written $(T_{22} - T_{12}) - (T_{21} - T_{11})$. This is of course the difference between the (simple) effects of the SQ_i factor at the two levels of the MF_j factor. In terms of the differencing operator used and defined in Table 2, this interaction contrast can also be written $D_{ij}(T_{ij})$

The third factor that I varied in the experiment was the number, *NA*, of different possible numerals that might be presented during a series of trials, that is, the number of alternative S-R pairs. In some conditions $NA = 2$, and in others, $NA = 8$. It is well-known that increasing *NA* causes an increase in $\overline{RT}$. (This is a case where the set of signals that *might* occur on a trial, but don't, influence the response to the signal that does occur, a possibility that psychologists became interested in when the theory of information was invented by Claude Shannon (1948). With more uncertainty about what *might* happen, what *does* happen conveys more "information," and the time it takes a person to make the same response to the same stimulus increases with the amount of "information" it conveys.)

This was a "complete" factorial experiment, which means that all eight possible combinations of the $2 \times 2 \times 2$ levels of the three factors were studied. Given the three factors, there are three pairs (*SQ*, *MF*), (*SQ*, *NA*), and (*MF*, *NA*) about which we can ask whether one factor modulates the effect of another. When you examine the data, or plot them, you will discover that the answer differs for the different pairs. One can also ask whether the interaction of any two of the factors is modulated by the third factor. If so, there is a "three-way interaction": just as a (two-way) interaction contrast is the difference between the

effects of one factor at different levels of another, so the three-way interaction contrast is the difference between the two-way interaction contrasts of two factors at different levels of a third. And just as a two-way interaction is symmetric, so is the three-way interaction. Thus, if *NA* modulates the interaction of *SQ* and *MF*, then each factor modulates the interaction of the two others. And if the three-way interaction contrast is zero, then none of the three factors modulates the interaction of the other two.

After you decide what the data from the numeral-naming experiment tell you about how the effects of the three factors combine, consider what your conclusions might suggest about the underlying process.

Table 1
Values of $\overline{RT}_{ijk}$ for five subjects and their mean in a $2 \times 2 \times 2$ choice-reaction experiment with digits as stimuli and digit names as responses. Values are in milliseconds.

Subject	NA_k	MF_j:	Familiar		Unfamiliar		Mean
		SQ_i:	Intact	Deg.	Intact	Deg.	
BN	2		300	314	300	326	354.1
	8		330	368	425	470	
DH	2		302	332	311	342	363.8
	8		339	396	415	473	
SS	2		329	354	353	369	382.0
	8		353	401	427	470	
AP	2		354	383	384	423	436.5
	8		399	468	500	581	
PM	2		363	405	396	439	469.9
	8		436	488	594	638	
Mean	2		**329.6**	**357.6**	**348.8**	**379.8**	401.2
	8		**371.4**	**424.2**	**472.2**	**526.4**	

Table 2
Main effects and interactions of factors SQ_i, MF_j, and NA_k in the data of Table 1, for five subjects, with means and standard errors (s.e.).

Subject	Main Effects			Interactions			
	$D_i(\overline{RT}_{i\bullet\bullet})$	$D_j(\overline{RT}_{\bullet j\bullet})$	$D_k(\overline{RT}_{\bullet\bullet k})$	$D_{ij}(\overline{RT}_{ij\bullet})$	$D_{ik}(\overline{RT}_{i\bullet k})$	$D_{jk}(\overline{RT}_{\bullet jk})$	$D_{ijk}(\overline{RT}_{ijk})$
BN	31	52	88	9.5	22	92	-5.0
DH	44	43	84	1.0	27	67	0.0
SS	33	46	62	-7.0	25	52	4.0
AP	54	71	101	11.0	41	72	2.0
PM	45	94	138	-3.5	6	120	-9.0
Mean	**42**	**61**	**95**	**2.2**	**24**	**81**	**-1.6**
s.e.	4	10	13	3.5	6	12	2.4

Notation

i, j, and k index the levels of the three factors SQ_i (stimulus quality, intact vs. degraded), MF_j (S-R mapping familiarity, high vs. low), and NA_k (number of alternative S-R pairs, 2 vs. 8). The mean reaction time for particular levels i, j, and k of the three factors is denoted $\overline{RT}_{ijk}$. When the mean is taken over levels of a factor, the index for that factor is replaced by a dot. D is a differencing operator; the subscript indicates the factor whose levels contribute to the difference. In this case, where there are only two levels, the result of applying the operator is a single number, the difference between mean RTs at the two levels. Thus, $D_i(\overline{RT}_{ijk}) = \overline{RT}_{2jk} - \overline{RT}_{1jk}$. A D operator with more than one subscript corresponds to applying the operator successively: $D_{jk}(\,) = D_j(\,D_k(\,)\,)$. Because the D operator is commutative, $D_{jk}(\,) = D_{kj}(\,)$. See Sternberg (1998b, Sections 14.3, 14.4) for more on the use of the D operator to describe interactions.

