On Performance Measurement in Psychology and Other Fields

Yves Guiard

On Performance Measurement in Psychology and Other Fields

Yves Guiard

Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS,

Université Paris-Saclay, France

yves.guiard@lri.fr

Laboratoire de Traitement et de Communication de l'Information, Télécom Paris,

Institut Polytechnique de Paris, France

yves.guiard@telecom-paris.fr

Abstract

The concept of quantitative performance has been used increasingly outside work management, its main field of origin, since the advent of the industrial era, pervading not just experimental psychology and many domains of science and engineering, but virtually all sectors of social life. Surprisingly, the key defining characteristic of performance measures seems to have systematically escaped notice: a performance is a numerical score subject to a deliberate extremization (i.e., minimization or maximization) effort exerted by a human agent against the resistance of a limit. Because of this characteristic performances must be recognized to constitute measures of a very special kind, where the numerical is marked axiologically. The paper contrasts the extremized scores of performance measurement with the optimized measures of feedback-controlled, regulated systems. In performance measurement the best numerical values are extrema, rather than optima, and the function that links the axiological value to the numerical value is strictly convex, rather than strictly concave.

One-dimensional performance measurement is analyzed in the extremely simple case of spirometry testing, where forced vital capacity, a measure of respiratory performance, is shown to be determined by the interplay of two variables, neither of which can be directly measured: the maximization effort, which varies haphazardly from trial to trial, and the patient's total lungs capacity, a personal upper bound, whose inductive estimation is the goal of spirometry testing. The paper shows that the magnitude of the estimation error decreases linearly with the magnitude of the patient's effort, explaining why respirologists so strongly urge their patients to blow as hard as they can into the spirometer.

The paper then turns to two-dimensional performance, analyzing distributional data from a psychology experiment on speeded aimed-movement. The variation of the speed/accuracy balance is shown to entail systematic changes in the markedly asymmetrical shapes of movement time and error distributions: The stronger the directional compression effect observable on one performance measure, the weaker this effect on the other.

These observations are hard to reconcile with the traditional view that performance measures are random variables and raise doubts on the suitability of the classic descriptive tools of statistics, whether parametric or nonparametric, when it comes to the decidedly special case of performance data. One possible direction for a more appropriate statistical approach to performance data is tentatively outlined.

Keywords

Performance; measurement; axiology; extremization effort; capacity of performance; spirometry; speed-accuracy trade-off; speeded aimed movement; quasi-deterministic variable; statistical descriptors.

Contents

1. *Quantitative Performance : A Stowaway Concept*

This paper is about quantitative *performances*, whose measurement is widespread not just in experimental psychology and many other fields of academic research, but also in virtually all sectors of social life. Despite a high frequency of occurrence, in general the concept is used more or less thoughtlessly with its meaning taken for granted. We start with a definition making explicit one defining characteristic of performances that seems to have escaped so far the attention of both practitioners and theorists of performance measurement. While the word "performance" is polysemous, as can be checked in any general dictionary, here we focus on the quantitative sense of the word, namely, performance as a numerical score.

Definition: *A performance is a measure subject to a deliberate minimization or maximization effort exerted by a human agent against the resistance of a limit, a lower- or upper bound, respectively*.

Below we will recurrently use the neat example of track and fields athletics, where runners try their best to obtain as *low* time marks as possible while throwers and jumpers try their best to obtain as *high* length marks as possible. These marks, like the vast majority of performance scores, are numerical in essence, constituting genuine *measures* in the strict sense of classic measurement theory (Hölder, 1901; Michel, 2005).

The demand that performers do their best to either minimize or maximize a certain numerical score is the hallmark of performance-testing situations. However, it should be noted from the outset that extreme deviations from medians are less likely in performance testing than anywhere else. The reason is because the performer's effort is typically exerted against the resistance of a limit working as a global attractor and a local repeller, so that a performance score is quite unlikely to fall a long way away from that limit (Gori, Rioul, & Guiard, 2018; Gori & Rioul, 2019; Guiard, Olafsdottir, & Perrault, 2011; Guiard & Rioul, 2015). For example, Usain Bolt's record-breaking time mark of 9.58 s on the 100 meters is indeed an extremum (i.e., an empirically observed minimum), yet statistically speaking it is certainly *not* an extreme value. The numerical values expected in performance testing must tend to cluster against the personal limits of performers and so the best performances (sample minima in minimization tasks, sample maxima in maximization tasks) must be expected to fall rather *close* to the medians.[1] Thus extreme value theory (e.g., Gumbel, 1935), all about

[1] Unfortunately it is typically impossible to estimate median performance in the remarkably abundant though wildly trimmed data sets from athletics, which only include what they call "best" marks. For example in the list

extreme deviations from medians, is essentially irrelevant to the problem of extremized performance.

The role of the effort in performance testing can be understood using the analogy of Newton's third law, which says that every force in mechanics encounters a reaction force of equal magnitude and of opposite direction. Absent the resistance opposing the minimizing or maximizing of a measure, attributable to the existence of a lower or upper bound located somewhere on the measurement continuum, it would make no sense to ask people to try their best to extremize the measure.

The fact that performance testing demands an extremization effort on the part of the performer seems so obvious that more often than not it goes without saying. In athletics there is no need to recall jumpers or throwers that the greater the distance they will cover, the better. Nor is it useful to recall business people that the higher the profit they will make, the better. There are some counterexamples, however. One is spirometry testing, to which we will return shortly, which measures respiratory performances in the form of maximized air volumes or flows: testees are always explained that they must blow into the spirometer as hard as they possibly can. Likewise, it is common practice in psychology laboratories to instruct participants quite explicitly to perform the proposed experimental task as fast and as accurately as they can, that is, to minimize both a time measure and an error measure. Thus sometimes the consensus on both the requirement of an effort and its direction is explicitly recalled to the performer. More often than not, however, it remains entirely tacit and this is because in performance testing it is generally perfectly obvious to everyone whether the key measure should be minimized or maximized.

The equivalence class covered by the above definition has an impressively large extension, cutting across most fields of science and engineering and most sectors of social life. For example, beside the time duration of an athletic run (to be minimized) or the length of an athletic jump (to be maximized), the equivalence class also includes response latencies and task-completion times which experimental psychologists ask their participants to minimize, the volume of air blown into a spirometer which respirologists ask their patients to maximize, the time spent at making an item in a factory chain which work managers ask workers to minimize, the amount of money earned with a market share which share owners

available on http://www.alltime-athletics.com/men.htm, a website that tabulates the 3,374 "bests" achieved since 2007 by no fewer than 360 sprinters, no time mark on the 100 meters exceeds the arbitrary value of 10.09 s (August 17, 2020).

try to maximize, the temperature of super-conductivity and super-fluidity which solid-state physicists try to maximize, the number of Nobel prizes won by a university which managers of academic institutions want to maximize, etc. This long and rather motley list ignores innumerable social, engineering, and scientific fields where performance measurement is no less overwhelmingly present, such as games of all sorts, chemistry, pharmacology, administrative science, mathematics, algorithmics, or computer science.

Surprisingly, despite its impressive frequency of use, the specific quantitative sense of the word performance to which attention is called here is manifestly overlooked by general dictionaries. More surprisingly yet, scientists who use the concept seem to be unaware of the fact that performances involve a maximization or minimization effort against a limit.[2] Thus the concept looks somewhat like an intellectual stowaway. It sounds familiar, but because it lurks in the background it is not clearly understood and apparently its most salient characteristic escapes the scrutiny of the analytical mind.

The concept of performance is of considerable import for experimental psychology. Since 1975 the American Psychological Association has been dedicating one section of its *Journal of Experimental Psychology* to the theme of *Human Perception and Performance*. One possible understanding of the term performance in this section title is that it refers to the motor side of human behavior—such as movement, language expression, action in general—thus pointing to a domain of inquiry complementary to that of perception and cognition. But this cannot be the whole story, as experimental psychologists measure human performances in every conceivable sort of tasks including ones designed to cast light on the mechanisms of not just action, but also perception, cognition, and memory. For example, reviewing one full issue of *JEP: HPP* (N° 41(5)), this writer found that 23 of the 25 articles that compose the issue (i.e., 92%) report to have considered at least one measure of extremized performance (see Annex 1).

[2] A Google Scholar search with the expression “concept of performance” was conducted to get a sense of the areas of Academia where the concept has been subject to reflective thought. The first 100 results were considered and classified. Seventeen references to arts studies (drama, music, verbal arts) and three references to linguistics had to be put aside, the term performance being taken in these fields in different senses (public entertainment and overt manifestation of linguistic competence, respectively). The remaining 80 studies were found to consider the sense of interest here. While only one minor item was found in the psychology field, the vast majority (66 of 80 = 83%) turned out to belong to the field of *management* (e.g., work management, business management, financial management, school management, administrative science). A cursory exploration of the listed publications suggested that they do refer to performances qua measures but apparently with no awareness of the fact that the measures in question are subject to extremization efforts (date of search: April 8, 2020; settings: any time, sort by relevance, all languages, patents not included, citations not included).

2. Genuine Measures

Measurement is consubstantial with performance testing. Imagine that for some reason measurement were suddenly prohibited in track and field athletics. Everything that happens in stadiums would become pointless, and so would all institutions such as the International Association of Athletics Federations, whose role is to organize the competition and to officially record quantitative performances. But athletics is no exception: in their vast majority, performances are, strictly speaking, *measures*. To make this point clear, a quick reference to measurement theory is in order.

According to classic measurement theory along the lines of Hölder (1901; see Michell & Ernst, 1996), measurement is just an instance of rational inquiry, one specifically aimed at discovering magnitudes on definite attributes of things (Michell, 2014). Specifically, measurement consists in procedures aimed to discover the quantitative relation linking a magnitude on a dimension of interest to a conventionally fixed unit of magnitude on that same dimension, and the quantitative relation to be discovered is no other than a real number (Michell, 2005).

Unlike S.S. Stevens's popular operationist theory, which holds that measurement is a (possibly arbitrary) operation just consisting in the "assignment of numerals to things or events according to rules" (Stevens, 1946 p. 667), and unlike Russell's representationist theory (see Michell, 1997, 2005), which postulates that numbers and attributes of things exist in separate spheres of reality, hence forcing measurement to be defined as a mapping of the former to the latter, the classic theory holds that numbers are objective relational characteristics of the real world that measurers have to *find out* (Michell, 2005). This philosophically realist perspective on measurement, to which this writer subscribes whole heartedly, has brought to the forefront the empirical question whether of not attributes of things intrinsically possess quantitative structure, that is, whether or not measurement is at all possible. In this philosophical research stream, accordingly, considerable efforts have been made to try to spell out the necessary and sufficient conditions for measurement (Hölder, 1901; Krantz, Luce, Suppes, & Tversky, 1971; Michell, 2014).

Below we will not have to bother about the so-called performance measures whose measurability has been controversial, such as those which form the matter of psychophysics since Fechner (1860) and, more generally, of psychometrics, a recurrent target of Michell's (2008) critical perspicacity. In fact only a negligible minority of the scores considered in the

sphere of performance measurement, whether scientific or else, actually raise a measurability concern. Not only is it the case that performance measurement nearly always rests on well-defined, if not strictly standardized procedures, it nearly always addresses simple physical dimensions such as time durations, lengths, volumes, temperature, or discrete counts.

3. Axiologically Colored Measures

Measurement leaves no room for judgments of value such are good and bad, better and worse, best and worst. It only considers magnitudes, and so the comparison between two measures involves nothing beyond the symbols "=" or "<", even though these symbols are often translated into ordinary language using various metaphors such as the altitude of numbers (a numerical value is often said to be said "higher" or "lower") or their size (a numerical value can be said to be "larger" or "smaller"). A number being just a magnitude, with no axiological connotation, it would make no sense, except for a numerologist, to ask whether 4 is preferable over 3. The statement that $4 > 3$ leaves no room for any of the questions that constitute the subject matter of axiology, the philosophical field that inquires into the 'goodness' of things from a diversity of viewpoint such as ethics, esthetics, hedonics, and utility (e.g., Hart, 1971).

Having acknowledged that the vast majority of performances are measures in strict parlance, we must now emphasize that performances are very *special* measures, in that they all have a manifest axiological coloration. In experimental psychology, for example, it is common to say that a performance score has "improved", say with practice, or that it has "impaired" in such and such condition. Likewise the minimum value of a sample of response-time measures is commonly called the "best" value. The same is true, in spirometry testing, of the maximum of a sample of measures of forced vital capacity recorded in a session: the maximum is often called the "best" value. These expressions have an unmistakable axiological coloration, which we must try to understand.

