Behavior Research Methods
2006, 38 (4), 569-573

Throughput: A simple performance index with desirable characteristics

DAVID R. THORNE
Walter Reed Army Institute of Research, Silver Spring, Maryland

Throughput is a corrected response rate measure giving the number of successes per unit of discretionary time. It is a simple but general index applicable to psychomotor, behavioral, and cognitive tasks in which response times are measured. This measure has several attractive features: (1) It allows comparisons to be made across various tasks in which speed and accuracy are meaningful and measurable, independently of temporal differences in hardware, software, and procedures; (2) under conditions in which both speed and accuracy decline (or improve), throughput will be a more sensitive index of performance than either alone will be; and (3) in those tasks in which the speed–accuracy trade-off phenomenon operates, throughput will tend to be relatively less variable than either component alone will be. The measure allows both behavioral and information-processing interpretations of data and may be useful as a simple composite index, a measure of effectiveness or of cognitive consistency in studies investigating performance degradation or enhancement.

Under conditions common to many psychological studies, the behavioral measures that are most sensitive to experimental manipulation are also the ones most sensitive to extraneous variables. If many of these extraneous variables are uncontrolled or uncontrollable, the data generated with such measures will be noisy. Conversely, measures that are stable across conditions or time tend to be relatively insensitive to random variables, but also to the independent variables of interest. Sensitivity and variability are usually inseparable.

A composite response measure (herein, called *throughput*) was originally devised for reasons of convenience specific to a particular set of experiments in our laboratory in which a computerized psychological test battery was used (Thorne, Genser, Sing, & Hegge, 1985). This measure subsequently proved to be more orderly than was anticipated, revealing effects that were larger and less variable than those shown by more conventional measures. Some of the reasons for this superiority have now been identified, and they suggest that the measure may be useful in a number of frequently encountered experimental situations.

Definition and Interpretation No. 1

Numerically, throughput is equal to the number of correct responses on a task, divided by the cumulative reaction times (RTs; both correct and incorrect). Conceptually, throughput may be viewed as a corrected rate measure giving the number of *hits* or *successes* per unit of discretionary time. Discretionary time is the time functionally available to and actually used by the subject for processing and responding, separated from any arbitrary or idiosyncratic delays imposed by the test mechanism or procedure itself. Choice of this measure was influenced by *response per opportunity* measures (Anger, 1956) and by the distinction between overall rates and running rates (Ferster & Skinner, 1957). Part of the rationale for its use is to maximize generality across tasks.

The term *throughput* is occasionally used in information theory as a synonym for channel capacity. It describes the amount of information that a given system can transmit *without error*, measured in bits per unit time. *Successes* or *hits per unit time*, as used here, express a similar concept.[1]

Definition and Interpretation No. 2

Numerically, the throughput measure is also equal to percent correct, divided by mean RT, times a scaling constant.[2] Conceptually, throughput may thus be viewed as a speed–accuracy product, where speed is the reciprocal of RT. This product may be considered a measure of useful or effective performance. In most tasks in which speed and accuracy are meaningful and measurable, subjects cannot respond both at 100% accuracy and at maximum possible speed simultaneously but can make various compromises (trade-offs) between the two, trading higher accuracy for lower speed or vice versa. This behavioral phenomenon can be represented as a shift in the operating point on a *speed–accuracy trade-off function* (Pachella, 1974; Pew, 1969). Trade-off functions have been widely studied and, under set conditions, are relatively stable for a particular individual and task. The operating point on the function

The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense. Correspondence should be addressed to D. R. Thorne, Department of Behavioral Biology, Division of Neuroscience, Walter Reed Army Institute of Research, 503 Robert Grant Ave., Silver Spring, MD 20910-7500 (e-mail: david.thorne@us.army.mil).

can be manipulated experimentally by instructions, deadlines, penalty/payoff protocols, and other variables (Wickelgren, 1977; Wood & Jennings, 1976). More typically, the operating point shifts back and forth from one group of trials to the next, due to random uncontrolled variables.

Generality

Because it separates available time from overhead time, throughput values are not determined or directly influenced by the execution times of a particular computer, operating system, programming language, interpreter, or algorithm or by the operating times of relays, slide projectors, disks or tapes, and so forth. Neither are they directly determined by procedural differences in the intertrial interval, the duration of a feedback interval or reinforcement, or the length of the retention interval in tasks of immediate or short-term memory. Thus, the measure allows comparisons to be made between different computer and instrumental implementations of a task. It also allows comparisons with paper-and-pencil versions where the full test duration may be considered discretionarily available. Its use also helps to avoid the confounding of dependent and independent variables that can create ambiguous or misleading conclusions, such as "the subjects responded twice as fast on a task with a 2-sec intertrial interval (ITI) as on one with a 4-sec ITI."