Appendix 2: Some Readings Relevant to RT Measurement and Interpretation

Excellent introductions to the use of RT in research on human information processing are provided by Pachella (1974) and Meyer, Osman, Irwin, & Yantis (1988). Books by Luce (1986), Townsend & Ashby (1988), Sanders (1998) and Welford (1980), are advanced treatments; the first two emphasize mathematical models. Reviews of basic RT phenomena can be found in Smith (1968), Keele (1986, pp. 2 - 20), Proctor & Vu (2003) and, for earlier work, Jastrow (1890) and Woodworth (1938, Ch. 14). Parts of Chase (1978) and Posner & McLeod (1982) and Massaro & Cowan (1993) are also of interest. Schweickert (1993) provides a helpful review, emphasizing theoretical ideas. For an entrée to models that attempt to explain both RTs and error frequencies, see Smith & Ratcliff (2004) and references therein. Skepticism is called for about claims of good fits and their implications; see Roberts & Pashler (2000). For tutorial treatment of some of the modeling issues in distinguishing parallel from serial processes, see Sternberg (1998a) and Townsend (1990). One view of the important issue of the "speed-accuracy tradeoff" is provided in Appendix 1 of the former (pp. 436-440) on "Error rates and the interpretation of reaction-time data". For a tutorial treatment of the method of additive factors, with discussion of a wide range of applications, see Sternberg (1998b). This method is placed in a more general context of the separate modifiability approach to identifying functional and neural processing modules by Sternberg (2001). For a discussion of mixture models, see Yantis et al. (1991). Van Zandt (2002) reviews many issues in the analysis of RT distributions; see also Roberts & Sternberg (1993) and Smith & Ratcliff (2004).

If you are interested in the early history of the subject, you will also enjoy the papers by and about Donders in Koster (1969), the proceedings of the Donders Centenary Symposium on Reaction Time, and also some of the papers by James McKeen Cattell (who did research at Penn) that have been collected in Cattell (1947). Hyman's (1953) study is perhaps the most impressive application of Shannon's (1948) information theory in psychology. Reports of some of the high points of recent years in the use of RT to learn about human mental processes can be found in the series of *Attention and Performance* volumes, published approximately every two years since 1967. These volumes also contain very useful tutorial reviews.

Readings

Cattell, J. M. (1947). *James McKeen Cattell, 1860-1944: Man of science.* Lancaster: Science Press.

Chase, W. G. (1978). Elementary information processes. In W. G. Estes *Handbook of learning and cognitive processes, Volume 5: Human information processing*. Hillsdale, NJ: Erlbaum. Pp. 19-90.

Hyman, R. (1953). Stimulus information as a determinant of reaction times. *Journal of Experimental Psychology, 45*, 188-196.

Jastrow, J. (1890). *The time-relations of mental phenomena. Fact and theory papers No. VI*. New York: N. D. C. Hodges.

Keele, S. W. (1986). Motor Control. In K. R. Boff, L. Kaufman, and J. P. Thomas *Handbook of perception and human performance, Vol. II: Cognitive processes and performance.* New York: Wiley.

Koster, W. G. (Ed.) (1969). *Attention and performance II. Acta Psychologica* 30

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Massaro, D. W. & Cowan, N. (1993) Information processing models: Microscopes of the mind. *Annual Review of Psychology, 44*, 383-425.

Meyer, D. E., Osman, A. M., Irwin, D. E., and Yantis, S. (1988). Modern mental chronometry. *Biological Psychology* 26, 3-67.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.) *Human information processing: Tutorials in performance and cognition.* Hillsdale, N.J.: Lawrence Erlbaum Associates. Pp. 41-82.