At first sight the irruption of axiology in the domain of measurement, which needs to get rid of value judgments, is a puzzling characteristic of the subclass of measures we call performances. But on second thoughts there is no paradox. It is the direction in which the measure is expected to be extremized by the performer that adds an axiological connotation to the symbols ">" and "<". If the measure is supposed to be minimized by performers (as, e.g., a run time, a response time, or an error rate), then the lower the numerical value, the better axiologically speaking; if, alternatively, the measure is supposed to be maximized by

performers (as, e.g., the number of correctly recalled items in a memory experiment, or the volume of expired air blown into the spirometer), then the higher the numerical value, the better.

Although axiology, a domain of philosophical inquiry, is very much concerned with subjective value judgments (Hart, 1971), it is important to note that the sort of axiology that colors performances involves no subjectivity whatsoever. The common observation of an "improved" or "impaired" measure of performance reflects nothing but the implicit recognition that, the measure being pressurize by an effort whose direction is known, the observed difference is in the desirable direction or in the opposite direction. The statement that the shorter an *RT* measure, the better involves no subjective judgment—it is true by definition.

4. Deliberately Biased Measures

Performance measures must be recognized to differ in another fundamental respect from the sort of measures that have constituted the main focus of measurement theory. Measurement theory typically considers the case of a single agent, namely the measurer in the face of nature, and in the measurement process nature is assumed to remain passive. In contrast, the measurement of a performance always involves *two* agents in tight connivance with each other, namely, the measurer in charge of the measurement procedure and the performer in charge of the minimization or maximization pressure.

To illustrate the contrast between an ordinary measure and a performance measure let us return to the case of athletics and compare two measures both taken on August 16, 2009 during the World Athletics Championships in Berlin, during which sprinter Usain Bolt famously broke the world record on the 100 meters. One measure was the time mark of Bolt, 9.58 second, the best ever recorded. But during the course of the all-important run *wind speed* was also discreetly measured,[3] yielding an average speed of +0.9 m/s. Anemometry is an instance of measurement of the ordinary kind, where care is taken by the measurer to avoid influencing the measured magnitude. There is every reason to believe that wind speed would have been the same absent the measurement device, but this is not the case of Bolt, who for sure would have *not* run absent the timing device. In connivance with the measurers, Bolt's

[3] +2m/s is the official upper limit above which the performer will be considered to have benefited from significant wind assistance, precluding the registration of any record.

task was to *struggle* against the timing device to obtain as short a time mark as possible. What we face here—and in performance testing in general—is a striking departure from the principle that measurement aims to discover magnitudes without interfering with them.

While athletic is all about the measurement of genuinely quantitative attributes of precisely defined human actions, the intriguing fact is that the agents subject to the measurement are officially supposed to try their best to *bias* the measure in the direction of lower or higher numerical values. The tested object being a human provided with such a considerable degree of control over the numerical result, how can a non-arbitrary measure obtain?

One central thesis of this paper is that performance measures are determined by the interplay of two background quantities, neither of which can be measured. The performance obviously depends on the magnitude of the performer's effort, which we must suppose to vary erratically from trial to trial; but that effort is not exerted in the void—it encounters the resistance of a lower or upper bound, characteristic of the performer, a personal constant subject to considerable between-individual variability.[5] Thus any attempt to understand a performance must take into account *three* variables, not just (*i*) the performance measure proper, but also (*ii*) the intensity of his/her minimization or maximization *effort*; and (*iii*) the performer's *capacity of performance*, his/her current limit.

To progress in the understanding of the relation borne by these three variables, in Section 6 below we will turn to spirometry, a remarkably simple and enlightening instance of performance measurement.

5. Extremization vs. Optimization

The axiological value of things—i.e., their degree of 'goodness', to use the terminology of philosophers specializing in axiology, or valuation theory (e.g., Hart, 1971)—is definitely not measurable. Notice, however, that we are now considering the axiological value of numerical values. Perhaps the language of mathematical functions may help us describe the relation borne by the axiological and the numerical aspects of performance measures.

[5] It is only at a relatively short time scale (e.g., during a testing session) that the capacity of a performer can be considered a constant. In the long run performance capacities improve with development, are affected by practice, doping, and health factors, and decline with aging.

Assuming that the axiological is dependent on the numerical, and letting $f(x)$ denote the axiological value of a numerical value x, we can write

$\forall(x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ such that $x_2 \geq x_1$,

either $f(x_1) \geq f(x_2)$ *minimization contexts* (1)

or $f(x_1) \leq f(x_2)$. *maximization contexts* (2)

In words, the function is decreasing in minimization contexts, where the lower the number, the better the performance; and it is increasing in maximization contexts, where the higher the number, the better the performance. Formulas 1 and 2 may be viewed as just formal definitions of what we call performance measures.

One immediate consequence of these inequalities is that there always exists on the x axis one extreme value—known as the 'best' value, sometimes as the 'record'—whose axiological value is unrivalled. For example respirologists, following the recommendations of the international standard of spirometry, discard all their measures of respiratory performance but the best, the session's maximum (see Section 6). Likewise, the current value of the world record on the 100 meters enjoys a very special degree of popularity: All fans of athletics have memorized the record value (9.58 s) along with the name of the record holder (Bolt), but one can safely conjecture that few of them would be able, on request, to cite the *second* best time mark of all times (9.69 s, obtained in 2009 by US sprinter Tyson Gay), not to mention the third best, fourth best, etc., time marks.

In fact we may take one more step and say that the function that relates the valuation and the numerical value of performances is *convex*.

$\forall(x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ such that $x_2 \geq x_1$ and $\forall d > 0$, we have

either $f(x_2 + d) - f(x_2) \leq f(x_1 + d) - f(x_1)$, *minimization contexts* (3)

or $f(x_2 + d) - f(x_2) \geq f(x_1 + d) - f(x_1)$. *maximization contexts* (4)

Returning to the example of athletics running, there is little doubt that the axiological value of the conventionally fixed time unit, namely the 100th of a second on the 100 meters, decreases monotonically upward on the time continuum. When someone is timed in 9.57 s, thus improving the current world's record by just one unit, that will be a major international event with presumably considerable press coverage. But suppose now that tomorrow a US sprinter obtains a 9.68 s mark, thus improving by the same 0.01 s over the *second* best time

mark of all times, Tyson Gay's record of 9.69 s, that will be a far less considerable event, even though that means braking the national US record. If Jimmy Vicaud, currently the holder of the French record (9.86 s), which ranks 109th in the list of all times performances, runs the 100 meters in 9.85, the event will enjoy still less coverage.

To help appreciate that the convexity of the relation linking the axiological to the numerical is a hallmark characteristic of performance measures, it is useful to consider another quite different family of measures, those subject to more or less automatic feedback correction in the service of some homeostasis process, for example the internal temperature of the human body, maintained roughly constant by a complex physiological machinery. In this case we observe the existence on the numerical continuum of an *optimal* region, a numerical interval most favorable to the organism's health, located around 37°C.

Notice in passing that the statement of optimality implies recourse to axiology: from a certain viewpoint (with regard to health in the chosen example) the numerical values falling in the optimal region are indeed best. And here again, just as is the case with performance measures, axiology has nothing to do with subjectivity: It is an objective fact that the larger the departure from the optimum, the more dangerous for the organism. But now comes the important difference: In this other family of measures the dependency of the axiological upon the numerical is *concave*. Continuing to conceptualize the axiological dimension of the measure as a function of its numerical value, we may write

$$\forall (x_1, x_2) \in \mathbb{R} \times \mathbb{R} \text{ such that } x_2 \geq x_1 \text{ and } \forall d > 0,$$

$$|f(x_2 + d)| - |f(x_2)| \geq |f(x_1 + d)| - |f(x_1)|. \qquad (5)$$

In words, the health impact of a given temperature variation increases monotonically in both directions away from the optimum. A given variation, of little consequence near the optimum, becomes more and more critical as the absolute value of the deviation increases.

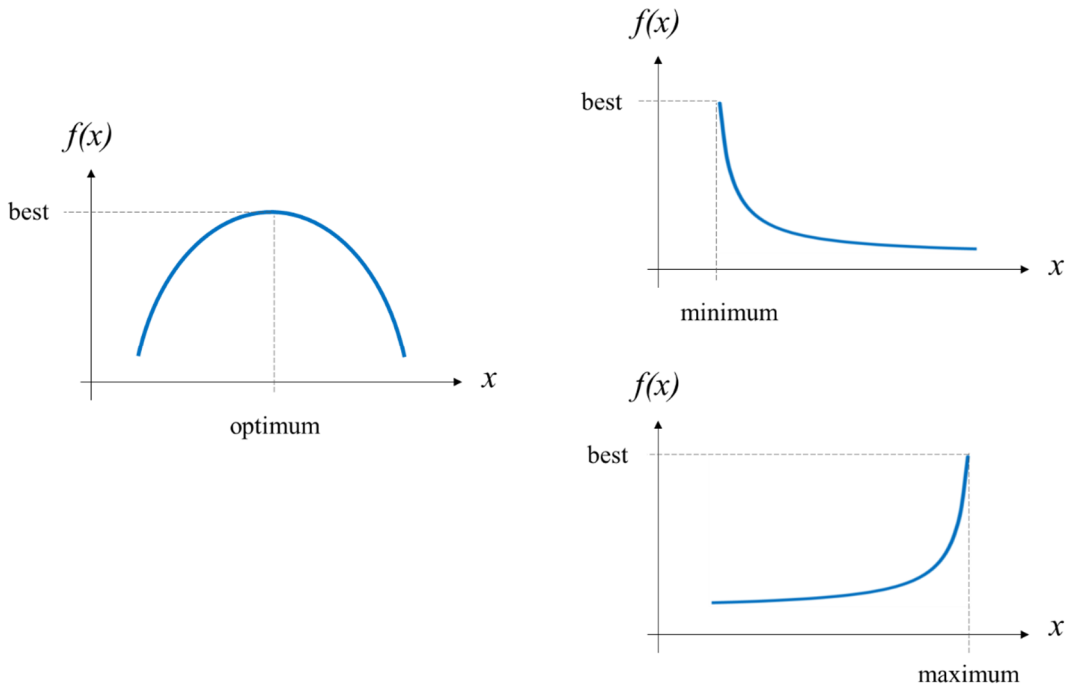

Figure 1. Schematic shape of the axiological/numerical dependency in feedback-controlled measures (left) vs. performance measures (right).

Formulas 1-5 are represented graphically in Figure 1.[7] The figure helps understand in what essential way the extremized measures of performance testing contrast with the optimized measures of feedback-controlled systems. In both cases, regulation vs. extremization of a quantity, some processes concur to maximize the dependent variable $y = f(x)$. There are two essential differences, however. In performance testing the axiological best is obtained at an extreme numerical value, either a minimum or a maximum, and the function is strictly convex, whereas in a feedback-controlled system the axiological best is reached at an optimal numerical value, and the function is strictly concave.

Formulas 1-5 are not equations, they just express *inequalities*. Only an order relation is assumed, once in Formulas 1-2 and twice in Formula 3-5. These propositions state that an order can be identified amongst differences expressed non-quantitatively, without taking it for granted that the y variable, the axiological value of quantity x, has the quantitative structure needed to make it measurable (Michel, 2014). But one can contemplate one promising alternative approach (to be explored below, see Section 9 and especially Figure 10). On can

[7] It is just for simplicity that the inverted U is drawn symmetrical, of course there is no reason to expect this relation to exhibit bilateral symmetry.

estimate the axiological value of a quantity by its impact on another quantity, provided that the latter can be plausibly considered an estimate of axiology.

The analysis presented in this section implies a three-class taxonomy of measures of possible interest to measurement theory. One class is that of *inert* measures, the main focus of classic measurement theorists, which offer no room for any objective axiological consideration—even though one is always free to express subjective preferences. Unlike Class 1, Classes 2 and 3 both have something to do with axiology. In Class 2, we find *regulated*, or feedback-controlled measures, characterized by the existence of a *concave* axiological/numerical relation. And in Class 3 we find *extremized* measures, or performances, characterized by a *convex* axiological/numerical relation.

To illustrate, we can see for example that the temperature of local atmospheric air falls in Class 1 (inert measures), whereas the internal temperature of the human body falls in Class 2 (regulated measures) and the temperature of a superconductive or superfluid piece of material in Class 3 (extremized measures).