Sensitivity

Under conditions in which an external or experimental manipulation causes parallel, rather than reciprocal, changes in speed and accuracy (i.e., both degrade or both improve), throughput will be a more sensitive measure of performance than either alone will be. In all cases in which both speed and accuracy decline, their product must decline even more. If the individual speed and accuracy values were changed by some multiplicative factor m, the resultant throughput values would, of course, change by a factor m^2.

There are many real-world variables that can cause both speed and accuracy to degrade (e.g., alcohol, drugs, or sleep deprivation), to improve (e.g., training, stimulants, or incentives), or to oscillate (e.g., circadian variation). These variables are often studied across time, treatments, or sessions, rather than within single sessions. It is under these conditions and with these kinds of variables that the improved sensitivity of the throughput measure may be most useful.

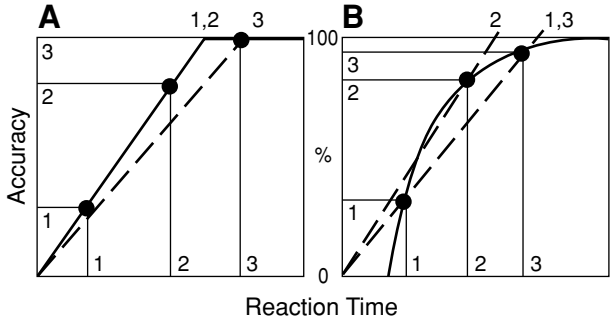

Figure 1. Schematized speed–accuracy trade-off functions. Filled circles indicate the operating point at three different points in time. The slopes of the numbered dashed lines correspond to throughput values at these operating points. In panel A, two of the three (dashed) lines are coincident with the function itself and, hence, appear as a solid line.

Stability

As a speed–accuracy product or an accuracy–latency ratio, the throughput index is related to the slope of the trade-off function and, therefore, is relatively unaffected by movements of a point along that function. More accurately, throughput is equal to the slope of a line connecting the origin to the operating point on the trade-off function. When the trade-off function approximates a straight line passing through the origin, as idealized in panel A of Figure 1, throughput will be completely isolated from fluctuations in accuracy and speed, up to the zero-slope asymptote.

When the function has a nonzero intercept, such as an irreducible minimum RT, or exhibits curvilinearity, as shown in panel B of Figure 1, throughput values will still be relatively constrained, in comparison with the corresponding fluctuations in accuracy and speed, for mathematical reasons elaborated below. Part of the rationale for using the throughput measure is that under approximately stationary conditions, particularly across trials within a session, throughput will tend to be less variable than speed or accuracy. This will hold for all cases in which accuracy is an increasing function of RT, regardless of the shape of the trade-off function. The function may be linear, convex, concave, continuous, or discontinuous, as long as slope(s) remain positive. It will not hold for any case in which the local slope is negative.[3]

The expected variability of the throughput measure is analytically specifiable. If we use variance (σ^2) as the index of variability and regard throughput (z) as a function of two random variables, accuracy (x) and RT (y), the variance of z will be related to the means and variances of x and y, to their correlation (ρ), and to the partial derivatives of the function with respect to each, according to the following equation:

$$\sigma_z^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\rho\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\sigma_x \sigma_y.$$

If, as here, the function is a simple ratio $z = x / y$, then

$$\frac{\partial f}{\partial x} = \frac{1}{y} \quad \text{and} \quad \frac{\partial f}{\partial y} = \frac{-x}{y^2}.$$

Substituting terms, evaluating at the mean (μ), and factoring gives

$$\sigma_z^2 = \left(\frac{1}{\mu_x}\right)^2 \left[\sigma_x^2 + \left(\frac{\mu_x}{\mu_y}\right)^2 \sigma_y^2 - 2\rho\left(\frac{\mu_x}{\mu_y}\right)\sigma_x \sigma_y\right].$$

This equation has no particular computational use. The important thing to note is that the third term in the summation is negative and that the variance of z will be reduced by

an amount proportional to twice the correlation coefficient. A high positive correlation between latency and accuracy is a defining characteristic of the trade-off phenomenon. Hence, where trade-offs operate, the variability in the combined throughput measure will be less than that otherwise expected for random variables. It is easier to conceptualize the direction and maximum size of this effect by selecting metrics where the two means are unity and then comparing the two cases in which the correlation is either zero or one. For uncorrelated random variables,

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

whereas when the correlation approaches one (e.g., Figure 1A)

$$\sigma_z^2 = \left(\sigma_x - \sigma_y\right)^2,$$

thus explaining the reduced variability.