Posner, M. I. and McLeod, P. (1982). Information processing models - In search of elementary operations. *Annual Review of Psychology* 33, 477-514.

Proctor, R. W. & Vu, K-P. L. (2003) "Action selection". Chapter 11 in I. B. Weiner (Series ed.), A. F. Healy & R. W. Proctor (Vol. eds) *Handbook of Psychology. Volume 4: Experimental Psychology.* New York: Wiley. Pp. 293-316.

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

Roberts, S. & Sternberg, S. The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.) *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience.* Cambridge, MA : M.I.T. Press, 1993. Pp. 611-653.

Sanders, A. F. (1998). *Elements of human performance: Reaction processes and attention in human skill*. Erlbaum, Mahway, NJ.

Schweickert, R. (1993). Information, time, and the structure of mental events: A twenty-five year review. In D. E. Meyer and S. Kornblum (Eds.) *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience.* Cambridge, MA : M.I.T. Press, 1993. Pp. 535-566.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423 and 623-656.

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin* 69, 77-110.

Smith, P. L. & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27,* 162-168.

Sternberg, S. (1998a) Inferring mental operations from reaction-time data: How we compare objects. In D. Scarborough & S. Sternberg (Eds.), *An Invitation to Cognitive Science, Volume 4: Methods, Models, and Conceptual Issues*. Cambridge, MA : M.I.T. Press. Pp. 365-454.

Sternberg, S. (1998b) Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *An Invitation to Cognitive Science, Volume 4: Methods, Models, and Conceptual Issues*. Cambridge, MA : M.I.T. Press, 1998. Pp. 703-863.

Sternberg, S. (2001) Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica, 106,* 147-246.

Towsend, J. T. and Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge: Cambridge University Press.

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

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

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

Yantis, S. G., Meyer, D. E., and Smith, J. E. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin* 110, 350-374.

Van Zandt, T. (2002) Analysis of response time distributions. In J. Wixted (Vol. Ed.) & H. Pashler (Series Ed.), *Stevens' handbook of experimental psychology, Third Edition: Volume 4: Methodology in experimental psychology.* New York: John Wiley & Sons. Pp. 461-516.

Appendix 3: References for Appendix 1 of Sternberg (1998a) above

Ashby, F. G. and Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology* 38, 423-466.

McElree, B. and Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General* 118, 346-373.

Meyer, D. E., Irwin, D. E., Osman, A. M. and Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review* 95, 183-237.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.) *Human information processing: Tutorials in performance and cognition*. Hillsdale, N.J.: Lawrence Erlbaum Associates. Pp. 41-82.

Ratcliff, R. and Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review* 83, 190-214.

Schweickert, R. (1985). Separable effects of factors on speed and accuracy: Memory scanning, lexical decision, and choice tasks. *Psychological Bulletin* 97, 530-546.

Wickelgren, W. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica* 41, 67-85.

Appendix 4: Some Applications of the Additive-Factor Method Related to Issues in Clinical Psychology

Archibald, C. J. & Fisk, J. D. (2000) Information processing efficiency in patients with multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology, 22*, 686-701.

Azorin, J.-M; Benhaiem, P; Hasbroucq, T; Possamaie, C.-A. (1995) Stimulus preprocessing and response selection in depression: A reaction time study. *Acta Psychologica 89*, 95-100.

Exposito, J; Andres-Pueyo, A. (1997) The effects of impulsivity on the perceptual and decision stages in a choice reaction time task. *Personality & Individual Differences. 22*, 693-697.

Dickman, Scott J; Meyer, David E. (1988) Impulsivity and speed-accuracy tradeoffs in information processing. *Journal of Personality & Social Psychology. 54*, 274-290.

Rogers, T. B. (1974) An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica 38*, 205-213.

Readings for the RT Methods Session of the Proseminar

[1] Pages 2-20 (on basic reaction-time phenomena) in: Keele, S. W. (1986) "Motor Control". In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. II: Cognitive processes and performance.* New York: Wiley.

[2] Pages 436-440 (Appendix 1: Error Rates and the Interpretation of Reaction-Time Data) in: Sternberg, S. (1998) "Inferring mental operations from reaction-time data: How we compare objects". In D. Scarborough & S. Sternberg (Eds.), *An Invitation to Cognitive Science, Volume 4: Methods, Models, and Conceptual Issues*. Cambridge, MA: MIT Press.
The references for this appendix are given above.