6. The Interplay of Effort and Capacity: Lessons from Spirometry Testing

Although not traditionally viewed as an instance of performance testing, spirometry meets our definition of performance testing with no reservation: what is being measured by the spirometer is a basic physical quantity, a volume (or a flow) of expired air, and that quantity is subject to a deliberate extremization (here maximization) effort on a part of a human agent. The test has been used for more than a century by medical practitioners to estimate the respiratory capacity of their patients, under the guidance of old international standards (Miller et al., 2005).

There is no question that the spirometry test addresses a quantitative, measurable attribute. Having the dimension of length to the cube, the volume of a gas is an attribute that obviously satisfies Hölder's (1901) axioms. Importantly for our present purposes, the test is *one-dimensional*, just as the time-minimization or distance-maximization tests of athletics, meaning that the test output is a single number. It should be borne in mind that many performance tests consider two dimensions simultaneously, raising the problem of the sharing and dilution of the effort between two measures that often trade with each other. Psychology (e.g., Norman & Bobrow, 1975) as well as work management studies (e.g., Hollagel, 2009)

have long been aware of the trade-off that links speed and accuracy in all sorts of performance tasks. The spirometry test is much simpler, confronting the patients with a single task, just maximizing a volume of air.

Incidentally, an intriguing volumetric metaphor is detectable in the word 'capacity' often used in performance measurement contexts. The word conjures up the idea of a container of some typically fluid resource, the performer's capacity being conceptualize pretty much like the inner volume of a container. In spirometry testing it happens to be the case that the capacity of performance *literally* denotes a volume—specifically the volume of air that *could* be contained in the lungs. In the context of spirometry it is particularly easy to view the capacity as an upper limit located somewhere on the measurement continuum, and to see that that limit works as a global attractor towards which the performer's effort biases all measures and at the same time as a local repeller, since the closer to the capacity limit, the harder the resistance.

Standard spirometry considers a whole variety of volumes and flows.[8] For the present purposes, however, it will suffice to consider two volumes: *forced vital capacity* (*FVC*),[9] "the volume of air delivered during an expiration made as forcefully and completely as possible starting from full inspiration" (Miller et al., 2005, p. 320), and *total lungs capacity* (*TLC*), the total volume of air that the patient's lungs would physically contain if it were completely filled.

Forced vital capacity (*FVC*) is the measure of the patient's *respiratory performance*, a quantity that varies more or less erratically from maneuver to maneuver. *FVC* is what the spirometer measures. Total lungs capacity (*TLC*), in contrast, is an unknown. It is a patient-specific anthropometric parameter that we must suppose constant at the time scale of the spirometry session. Obviously *TLC* represents the upper bound of the patient's respiratory performance, characterizing his/her pulmonary capacity—put differently, his/her *capacity of performance*—in the test. To estimate that constant inductively from a (typically small) sample of *FVC* measures is precisely the goal of a spirometry test.

---

[8] Besides *FVC*, a volume of dimension $[L^3]$, spirometry considers various flow measures $[L^3T^{-1}]$ such as forced vital capacity in 1 second (FVC_1) or peak expiratory flow (*PEF*).

[9] The traditional terminology of spirometry is slightly misleading. The expression "total lungs capacity" is quite appropriate to designate what must be called a *capacity* in both the metaphorical sense of a capability and the literal sense of an inner volume susceptible to be filled with a liquid or a gas. In contrast, the term "capacity" is rather unfortunate in the expression "forced vital capacity" (or, synonymously, "forced capacity") because the latter quantity is a capacity in neither sense—actually it is a performance *measure* to be read on the spirometer and it varies haphazardly from maneuver to maneuver, depending on the magnitude of the testee's effort.

Let us start by noting that the measured performance *FVC* depends on the capacity of performance *TLC*:

$$FVC = f(TLC) \tag{1}$$

with *f* denoting some increasing function: the higher the patient's pulmonary capacity *TLC*, the higher the *FVC* performance to be expected from this patient in a maneuver.

Being constant during a testing session, *TLC* cannot be the source of the haphazard variability that affects *FVC* performance across maneuvers. The source is the variable extent to which the patients fill and empty their lungs. Let *E*, for *effort*, denote the depth of this more or less random filling/emptying process. We may write

$$FVC = E \times TLC \tag{2}$$

where *E* specifies the proportion of *TLC* that enters the device during the maneuver. That proportion varies from 0% to 100%:

$$0 \leqslant E \leqslant 1.$$

In the absence of any effort (*E* = 0) we have *FVC* = 0. With a total engagement of the patient (*E* = 1) we would obtain *FVC* = *TLC* and it would be correct to say that the spirometer measures *TLC*. Unfortunately, however, the patient's engagement is never total. An important relation linking *FVC* and *TLC* is the double inequality

$$0 \leqslant FVC \leqslant TLC, \tag{3}$$

meaning that the performance measure *FVC* can only underestimate the patient's capacity of performance *TLC*. Letting ε denote the error made by the practitioner who takes *FVC* as an estimate of *TLC*, we necessarily have

$$\varepsilon = FVC - TLC \leqslant 0. \tag{4}$$

The reason why the possibility of an overestimation must be excluded is simply because—cheating left apart[10]—the spirometer cannot receive a volume of air larger than that containable in the patient's lungs; one cannot give more than one has. But one can give much

[10] Some cases of extra breath have been reported (e.g., in the large scale spirometry survey of NHANES III, see https://wwwn.cdc.gov/nchs/nhanes/nhanes3/default.aspx) but their incidence is negligible. Moreover, an extra-breath incident is easy to detect for physicians as well as computer algorithms.

less than one has, meaning that occasionally the underestimation error can be very large indeed.

Upon each maneuver the patient must combine two consecutive efforts, an initial inspiration effort E_{in} immediately followed by an expiration effort E_{ex}

$$E = E_{in} \times E_{ex}, \tag{5}$$

where all three quantities are percentages. Simply because the two components of the effort combine multiplicatively, alarmingly low values of E can obtain, leading by Equation 2 to the possibility of very serious underestimations of *TLC*. For example, with E_{in} = 80% and E_{ex} = 70% we obtain a rather low E = 56%, as indeed 70% of 80% of something is 56% of it. No surprise then that practitioners, who understandably want to make as small underestimation errors as they can in the determination of their patients' *TLC*, so insistently urge them to try as hard as they can to maximize *both* E_{in} and E_{ex}.

Not only can we say that the underestimation error ε about the value of *TLC* in spirometry testing is entirely determined by the magnitude of the patient's effort *E*, we can specify that dependence quantitatively. Taken together, Equations 2 and 4 tell us that the error ε made in estimating *TLC* from *FVC* is an affine function of *E*

$$\varepsilon = (E \times TLC) - TLC, \tag{6}$$

whose slope and intercept are both given by the constant *TLC*. The error ranges from ε = –*TLC* for E = 0 to ε = 0 for E = 1 (Figure 2).

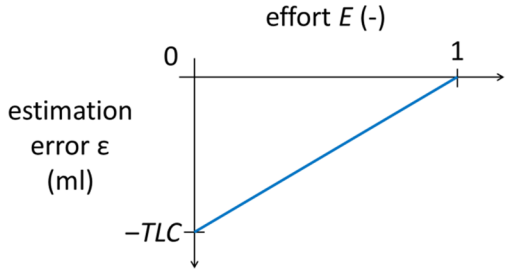

Figure 2. How the magnitude of the patient's effort *E* determines the error ε committed by the practitioner in estimating the patient's respiratory capacity *TLC* from the measured performance *FVC*.

We easily understand why spirometry patients should blow as hard as they possibly can into the spirometer on every single maneuver of the session: the deeper the respiratory

effort *E* in the maneuver, the smaller the underestimation error ε made in that maneuver in estimating *TLC*.

But the figure also provides a justification of the rather special way spirometry practitioners routinely handle their samples of *FVC* measures. International standards of spirometry (see Miller et al., 2005) have always asked practitioners to note down just the session's highest value of *FVC* and to flatly discard all others. This recommendation is easy to understand as the empirical maximum of a distribution of maximized performance measures must be expected to converge fairly quickly to its upper bound, and the convergence rate should increase with the strength of the biasing effort .

Notice, however, that such a recommendation is rather hard to reconcile with the standard principles of statistics. In the face of a measure that varies haphazardly from one observation to the next, standard statistics says that the measurer should take the whole sample into account and try, at the very least, to calculate a representative central-trend statistic (the mean, the median, or simply the mode of the sample). But notice that an extremum is, by definition, the *least representative* of all the measures of performance collected in a session. Thus the maximum value of *FVC* seems, at first sight, strikingly inadequate as a statistical location summary. The explanation is straightforward: the session's maximum of *FVC*—the *best* measure, to use an axiologically-colored term—is the *most valid* estimate of *TLC*, that which minimizes the practitioners' error in estimating the pulmonary capacity of their patients.

By the same token we also understand why it is standard practice in spirometry to dispense with any dispersion statistic.[11] Although a standard deviation (or an interquartile interval, or a min-max interval) computed on the collected sample of *FVC* measures might be useful to characterize the variability of the patient's effort during the session, for the practitioner such a characterization would pointless. Contrary to a widespread belief, the goal of the test is not to record such performance quantities such as *FVC*, but rather to estimate the *TLC* parameter, the upper bound of respiratory performance, inductively from a sample of respiratory performance measures.

[11] Some statisticians of spirometry recommend the computation of the difference between the best and the second-best measure of respiratory performance, and they interpret this statistic as an index of the 'repeatability' of the result (e.g., Enright, Beck, & Sherrill, 2004). This option will be critically discussed in Section 8 below.

If one overlooks that the gathered measures are, by construction, biased by a strong maximization effort, the goal of spirometry testing becomes unintelligible. In this regard a serious misunderstanding is manifest in this quote from recognized experts of medical statistics:

> "Let us suppose that the child has a "true" average value over all possible measurements, *which is what we really want to know when we make a measurement*. Repeated measurements on the same subject will vary around the true value because of measurement error. The standard deviation of repeated measurements on the same subject will enable us to measure the size of the measurement error" (Bland & Altman, 1996, p. 1654, emphasis added).

Apparently the authors have thoughtlessly tacked the standard statistical view onto the case of spirometry. Indeed there is a measurement-error problem in spirometry as everywhere else, but that sort of error can only account for a negligible proportion of the observed variability, in comparison with the proportion attributable to the variability of the testee's effort.

This being acknowledged, Equation 6 looks pretty much like a truism in the specific context of spirometry. It is more or less obvious that *FVC* measure is critically dependent on the patient's respiratory effort, that the respiratory effort can be defined as the ratio *FVC*/*TLC* (Equation 2), and that the goal of spirometry measurement is to estimate upper bounds, rather than averages. However, one justification for the above propositions is that they may possibly teach us a lesson of potential relevance to the study of human performance. Spirometry testing, in other words, may be considered an enlightening paradigm for the study of human performance in general.

The test constitutes a remarkably simple instance of a performance test. It is strictly one-dimensional, meaning that the performer's effort need not be shared and diluted. Second, the notion of capacity can be taken in the literal sense of an inner volume, making it particularly easy to see that the performance measure P is necessarily less than the capacity c, since the equality $P = c$ requires an impossible $E = 100\%$, and that the effort can be conceptualized simply as the proportion of the capacity that is actually converted into the performance.

The example of spirometry helps to see that performance testing always involves a crucial trio of variables, the *performance*, the *capacity* and the *effort*. Only the performance is measurable, the capacity and the effort being unknowns of the situation. This is not to say that no information is available to the practitioners concerning the lungs capacity of their patients,

because a bodily morphology conveys a great deal of information. For example the respirologist may sensibly expect a capacity on the order of 6 liters in this tall and healthy young adult, and much less, perhaps 3 liters, in this short and skinny person—but no direct measurement of the capacity of performance is possible. This personal parameter must be estimated from a sample of performance measures. Likewise, much information is available to the observer of a maneuver on the magnitude of the effort made in that maneuver—some maneuvers will be discarded by the practitioner based on the judgment that the effort was evidently submaximal—but the test does not measure the effort either.

7. Are Performance Measures Really Random Variables?

We have just seen that the unpredictability of performance in spirometry is primarily due to the haphazard variation, from trial to trial, of the strength of the testee's effort. The same process is presumably at work in all sorts of performance measurement situations. The question discussed in this section is whether or not a performance measure can be reasonably conceptualized as a random variable, as almost unanimously assumed.