Examples

An indication of the relative sensitivity and variability of the three performance measures is provided in Table 1. These data were derived from a 72-h sleep deprivation experiment involving 7 human subjects and eight performance tasks (Thorne, Genser, Sing, & Hegge, 1983). The tasks included visual search and recognition, mental arithmetic, logical reasoning, sustained attention, short-term memory, and pattern recognition. For each task, each subject's performance scores were first converted to percentage difference from baseline and then averaged. Performance declined with increasing sleep deprivation, and the maximum performance decrements for each of the tasks were used to construct the table.

The first row of numbers shows throughput to be more sensitive than accuracy or speed, since it exhibits the greater decrement. The second row of numbers shows throughput to be less variable than accuracy or speed. Even if all three measures had the same absolute range of variation, the smaller relative variability of throughput would make it the preferred measure in this case, as is also shown by the coefficients of variation in the bottom row.

Only the relative variability in throughput was expected to be smaller. The smaller absolute variability obtained suggests that yet another factor was operating. The probable source can be identified, although not completely quantified. With RTs, the mean and standard deviation are positively correlated, usually by a simple direct proportionality (Chocholle, 1945). With accuracy, the mean and standard deviation are negatively correlated, with variability greatest near chance and decreasing as the mean approaches 100%, although the form of the inverse relationship is not generally specifiable. When the two means are linked by the speed–accuracy trade-off effect, the two sources of variability are also linked but change in opposite directions and, thus, tend to compensate for one another.

This greater sensitivity and/or reduced variability will increase the computed statistical significance values of experimental differences, or how quickly they may reach a given alpha level. This is indicated in Table 2 with data from a study by Kay et al. (1997). After a baseline training day, subjects were given either a placebo or 100 mg of diphenhydramine, a drug and dosage known to have sedative effects, and were tested 90 min later on tasks from the CogScreen battery (Kane & Kay, 1992; Kay, 1995). The table shows both significant and nonsignificant *p* values for the three performance measures across four cognitive tests, with throughput typically showing the largest effect.

Williams (1994) evaluated the relative sensitivity of 24 dependent variables to four classes of independent variables (drugs, exercise, heat stress, and sleep deprivation), using 13 cognitive tests, and concluded that

> across all the dependent measures, the most sensitive measure was [throughput]. This single measure of performance produced substantially more significant findings than the standard measures used in analyzing this type of data . . . 57% more than Reaction Time and 267% more than Percent Correct. (p. 18)

In a factor-analytic study evaluating the construct validity of four cognitive functions (attention, processing efficiency, cognitive flexibility, and working memory), using tasks from the ANAM battery (Kane & Kay, 1992; Reeves, Kane, & Winter, 1995), Short, Cernich, Wilken, and Kane (in press) concluded that

> choice of a specific index for a given construct is currently a function of historical practice. . . . However, we believe there are strong psychometric reasons for relying predominantly on throughput values. Response latencies and accuracy scores may be helpful for interpreting why throughput behaves as it does, but the information contained in throughput best reflects the processes at work in the ANAM tasks we evaluated.

Limitations and Applications

Throughput is a measure of useful or effective performance, an index having both theoretical and practical implications and applications. It is more of a descriptive measure than an analytic one. Although it serves a data reduction function by combining two measures into one, it thereby confounds them. It cannot replace accuracy or RT for purposes of detailed analysis.

The throughput measure separates discretionary time from overhead time and, thus, is mathematically (not necessarily behaviorally) independent of the usual temporal variations that arise from different implementations of a given task. However, both throughput and RT measures combine stimulus-processing time with response execution time and cannot distinguish between the two without additional information and assumptions.

Table 1
Effects of ~72 h of Sleep Deprivation Shown by Three Performance Measures

Statistic	Accuracy	Speed	Throughput
Mean change from baseline	−35%	−66%	−77%
Range of variation	53%	31%	26%
Coefficient of variation	52%	18%	12%

Note—Averaged across 7 subjects and eight tasks.