In empirical applications of probability theory random vs. deterministic variables are distinguished in an all-or-none fashion. Thus many authors who take probability theory seriously use an uppercase vs. lowercase symbol notation to conspicuously distinguish random variables from fixed quantities (e.g., Rioul, 2008, to cite a mathematician). So used to do mathematical psychologist Luce, who dedicated a whole introductory chapter of his treatise on Response Times (Luce, 1987) to a detailed primer on what probability theorists call a random variable, response times obviously constituting in Luce's view a representative instance.

It does not seem too risky to say that experimental psychology, a field that has developed, since the second half of the nineteenth century, a high degree of expertise in the study of human performance (e.g., Fechner, 1860; Donders, 1868; Ebbinghaus, 1880; Woodworth, 1899), has constantly conceptualized performance measures as random variables. Nevertheless, taking into consideration the strong and systematic efforts that bias this kind of measures in the upward or downward direction, there is reason to feel uncomfortable with the idea that performances are random.

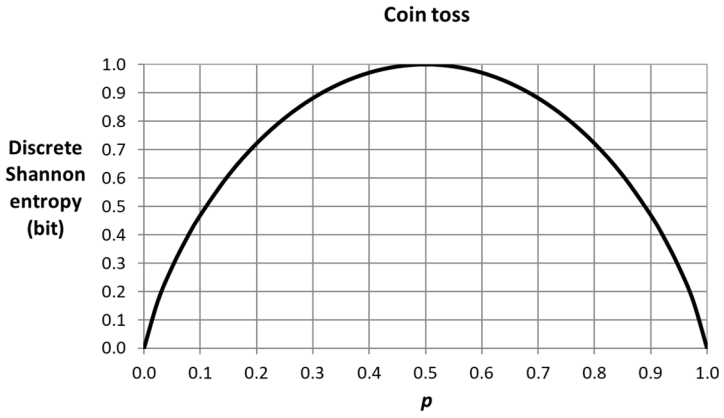

Figure 3. The entropy of the Bernoulli experiment vs. *p* (redrawn after Shannon, 1948, Figure 7, p. 11).

Figure 3 shows how in the coin toss, the classic illustration of the so-called Bernoulli experiment, a bias on the probabilities of the heads (p) and tails ($q = 1 - p$) affects the uncertainty of the experiment, which we measure here using Shannon's (1948) discrete entropy

$$H = -K \sum_{i=1}^{n} p_i \log \ p_i.$$

With a fair coin one faces total binary randomness with p = q = ½ and entropy at its maximum value $H = \log_2(2) = 1$ bit. If we eliminate randomness altogether by means of some contrivance (e.g., an appropriate layout of magnets) ensuring that $p = 1$ and $q = 0$ or vice versa, the coin toss turns into a purely deterministic, zero-entropy experiment. Between these two extreme cases we have the case where $p \neq q$, but with neither equal to zero. What if the coin is *very* strongly biased with, say, $p = .95$ and $q = .05$, meaning a very low level of uncertainty ($H = 0.25$ bit)? The system is still considered random. It is only at the limit where the entropy is *exactly* zero, with either $p = 0$ or $p = 1$, that it is considered deterministic.

Let us suppose we toss the coin n times and calculate the entropy of the sequence of heads and tails. Let $B = 2p - 1$ denote the bias: we have $p = 0$ for $B = -1$ (perfectly successful minimization effort), p = ½ for $B = 0$ (no effort whatsoever), and $p = 1$ for $B = +1$ (perfectly successful maximization effort). From the probability mass function of the binomial law we can compute the entropy of the Bernoulli process, a sequence of n independent Bernoulli experiments. Figure 4 shows the effect exerted on the entropy of the process by a gradual variation of the bias from $B = -1$ to $B = +1$, for $n = 1$, 10, and 100 tosses. Interestingly, the

shape of the function converges, as *n* is increased, on the shape of an inverted U. With *n* = 100, the function exhibits a large plateau surrounded by two abrupt cliffs.

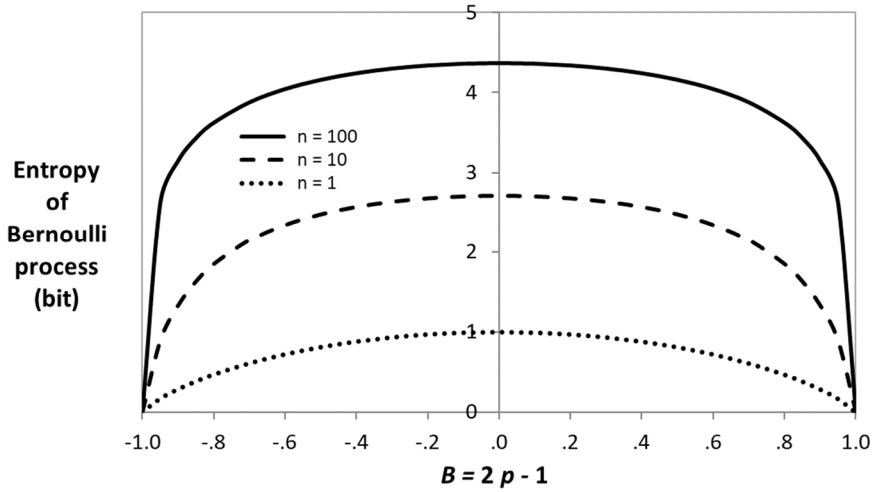

Figure 4. Effect of a probability bias on the entropy of a sequence of *n* coin tosses.

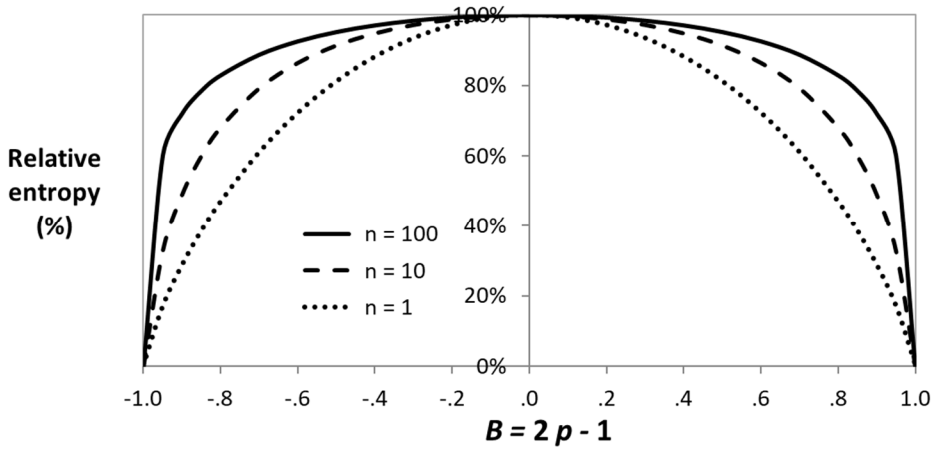

Figure 5. Effect of a probability bias on the degree of randomness, computed as relative entropy, of the Bernoulli process.

In Figure 5, the dependent variable is relative entropy, computed as Shannon's discrete entropy scaled to its maximum reached at $p = ½$ or $B = 0$, taken to measure the degree of randomness of the system. The curves show that with B close to −1 or +1, the system loses most of its randomness, and the greater n, the more true that observation. Common sense says that near the left and right edges of the plot the Bernoulli process looks much more like a

deterministic than a random process, but obviously some unequivocal criterion is needed. One such criterion is available in the shape of the probability mass function that can be calculated from the binomial law. Figure 6 show how a bias varied over its complete range from $B = -1$ up to $B = +1$ affects the first three central moments of the binomial distribution.

While the unbiased binomial distribution is a discrete approximation of the normal law, as often illustrated with the Galton's Quincunx, a moderate probability bias results in a shift of the distribution away from the central bin, with little change in the shape of the distribution. Even with a pretty pronounced bias like $B = -.9$ or $p = .05$ (Figure 7A), the distribution is still bell shaped, retaining its characteristic three parts separated by two inflection points, namely, a central concave body surrounded by two convex tails.

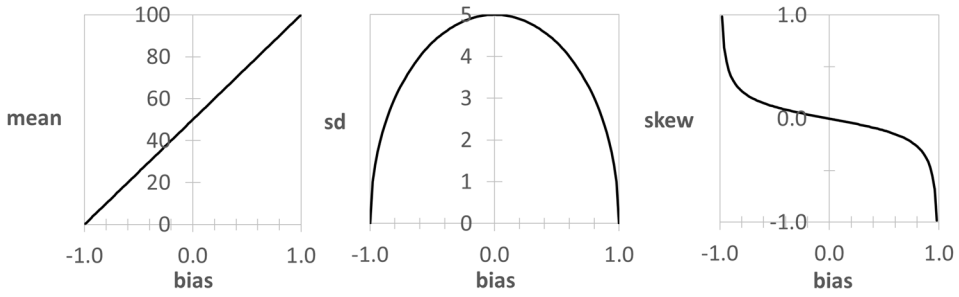

Figure 6. Effect of the probability bias on the arithmetic mean $\mu = np$, the standard deviation $\sigma = [np(1-p)]^{1/2}$ and the skew $\gamma = (1- 2p)/\sigma$ of the Bernoulli process with $n = 100$.

However, the bias also affects the spread and the skew of the distribution, and these effects are highly nonlinear (see Figure 6), so that some pretty abrupt changes take place in the shape of the distribution as B approaches −1 or +1. For example with $B = -0.95$ (Figure 7B), the lower tail has disappeared, with the histogram showing just one inflection point separating a concave body from an upper convex tail. And with $B = -0.99$ (Figure 7C), no more inflection point is observable, the distribution having turned entirely convex.

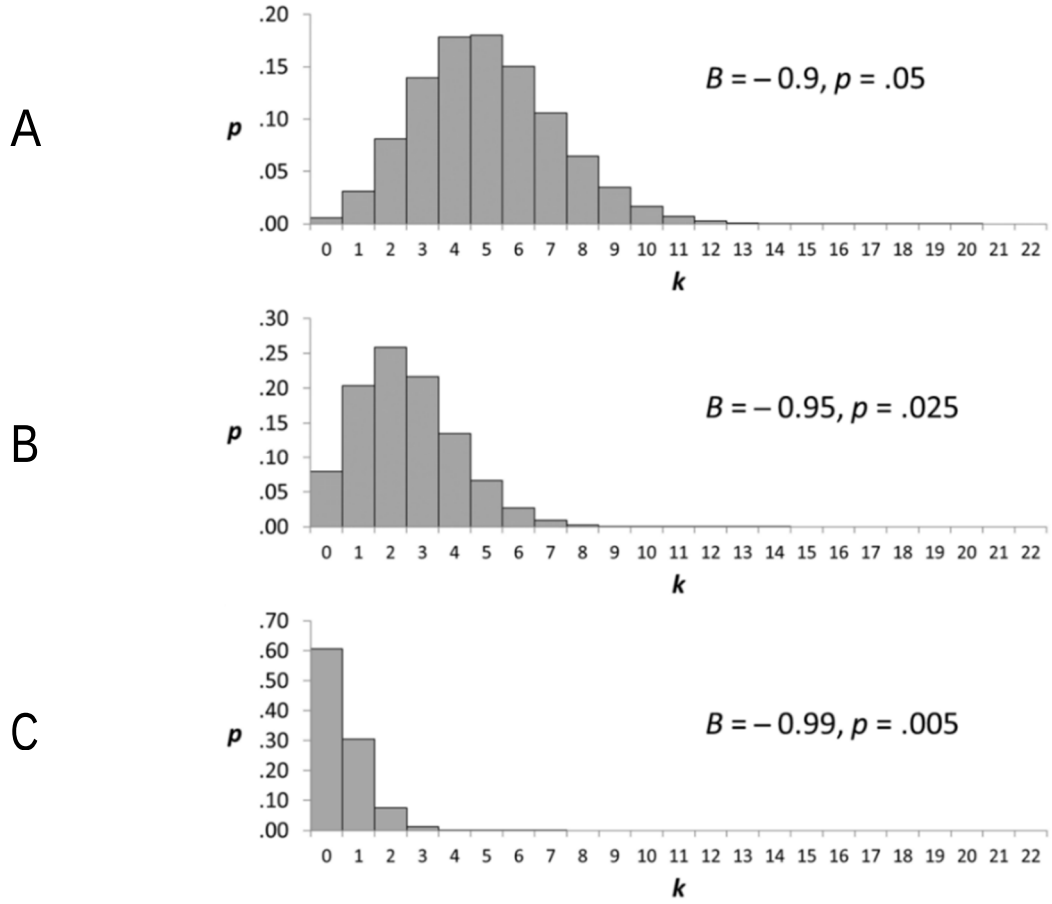

Figure 7. The crushing of probability against the lower edge of a binomial frequency mass function as a downward bias gradually approaches −1 in a Bernoulli process with *n* = 100.

The changes shown in Figure 7 provide us with an objective criterion to distinguish two non-overlapping intervals along the continuous dimension of randomness (Table 1). While the system is absolutely random at the limit where $B = 0$ and absolutely deterministic at the other limits where $B = \pm 1$, in between it makes sense to call it *quasi-random* if the distribution still exhibits its two original inflection points, and alternatively *quasi-deterministic* if it has lost at least one of its original inflection points.

Table 1. Degree of Randomness of the Bernoulli Process

	Bias	Probability	Number of inflection points
Random	$B = 0$	$p = q = 1/2$	2
Quasi-random	$0 < \|B\| < 1$	$0 < p < ½$ or $½ < p < 1$	2
Quasi-deterministic	$0 < \|B\| < 1$	$1 < p < ½$ or $½ < p < 1$	1 or 0
Deterministic	$B = \pm 1$	$p = 0$ or $p = 1$	0

This unconventional treatment of the binomial law, concerned with the range of variation of *B* from the limit of total randomness to the limit of total determinism, brings to the forefront what we may call distributional *edge effects*, illustrated in Figure 7. A mathematically-inclined mind might object that continuous distributions have no edges and that the gradual compression effect of Figure 7 are just artifacts of our discrete treatment of distributions. Indeed, there is some degree of arbitrariness in the choice of bin size in the making of a histogram. It is true that at the limit, with *n* tending to $+\infty$, meaning infinitely many and infinitely small bins, the distribution of an infinite population of measures will remain bell-shaped, no matter how strong the bias. Perhaps discrete modeling misses some important properties of the real world, but empirical distributions are always discrete, involving a finite number of bins. And discrete modeling has the merit of delivering an objective criterion allowing a four-level ordinal description of the degree of randomness that seems more realistic than the usual dichotomous description.

Below we will take the risk of treating the above-described theoretical edge effects seriously. We will see that they actually help understand the way in which variations in the intensity of the effort modulate the shape of discrete distributions of empirical performance measures (Section 7). To anticipate, the data will show that performance measures are more faithfully understood as quasi-deterministic in nature, than random. But before that, we must briefly consider the problem raised by two-dimensional performance, which complicates the matter.

8. Effort and Capacity in Two-Dimensional Speed/Accuracy Performance

In performance testing the performance is very commonly measured simultaneously in two dimensions between which the performers must share their effort—typically speed and

accuracy. For example many psychology experiments ask the participants to minimize their response time (*RT*), the latency of their response to some stimulus, while making as few errors as possible. Such experiments, like many situations of real life, confront the participants with the necessity of a speed/accuracy compromise. This section will put forward a tentative explanation of why a distribution of time performance is likely to look quasi-random, rather than quasi-deterministic. Two necessary conditions must be satisfied for the appearance of a quasi-deterministic distribution: (*i*) The performance measure must be lower-bounded, which happens to be the case with a movement time (*MT*) but not with an *RT*, and (*ii*) the time-minimization effort must be strong enough, and hence the concurrent error minimization effort weak enough.

Beside response time, by far the most common speed measure in psychology (see Luce, 1987; see also Annex 1), an alternative measure is task-completion time (*TCT*). The measurement of *TCT* appears in particular in studies of *speeded aimed movement* (Woodworth, 1899; Fitts, 1954), where the action required of the participants is typically an elementary movement of the hand. Provided with a pointing device, the participants are to repeatedly reach a target as fast and as accurately as they can. Here the two dependent variables are movement time (*MT*)—the time elapsed from the start to the cessation of motion—and some measure of endpoint error. Both constitute numerical performance scores that the participants are asked to try their best to minimize.

In the analysis to follow we will consider the time and the error measures in parallel. More often than not experimental participants are urged to perform in every single experimental condition with low enough error rates that they can be considered negligible. Such a tactic aims to simplify the analysis of the data by forcing inherently two-dimensional tasks into a one-dimensional mold. Prominent advocates of this classic approach are S. Sternberg (1966) in the *RT* field and Fitts (1954) in the aimed-movement field. Unfortunately, however, the tactic involves quite some wishful thinking. It is easy to check in virtually every data set (e.g., in Fitts's, 1954, own data) that in actual practice it is almost impossible to obtain a constantly negligible error measure: irrespective of instructions, the error measure is almost always found to correlate negatively with the time measure. Moreover, due to the highly nonlinear shape of the time vs. error trade-off, the closer to zero the average level of error, the larger the perturbation caused on the time measure, be it *RT* or *MT*, by an arbitrarily small variation of the error, as eloquently explained by Pachella (1973), Wickelgren (1977), and Luce (1986).

Rather than an undesirable technical complication, the speed/accuracy trade-off is a phenomenon we want to confront here. Specifically, we want to understand the interplay of the performance capacity and the extremization effort as determinants of the measured performance, and we will see that the two-dimensional context is in fact favorable to this inquiry.

The speeded aimed-movement paradigm introduced by Fitts (1954) is similar to the more widespread *RT* paradigm in the sense that it involves both a time measure and an accuracy or error measure, but in one regard it is much more suitable to our purpose. The precious advantage of the *MT* measure over the RT measure is that it is constrained by a strictly positive lower bound. Obviously, there is no such constraint in a latency measure, which may take arbitrarily low values and even occasionally turn negative ('absolute' anticipation) if the level of event or time uncertainty is low enough. *MT* measures the duration of an episode of overt motion. Unlike an *RT*, an *MT* cannot be indefinitely reduced, however strong the performer's minimization effort, simply because for a given value of prescribed movement amplitude the shorter the *MT*, the higher the demand in energy expenditure, and muscular power is limited.

A comparison with athletic runs may again help. Perhaps Usain Bolt, whose personal record is 9.58 s, could have covered the 100 m in 9.50 s, but almost surely not in 9 s. Bolt's performances have been constrained by a personal limit, a lower bound located in some region of the time-mark continuum. Let T_0 denote this theoretical lower bound, an unknown, and T_{min} Bolt's record, a measured value. We can safely say that $T_{min} \geq T_0 > 0$. Similarly we have seen that in spirometry the performance measure *FVC* is bounded on the upper side by the testee's upper bound *TLC,* so that $FVC_{max} \leq TLC$. The same is true in speeded aiming, where participants instructed to minimize *MT* as much as they can are constrained by their personal power limit, so that we necessarily have $MT_{min} \geq MT_0 > 0$.

Turning to the other dimension of aimed-movement performance, the basic measure of aiming accuracy is the distance from the endpoint of the movement to the target point—or to target center if the target is visualized as a tolerance interval. As visible in the example of Figure 8B, a typical distribution of endpoint error is bell shaped, with little or no skew, and

with its mean close to zero—reflecting a negligible amount of systematic error, with undershoots about equally frequent as overshoots.[12]

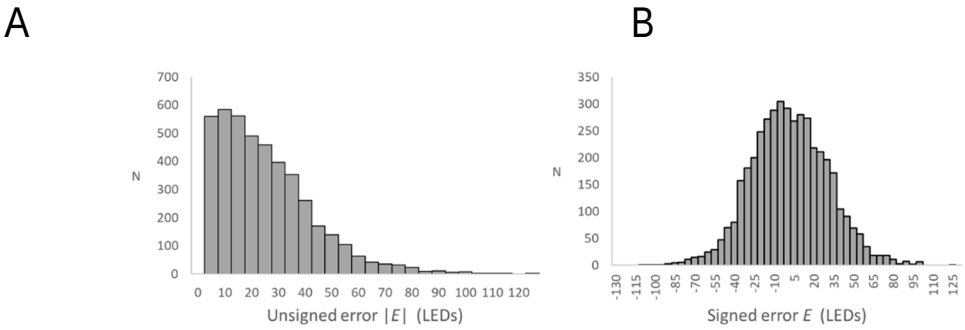

Figure 8. Distribution of endpoint error taken as an unsigned (A) vs. signed (B) length measure in a representative participant of Ferrand (1997) (see explanations in next section).

It is important, however, to realize that the undershoot vs. overshoot distinction is irrelevant to the evaluation of performance accuracy in speeded aiming: what the participants are supposed to minimize in this sort of task is always an *unsigned* length, the distance separating the movement endpoint from the target. Thus the relevant data is an inherently asymmetric distribution, as in the example shown in Figure 8A. One might be tempted to view that distribution as a half-Gaussian obtainable by folding the Gaussian of Figure 8B over its mean, but this view would be mistaken. In no way is the considerable degree of asymmetry visible in Figure 8A an artifact of a Gaussian folding. The accuracy measure under consideration is a performance, meaning by definition a measure that is being pressurized by an extremization (here minimization) effort, and so asymmetry is a property that we in principle *must* have in our distribution. It is by construction, and not as the result of a transformation of the performance-irrelevant distribution of Figure 8B, that the distribution of Figure 8A exhibits no bilateral symmetry.

[12] A systematic undershoot whose size increased monotonically with target width in a Fitts task was reported by Slifkin (2017), but this result is an artefact of atypical instructions. Slifkin asked his participants to hit the targets without asking them to aim at their centers—put differently, he allowed what we may call marginal aiming. Unsurprisingly then, his participants' movements tended to stop closer to the inner than outer edge of the target. For the Fitts paradigm to work properly—i.e., to allow experimenters to control, via the manipulation of *D* and *W*, mean movement amplitude and endpoint spread, respectively—the participants should be instructed to continue to aim at target centers even in the face of very large targets. Although neither Fitts (1954) nor subsequent users of his paradigm have apparently realized this, it is useful to note that with marginal aiming allowed much confusion is likely to arise in the effects of the two independent variables *D* and *W*.

In fact Figure 8B, which represents in the usual way a distribution of errors in aimed movement tasks (e.g., Wobbrock, Cutrell, Harada, & MacKenzie, 2008), provides just another eloquent verification of the classic law of errors of Quételet and Galton, which has open the way to the central limit theorem of de Moivre, Laplace, and Gauss. The law of errors says that the larger the deviation, whether downward or upward, from the 'true' expected value, the lower the frequency of occurrence. But here we are concerned with a performance measure and the sign of the error is irrelevant. The quantity subject to the minimization effort is the absolute magnitude of the error. Thus Figure 8B is 50% redundant and it is Figure 8A that captures the empirical information needed to characterize the accuracy component of the movement performance.

That there can be no left-right symmetry in the distribution of Figure 8A does not mean that that distribution must be 'skewed', skewness denoting a deviation from bilateral symmetry. Given that by definition a performance measure is biased by a deliberate effort, why should one take as one's implicit reference the symmetrical limit—i.e., the case of *total randomness* with no biasing effort whatsoever ($B = 0$, hence $p = q = ½$ in the binomial model)? It seems more reasonable to take as one's reference the opposite limit, that of *total determinism* ($B = \pm 1$ and $p = 0$ or 1). Taking that approach, we will not say that the distribution of Figure 8A is strongly skewed; rather we will say it is very different from the deterministic limit, exhibiting a relatively long tail on the right-hand side and a very conspicuous inflection point.

In the context of Figure 8A the idea of symmetry is just incongruous. We know for sure that a deliberate effort—opposed by the resistance of a lower limit located at zero, an impassable limit since a distance cannot be negative—has pushed the error data downward. Thus, if a convex *tail* is possible on the upper side of the distribution, which offers free space for a gradual decline of frequencies, on the lower side of the distribution one must expect a more or less abrupt *front*, with the data more or less clustered against the lower bound (Guiard, Olafsdottir, & Perrault, 2011; Guiard & Rioul, 2015). The point being made here is that when it comes to the measurement of a performance, by definition pressurized by a strong effort bias, distributions are better understood as organized along a *front-tail* rather than left-right axis. The merit of such an understanding is to dispense with any implicit reference to the hypothesis of symmetry, of no relevance in the present context.

9. *Distributional Analysis of Some Data from a Speeded Aimed Movement Experiment*

This section examines the shape of some empirical distributions of performance. Data from an old experiment on speeded aimed movement (Ferrand, 1997) is reanalyzed from the new perspective just outlined. Our main goal is to identify the necessary and sufficient conditions for a performance distribution to take the markedly asymmetrical shape shown in Figure 7C and by the same token to understand why performance distributions so often *look* random. Our strategy will be to visualize in parallel how the shapes of the distributions of *MT* and error are affected by the variation of the speed/accuracy balance.