Table 2
***p* Values for Diphenhydramine* − Placebo† Differences on Four Tasks Shown by Three Performance Measures**

Task	Accuracy	Reaction Time	Throughput
Continuous performance	.010	.063	.002
Shifting attention instruction	.062	.117	.015
Mathematical processing	.115	.136	.003
Running memory	.008	.000	.000

Note—Subanalysis of data from Kay et al. (1997), Table 4, courtesy of the authors. *100 mg, $n = 32$. †$N = 33$.

Throughput can give misleading values if the subject is deliberately making errors, responding randomly or by some arbitrary rule, dozing off, or otherwise changing the nature of the task. If the numerator approaches chance or the denominator approaches or falls below the *minimum RT* for the given task, the measure acquires a different meaning. Similar qualifications hold for speed and accuracy. It is the experimenter's responsibility to ensure that the subject is performing the task as intended or to detect departures from the intent.

Throughput is a general performance index applicable to a large number of psychomotor, behavioral, and cognitive tasks—specifically, those that are instrumented or computerized to provide RT measurements. It is functionally similar to a measure frequently used with paper-and-pencil tasks: number correct per qualified time unit. This qualifier may not be stated but usually means *a period of time fixed by the experimenter (e.g., 30 min) and wholly available for subject responding*. However, throughput can give the same sort of measure while allowing manipulations that are not practical or possible with most paper-and-pencil tasks, such as experimenter-controlled pacing, subject-controlled termination, and/or termination by item count, as well as controlled interitem intervals, retention intervals, feedback durations and delays, and other intervals that are *not available for subject responding*. It allows for these without confounding the measure and the manipulation, as would be the case for number correct per total time, both available and not.

Because it is multiplicative, throughput will be particularly sensitive (and variable) to manipulations that cause speed and accuracy to change in the same direction. Such shifts correspond to changes in the slope or shape of the trade-off function itself. Note that such parallel changes in speed and accuracy may be called *gains* or *losses*, but not *trade-offs*. Examples of variables that typically produce such changes are sleep deprivation, circadian rhythms, alcohol, drugs, and training.

Throughput will be rather insensitive (and stable) to manipulations, variables, and phenomena that cause speed and accuracy to change in opposite directions (e.g., trade-off effects). Such effects generally correspond to shifts in the subject's criterion operating point on the current trade-off function. Although these shifts may be induced by experimental manipulations (the most common of which may be the instruction to respond both as rapidly and as accurately as possible), usually they are neither conscious on the subject's part nor deliberate on the experimenter's part but are attributed to unidentified random variables. It is these kinds of random variables to which throughput will be relatively insensitive.

In situations in which speed changes but accuracy remains the same or vice versa, the throughput measure will be "safe," in the sense that it will detect the effect, but will be no more or less sensitive than the component measures themselves. One example of such would be an overlearned task with accuracy constant at 100%, where throughput would simply be inversely proportional to RT.

Although the stability (but not the generality and sensitivity) advantages of the throughput measure arises from the existence of the speed–accuracy trade-off phenomenon, the measure would be of no particular use to those interested in obtaining or studying trade-off functions themselves. However, since the function may be linear, curvilinear, or even discontinuous, those using throughput as a performance measure need not know the shape of the function in order to take advantage of its existence. They do need to make two simple assumptions: (1) Some form of trade-off is operating, and (2) the operating point varies over only a moderate or limited range during a typical single test session. Both assumptions are reasonable with the tasks and variables most often used in performance testing and with the degree of experimental control employed in most laboratory studies. Indeed, it usually requires considerable experimental effort and ingenuity to force the operating point over the full range of the function (Wickelgren, 1977; Wood & Jennings, 1976).

The data in Table 1 show percent changes in throughput, relative to rested baseline. In theory, absolute throughput values or their channel capacity transforms (see note 1) could be used to rank order qualitatively different tasks on a single dimension. It remains to be seen whether such a ranking scale would have any practical or theoretical utility. A more conventional use would be for scaling and quantifying the effects of parameter changes or variable manipulations in qualitatively identical tasks in which trade-offs are known or expected to occur.

In the introduction, it was said that sensitivity and variability are usually inseparable. Not even the throughput measure can separate the two within a single experimental setting if the source of variation arises from the actions of a single variable or phenomenon. However, across situations, it can weight the two differentially. Typically, throughput appears to provide increased sensitivity to several kinds of systematic variations that operate across sessions and decreased sensitivity to some of the random variations that operate within sessions. This would seem to lend it best to the study of time-acting variables and repeated measures designs.