9.1. Experimental Task and Paradigm

The data come from an experiment using the classic *time-minimization paradigm* of Fitts (1954; Fitts & Peterson, 1964). The target, which the participants are to reach as fast as they can using a screen cursor, is visualized as an interval of width *W* whose center is located at distance *D* from the start point of the movement. While the critical performance measure is movement time (*MT*), to be minimized, task instructions typically urge the participants to invariably make a very small (i.e., negligible) proportion of target misses, even with very small and very far targets.[13] The paradigm assumes that if that condition is satisfied, meaning that movement accuracy is under control via the manipulation of *D* and *W*, then one will be in a position, at the stage of data processing, to entirely focus on *MT* while flatly ignoring error data.[14]

The belief, widespread among users of the Fitts paradigm, that the difficulty of an aimed-movement task is captured by the so-called index of difficulty, which Fitts (1954) proposed to compute as $ID = \log_2 (2D/W)$, has the serious drawback of overlooking half of the story. On the accuracy front it is true that the higher the *ID*, the more 'difficult' the task, but notice that on the speed front the *opposite* is true: namely, the *lower* the *ID*, the more difficult the task. Experienced Fitts' law experimenters know that participants are often reluctant to

[13] Today it is widely admitted that 4% errors is an ideal error rate in the Fitts paradigm, in keeping with the explicit recommendations of an ISO standard (Soukoreff & MacKenzie, 2004). However, as recently emphasized by Gori, Rioul, and Guiard (2018), the rationale for this norm is lacking.

[14] The time-minimization paradigm of Fitts (1954) led him to the discovery of the empirical regularity known today as Fitts' law, which says that *MT* varies as a linear function of an index of difficulty (*ID*) that combines *D* and *W*, namely, $MT = k_1 + k_2 \log_2 (D/W + 1)$, where k_1 and k_2 are empirically adjustable coefficients. The paradigm assumes, in fact optimistically (Guiard & Olafsdottir, 2011), that the accuracy dimension of the movement is under full experimental control via the manipulation of the *ID*.

perform the extremely rapid movements required of them in the conditions with so high a tolerance that the risk of a target miss is virtually zero. Such conditions actually demand a high amount of physical effort on the part of the performer, thus shifting the effort demand from the accuracy front to the speed front (Guiard & Ferrand, 1998). Because whenever difficulty decreases on one front, it increases on the other there is no reason to depart from the view that overall task difficulty—i.e., the magnitude of the total effort to be invested in the task—is constant across the variations of the *ID*. To manipulate the *ID* in the Fitts paradigm is in fact to manipulate the *balance* between two mutually incompatible minimization efforts.

Ferrand (1997),[15] whose data we will revisit in this section, used a simplified variant of the Fitts paradigm, where the experimenter manipulates *W* but not *D*, kept constant. This variant was recommended by this writer to preclude the spurious correlation of the *ID* with movement scale that corrupts the usual variant of the paradigm (Guiard, 2009; see also Gori, Rioul, Guiard, & Beaudouin-Lafon, 2018), and thus to avoid a factor confound that has introduced noise in many Fitts' law experiments, including Fitts's own (Guiard, 2019). The manipulation of the tolerance of a target whose center remains at the same location is essentially equivalent to instructions to adopt different speed/accuracy compromises: the difference is simply that task instructions are given visually in the former case and verbally in the latter.

Ferrand (1997) used the *reciprocal* protocol (Fitts, 1954), in which the participants are to alternatively reach two targets of width *W* separated by a distance *D*. Beside some minor drawbacks (see Guiard, 1997), the reciprocal protocol has the advantage of delivering many measures per experiment. Unlike respirologists, who can record only three to five measures of respiratory performance from their patient (Miller et al., 2005), students of aimed movement who use the reciprocal protocol can gather in a single session hundreds of measures from each participant. Collected from six participants who completed each about 10,000 movements in four sessions, the Ferrand data seems quite suitable to a study of distribution shape.

Ferrand used the traditional set of task instructions, asking his participants to perform the reciprocal movement as fast as possible under the constraint of a constantly small

[15] Warm thanks are due to Thierry Ferrand for a considerable amount of help to unearth and reprocess anew the data he collected himself a quarter of century ago, during his PhD completed under the supervision of this author (Ferrand, 1997. *Coopération Bimanuelle Intra- et Inter-Individuelle dans une Tâche de Pointage*, unpublished doctoral dissertation, Université de la Méditerranée). Procedural details about this experiment can be found in a subsequent treatment of the same data published by Mottet, Guiard, Ferrand, and Bootsma (2001).

proportion of target misses. He manipulated *W* (4, 8, 16, 32, 64, and 128 LEDs)[17] at a constant level of *D* (256 LEDs).

9.2. Apparatus, Procedure, and Data Processing

The experimental apparatus consisted of a visual display and a pair of manipulandums.[18] The display showed two contiguous vertical columns of 512 light-emitting diodes (LEDs). The left column of displayed four luminous dots forming the two targets of the reciprocal protocol, which could be moved up and down as a whole using the left manipulandum. The right column displayed a single luminous dot representing the pointer, whose position was controlled by the right manipulandum. The manipulandums were two identical carriages that could be moved back and forth along a linear course of 308 mm. For the pointer to cover the constant distance between the two target centers the movement had to cover an amplitude of 154 mm.

This apparatus allowed Ferrand (1997) to investigate a number of task variants with the aimed movement performed either with one hand or with two hands.[19] Here we will indiscriminately pool together the various conditions of hand assignment, whose effects were very small and statistically independent of the effect of *W*. We will consider just one independent variable, *W*, interpreted as a manipulation of the balance between the speed and the accuracy efforts, and two dependent measures, *MT* and $|E|$.

Reciprocal pointing in this setting taking the form of a one-dimensional oscillation, *MT* was measured as the horizontal peak-to-peak distance in the time profile of pointer displacement, using a sampling frequency of 50 Hz (20 ms period) and $|E|$ was measured as the distance separating the pointer from target center at the reversal point of the oscillation.[20]

The oscillatory form taken by the target-acquisition movement in the reciprocal protocol precludes the possibility of independence between successive measures, thus

[17] The light-emitting diode (LED) serves here as the unit of measurement for pointing error. The LED measured 2.34 mm and was seen from a distance of 2 m, thus corresponding to a visual angle of 0.07° and a displacement of 0.6 mm.
[18] For a detailed description, see Guiard (1993).
[19] The bimanual condition involved either the left and right hands of an individual participant or the two right hands of two different participants. For simplicity the present reanalysis ignores the performance of dyads, focusing on the data of individual performers, but it is interesting to note that essentially the same patterns of results obtained with dyads as with individuals (see Mottet, Guiard, Ferrand, & Bootsma, 2001).
[20] Instructions made it clear to participants that the extrema of pointer excursion in either direction would be treated as the endpoints of their movements.

violating one important criterion of the random variable according to probability theory[21] (e.g., Rioul, 2008). Luce (1986) has remarked that in response-time studies the existence of sequential effects, found to be strong and systematic in every single study that has investigated them (see, e.g., Kornblum, 1969), is very problematic for the applicability of the random model to such measures. In fact the case is still worse for time and error measures in the reciprocal Fitts protocol, which elicits a rhythmic sort of behavior. Neither the period nor the amplitude error of a sustained oscillation, best tackled with the concepts of nonlinear dynamic systems theory (Kelso, 1995; Kugler & Turvey, 2015), is likely to behave like a random variable. However, the assertion that each of these two quantities is subject to a deliberate minimization effort is immune to the objection.

9.3. Results

This section presents an attempt to jointly understand the changes that take place, in parallel, in the shapes of empirical distribution of *MT* and of |*E*| in response to the manipulation of the balance of the performers' effort in the face of two incompatible minimization demands. Bearing in mind that the lower bound of *MT* is a personal constant whose value differs from participant to participant, we will refrain from pooling the data from different participants and will exclusively consider within-individual distributions. We will resort sparingly to any statistical compression of empirical information, deliberately dispensing with the summaries of classic parametric statistic, if only because most of the distributions exhibit no symmetry.

Below the data is visualized in three steps. We start with a raw representation of two-dimensional performance, with all levels of tolerance pooled together (Figure 9). Then we visualize the two one-dimensional distributions side by side, with target tolerance shown as an independent variable (Figure 11). And finally we use box plots to obtain prudent non-parametric descriptive summaries of the results (Figures 12-13).

Figure 9 shows for each of the six participants of Ferrand's experiment a raw scatter plot of |*E*| vs. *MT*. Note that this plot involves no statistical compression of empirical information whatsoever, each data point representing the duration and the error of one individual movement (i.e., one half-cycle of pointer oscillation). All the data is represented, with all six levels of *W* pooled together.

[21] The so-called iid condition, meaning independent and identically distributed.

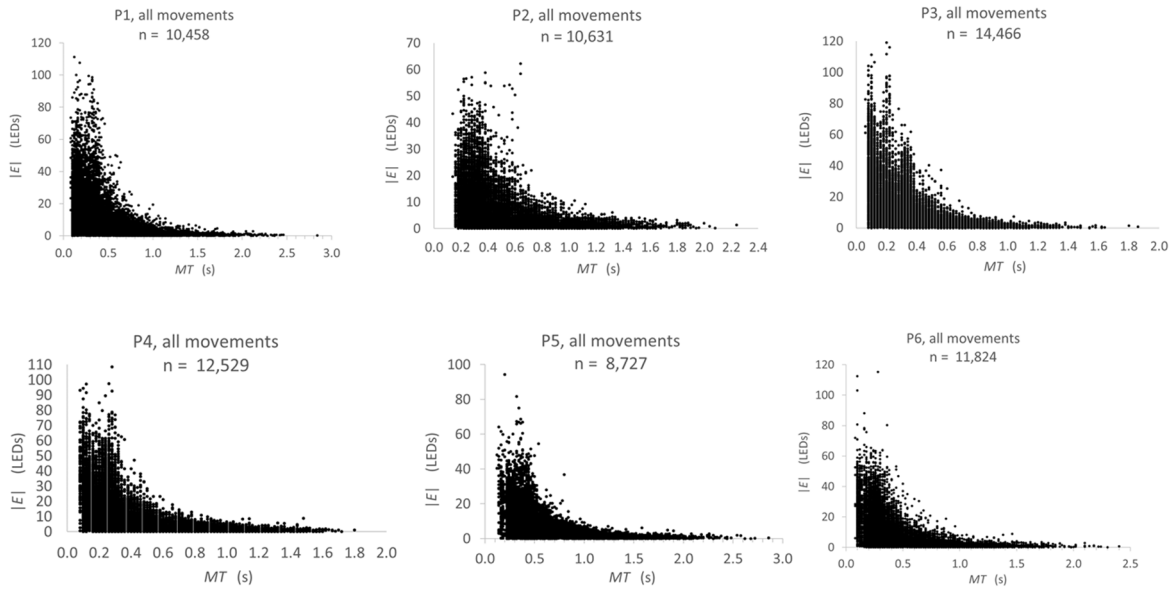

Figure 9. Raw two-dimensional distribution of performance. All conditions of target tolerance are pooled. Each data point corresponds to one individual movement, *n* denoting the total number of movements performed by the participant in the experiments, all visualized here.

This elementary plot visualizes quite clearly the way in which the encounters of capacity and effort taking place simultaneously on the two fronts influenced each other. The global shape of the two-dimensional distribution is remarkably invariant across individuals. Consider first how the vertical distribution of data points changes from left to right: the longer the measure of *MT*, the more marked the clustering of the measures of |*E*| against their lower bound—zero by construction. Then consider how the horizontal distribution changes from bottom to top: the larger the |*E*|, the more marked the clustering of the *MT*s against their lower bound—whose values are unknown personal parameters but which we must suppose to rest either at the observed minima (which range from 60 ms[24] to 140 ms in the sample of participants) or just below, as we learned for spirometry analysis (Section 4).

Thus, the more pronounced the downward clustering of the performance measures on one axis, the less pronounced on the other axis. In other words, the stronger the minimization effort in one dimension, the weaker the minimization effort in the other dimension.

[24] A 60-ms half-period of pointer oscillation corresponds to an instant frequency of over 8 Hz. That minimum—the experiment's best or 'record' performance score—was reached only three times by P3 in the condition of highest tolerance, but the data show that in that condition all participants except P2 and P5 were able to sustain a frequency of pointer oscillation of about 6 HZ (80 ms half-period) over episodes of a few seconds.

An interesting minor finding, with again high consistency across participants, is that the strength with which the data points gradually cluster against their lower bound is much less pronounced in the leftward than downward direction. In a speeded aimed-movement task it seems more costly for performers to reduce *MT* than |*E*| down to their respective lower bounds. Task conditions with lower indices of difficulty are decidedly not 'easier'.

A much more important point is made in Figure 10, which shows, in one arbitrarily chosen participant, how *MT* is affected by the variation of endpoint error expressed, for the sake of the argument, as a signed measure (i.e., *not* a performance measure, as emphatically explained in Section 8). Since *MT* can serve as a quantitative estimate of the time cost of error minimization, let us use it as an objective criterion for estimating quantitatively the axiological value of the error measure (see Section 5).

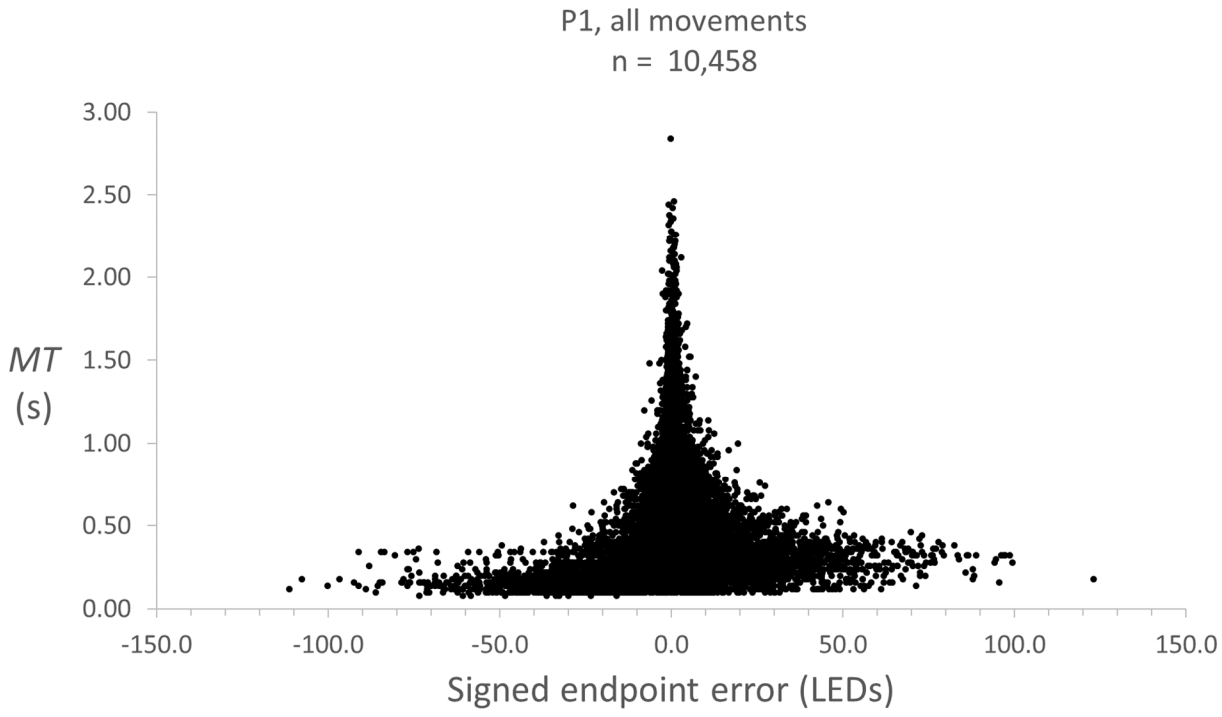

Figure 10. Raw two-dimensional distribution of performance in participant P1, with now MT plotted on the vertical axis and the error expressed as a signed measure.

The figure shows rather eloquently that the closer the measure to zero the higher the time cost. This empirical fact allows us to say that the closer the measure to zero, the higher

its axiological value—after all, were it not the case that the numerical value of zero is the best axiologically, we would wonder why this abscissa, where the temporal cost is so high, enjoys such a frequency of occurrence in the distribution.

Notice that the time cost increases at an accelerated rate as the signed error approaches zero, whether from above or from below. The axiological/numerical relation for the signed error of Figure 10 is indeed *double convex*, combining the two convex relations of Figure 1. One could possibly say that an aimed-movement task requires a double error extremization, in the sense that performers have to *maximize* the abscissas of their undershoots (up to zero) and at the same time to *minimize* the abscissas of their overshoots (down to zero). Obviously, it is much simpler to say that they have to zero out the absolute value of their terminal errors, and this is a potent argument for switching once and for all to an absolute-value representation of performance accuracy.

Figure 11 offers another visualization of the same data, showing distributions of *MT* and |*E*| side to side for one arbitrarily chosen participant. In this figure some compression of the empirical information has taken place, that required to make histograms, where slightly different measures are pooled in the same bins.

Now the six conditions of tolerance are separated, with *W* increasing geometrically from top (4 LEDs) to bottom (128 LEDs). What changes from top to bottom in the figure is not task difficulty—to reiterate, task difficulty must be considered invariant in the Fitts paradigm—but the balance of the performer's effort, forced to shift gradually from one extreme, pure accuracy effort, to the other extreme, pure speed effort. Having discarded the traditional misunderstanding about task difficulty, we are in a better position to see that the performance is two-dimensional in intermediate conditions of tolerance and essentially one-dimensional in the two extreme conditions.

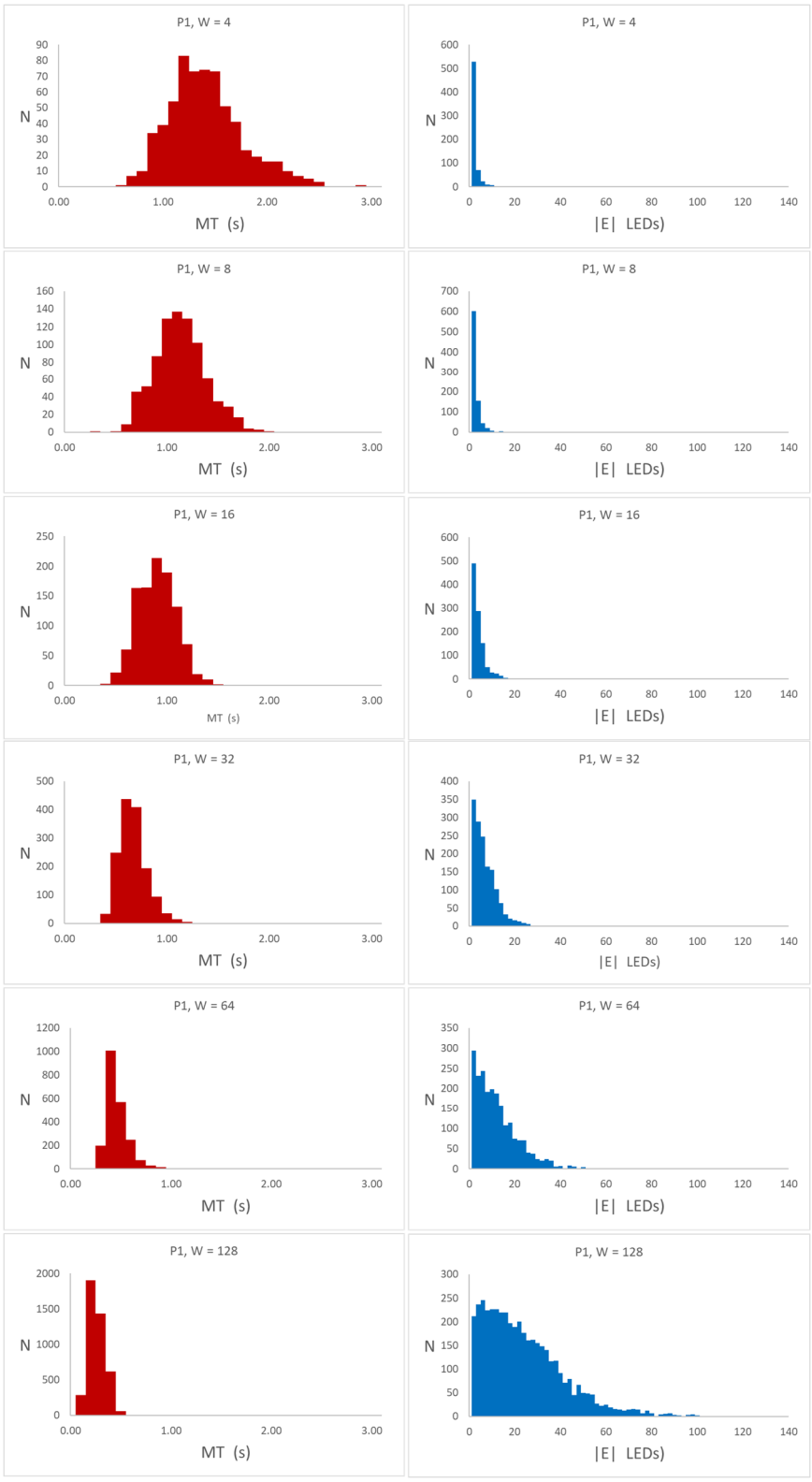

Figure 11. Parallel effects on the shapes of distributions of *MT* and |*E*| of the variation of the speed/accuracy balance entailed by the manipulation of target tolerance at a constant level of movement amplitude. The result, shown in one arbitrarily chosen participant, is essentially the same in others.

In both columns of the figure the shape of the distributions change so dramatically across the conditions that it would make no sense to ask about the 'true' shapes of these distributions. What we need to understand is how the two distributions shapes *change* dynamically under the systematic variation of the speed/accuracy balance, recalling that each of these distributions is pressurized by a minimization effort directed against a lower bound and that the strength of the effort can increase on one front only at the expense of a decline of the effort on the other front.

Let us start with the extreme condition with $W = 4$ LEDs (top), where the performer's effort was almost totally dedicated to the quest for accuracy. The shape of the histogram of endpoint error $|E|$ (top right panel) is reminiscent of the strongly biased binomial distribution of Figure 7C, with an abrupt, non-convex front on the lower-values side and a dramatically compressed convex tail on the higher-values side. The distribution exhibits no inflection point. Using the criterion of Table 1, the observed shape is characteristic of a quasi-deterministic variable. How about *MT* (top left panel) in this maximum-accuracy condition? Not only it is, unsurprisingly, rather long (over 1 s) on average, it is distributed as a quasi-random variable, exhibiting the shape of a nearly symmetrical bell with its three part separated by two inflection points. Thus we can see that when the task demands that the performer's effort be entirely focused on accuracy, the error is very small and distributed as a quasi-deterministic variable, while the time measure is very long and, for lack of a substantial minimization effort, distributed pretty much like a random variable.

In the other extreme condition with $W = 128$ LEDs (bottom), the picture is essentially the same, mutatis mutandis. Now it is the distribution of *MT* (bottom left panel) that is reminiscent of Figure 7C, showing a considerable amount of downward compression. In contrast the error distribution (bottom right panel) shows little compression with a distinct inflection point between a convex tail, located on the higher-values side, and a concave body. Now it is *MT* that has turned quasi-deterministic under the pressure of the speed effort while the error, neglected by the participant, has turned random.

To see that the bottom condition required an essentially one-dimensional, pure speed effort one must realize how high the tolerance for movement error was with $W = 128$ LEDs: for a target miss to be recorded in that condition the endpoint error had to exceed ±64 LEDs—i.e., one fourth of the required movement amplitude, 512 LEDs. In the particular participant of Figure 11 the spread of endpoint error resulted in 3% target misses, but that error rate was 1% or 2% in four of the other five participants—definitely less than the level of error explicitly

recommended by instructions (5%). Thus in the bottom condition the risk of a target miss was no longer a concern for the participants. Their main concern in that condition was to move as fast as possible, and accordingly to make as *many* target misses as possible, and that meant struggling with their personal speed limits.

We now turn to intermediate levels of tolerance with $W = 8$, 16, 32, and 64 LEDs. Here the task was genuinely two-dimensional, requiring of the participants various mixtures of speed and accuracy efforts. On the speed front, as target tolerance is raised from top to bottom in the figure one can see a gradual downward shift of the bulk of the distribution of *MT*, along with a gradual reduction of its spread, down to a final clustering of the data points against the lower bound. The picture is quite similar on the accuracy front. As the tolerance is reduced, the error data cluster more and more against the zero-error limit. One difference is that the measures of $|E|$ tend to accumulate against the physical limit of zero LED, a lower bound obviously common to all participants, whereas the measures of *MT* tend to accumulate near their personal lower bounds (in the vicinity of 80 ms in this particular participant).