As a corrected rate measure, a speed–accuracy product, and a channel capacity analogue, throughput allows both traditional behavioral and information-processing interpretations of findings. As a simple performance measure, the generality, stability, and sensitivity of throughput recommend it for consideration in studies of performance degradation and enhancement in humans or lower animals.

REFERENCES

Anger, D. (1956). The dependence of interresponse times upon the relative reinforcement of different interresponse times. *Journal of Experimental Psychology*, **52**, 145-161.

Chocholle, R. (1945). Variation des temps de réaction auditifs en fonction de l'intensité à diverses fréquences. *L'Année Psychologique*, **41**, 65-124.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York: Appleton-Century-Crofts.

Kane, R. L., & Kay, G. G. (1992). Computerized assessment in neuropsychology: A review of tests and test batteries. *Neuropsychology Review*, **3**, 1-117.

Kay, G. G. (1995). *COGSCREEN: Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Kay, G. G., Berman, B., Mockoviak, S. H., Morris, C. E., Reeves, D., Starbuck, V., et al. (1997). Initial and steady-state effects of diphenhydramine and loratadine on sedation, cognition, mood, and psychomotor performance. *Archives of Internal Medicine*, **157**, 2350-2356.

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* (pp. 41-82). Hillsdale, NJ: Erlbaum.

Pew, R. W. (1969). The speed–accuracy operating characteristic. *Acta Psychologica*, **30**, 16-26.

Reeves, D., Kane, R., & Winter, K. (1995). *Automated neuropsychological assessment metrics (ANAM): Test administrators guide Version 3.11* (Rep. NCRF-95-01). San Diego: National Cognitive Recovery Foundation.

Short, P., Cernich, A., Wilken, J. A., & Kane, R. (in press). Initial construct validation of frequently employed ANAM measures through structural equation modeling. *Archives of Clinical Neuropsychology*.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1983). Plumbing human performance limits during 72 hours of high task load. In *Proceedings of the 24th DRG Seminar on the Human as a Limiting Element in Military Systems* (pp. 17-40). Toronto, Canada: Defense and Civil Institute of Environmental Medicine.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1985). The Walter Reed performance assessment battery. *Neurobehavioral Toxicology & Teratology*, **7**, 415-418.

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

Williams, D. (1994). *The relative sensitivity of dependent measures and cognitive tests determined by hypothesis testing* (Rep. 95-44). San Diego: Naval Health Research Center.

Wood, C. W., & Jennings, J. R. (1976). Speed–accuracy tradeoff functions in choice reaction time: Experimental designs and computational procedures. *Perception & Psychophysics*, **19**, 92-102.

NOTES

1. Throughput, as defined here, can be formally transformed to channel capacity by multiplying it by the binary logarithm of the number of response alternatives. For the case of a two-choice task, the scaling factor is unity, so hits per unit time and bits per unit time are numerically identical. This (log) transform is of no particular advantage when one compares different tasks with the same number of alternatives or the same task with itself under different conditions, since it simply scales each equally. It may be useful, however, for information-theoretical interpretations across tasks with different numbers of stimulus–response alternatives.

Note that this conversion differs from the practice of plotting mean or correct-only RTs against log alternatives without regard to accuracy itself. These RTs are bought at the cost of a particular (often undisclosed) error rate *and* error RT, muddying interpretation and making the practice questionable.

2. Throughput units are typically reported as correct responses per minute of actual responding (cumulative RTs). The scaling constant converts percent to number and RT to minutes. For mean RTs in seconds, the scaling factor would be $60/100 = 0.6$; for milliseconds, 600.

3. Subtracting chance accuracy from the numerator and minimal RT from the denominator yields a measure that may be useful in some applications. Both constants attenuate change and, thereby, reduce sensitivity. Although this transform is no longer interpretable as a simple hit rate, it theoretically forces the trade-off function to pass through the origin. Some practical problems with this measure are the following: (1) In many tasks, the appropriate minimal RT is not easily specifiable; (2) in some tasks, the expected or chance value depends on arbitrary definitions or can vary with different subject strategies; and (3) amplified variability and spuriously low or even negative values can arise if tasks use a very small number of trials and randomization with replacement.

(Manuscript received October 10, 1984;
revision accepted for publication September 1, 2005.)