In Figure 12 the layout is the same but the distributional information has been massively compressed into the five statistics of the box plot, an ordinal technique that delivers a faithful description of deformable distributions under the minimalist assumption that they remain unimodal—which never ceases to be the case in this data set. The translation and compression/expansion effects of the previous figure are still quite apparent in this highly compressed statistically representation. Needless to say, the usual parametric approach, which presupposes symmetrical distributions, would have delivered a severely distorted view of the data.

Consider errors on the left-hand side. While the sample minimum Q_0, whose values never depart from the horizontal axis, is unaffected by *W*, the three central quartiles of the error $|E|$, Q_1, Q_2, and Q_3, scale proportionally with target tolerance, with increasing slopes. However, for the last quartile (i.e., sample maximum), the function turns affine, the non-zero intercept reflecting the fact that even with all their effort dedicated to the quest for accuracy the participants were unable to totally annihilate the trailing edge of their distribution of errors.

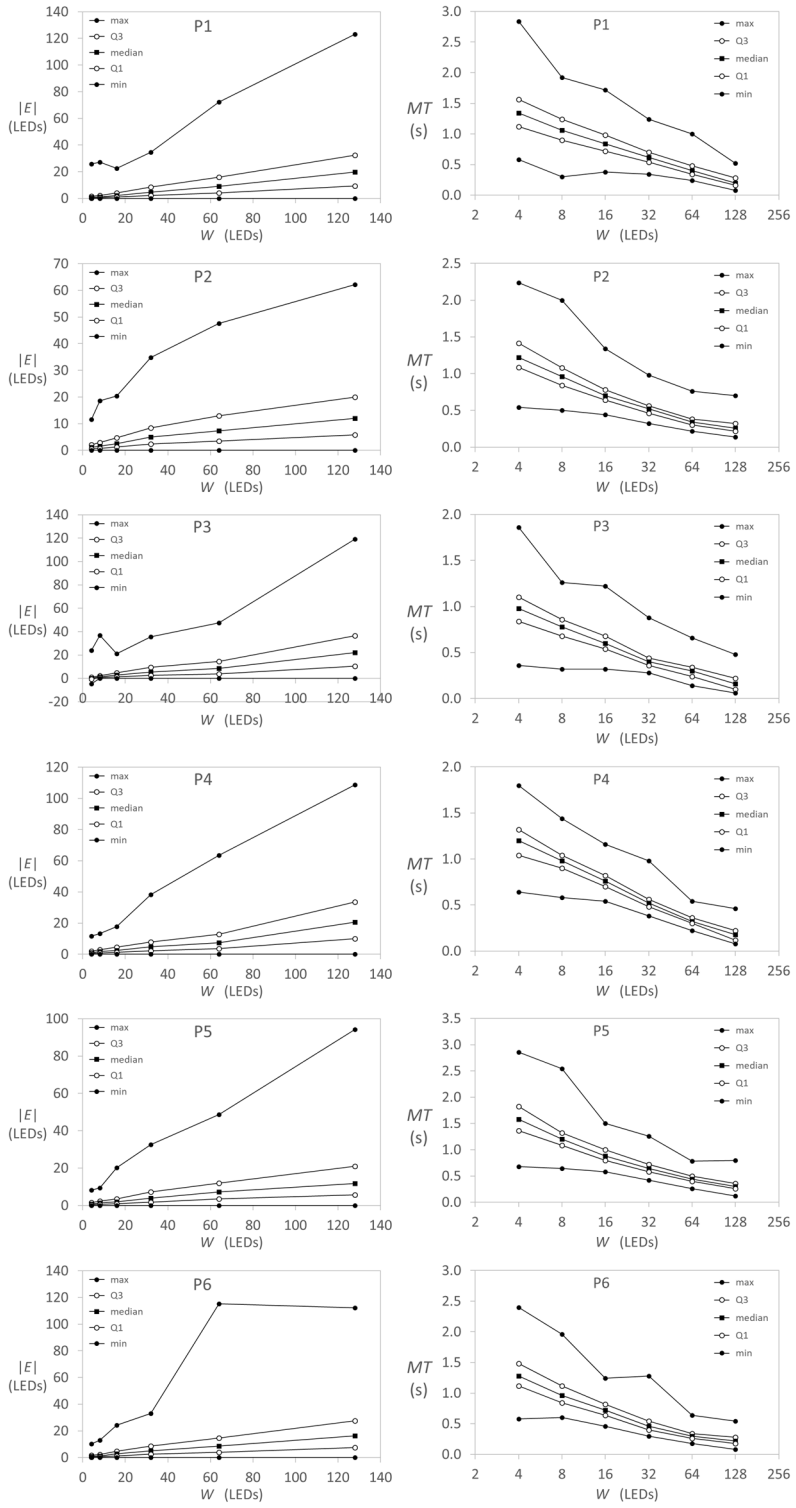

Figure 12. A boxplot description, for each participant, of the effects of the variation of the speed/accuracy balance on the shapes of the distributions of $|E|$ (left) and MT (right). Notice that in the plots of MT the horizontal scale is logarithmic. To avoid visual clutter the box and whiskers proper are not drawn, but the five quartile statistics (minimum or Q_0, first quartile Q_1, second quartile Q_2 or median, third quartile Q_3, maximum or Q_4) are shown for each level of W.

The right-hand side panels show that for all participants all the quartiles of *MT* vary as decreasing affine functions of the logarithm of *W*, in keeping with Fitts' law (Fitts & Peterson, 1964; Gori et al, 2019; MacKenzie, 1991).[26] Just as with errors, the slope increases with quartile rank, but the slopes of the minimum and maximum diverge from those of the central quartiles Q_1, Q_2, and Q_3. The slope is minimal for the distribution minima Q_0 and maximal for the distribution maxima Q_4 the slope is especially steep, reflecting the gradual compression of the trailing upper tail.

Recourse to these statistical summaries opens the way to a quantitative description of the morphological changes. Figure 13 quantifies one remarkably consistent aspect of the pattern of Figure 12—the especially steep slope of the linear function linking the maxima of $|E|$ to *W*. With the shift of the speed/accuracy balance so as to produce shorter *MT*s, the distribution of $|E|$ from Q_0 up to Q_3 expands upwards in an uniform way, an effect reminiscent of that observable on a linear spring subject to a gradually relaxed compression, where the measured displacement is proportional to the distance from the edge of the spring. The non-trivial finding is that sample maxima obey a different law, the distributions having their tails more affected by the manipulation of *W* than their bulks.

Incidentally, one may guess from the patterns of Figure 12 that they all verify Weber's law, an old rule of thumb of psychology which says that in general, particularly in time measurement contexts, the standard deviation of a distribution tends to vary proportionally with its means (Meyer, Abrams, Kornblum, Wright, & Smith, 1988; Schmidt, Zelaznik, Hawkins, Frank, & Quinn, 1979). This empirical regularity in fact holds in each panel of the figure, with particularly high qualities of linear fit in the distributions of $|E|$. It is worth emphasizing here that while the explanation of Weber's law has remained elusive (e.g., Dehaene, 2003), that regularity is very easy to interpret in light of a distributional compression phenomenon. If a distribution of minimized performance measures crushes against its lower bound, it comes as no surprise that the first and second moments of the distribution will tend to correlate positively. The immediate prediction follows that in maximization contexts Weber's law should show up in reversed form, with the correlation between the standard deviation and the mean turning negative. Preliminary data collected by this author appear to support this prediction.[27]

[26] Plotting on the horizontal axis the $ID = \log_2 (D/W + 1)$ in the place of $\log_2 W$ would essentially result in a horizontal translation of all data points.
[27] Guiard, Y. (in preparation). Inverted Weber law in respiratory performance distributions.

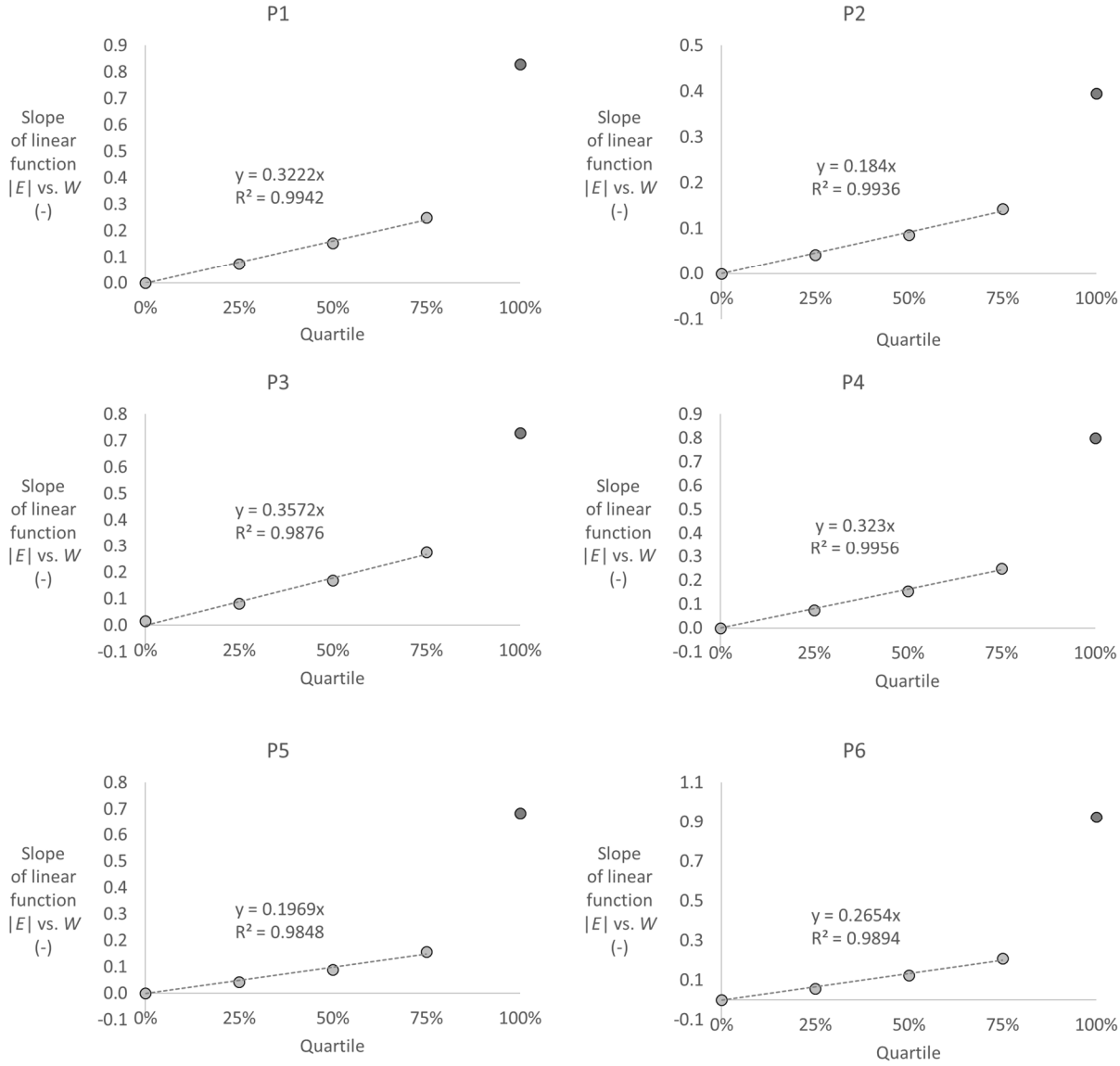

Figure 13. A quantitative description, in all six participants, of the way the distribution of error contracts and expands in response to the manipulation of target tolerance, with divergent distribution tails.

9.4. Discussion

Is there something of general interest for the understanding of performances to be learned from these results, obtained by reanalyzing the data of an old experiment on aimed-movement? It is noteworthy that several other data sets, including that collected in a markedly different paradigm by Guiard et al. (2011), have been submitted by this author to the same

descriptive analysis, producing essentially the same results. Apparently, the regularities visualized in Figures 9-12 hold for speeded aimed movement in general.

The results, obtained in the difficult and important case of two-dimensional performance testing, help understand how a minimization effort opposed by the resistance of a lower bound strains the movement-time and error measures. The experiment offered no direct measure of the magnitude of the effort, but the direction in which that magnitude varied was experimentally controlled on each dimension of the task, making it possible to confirm the close mutual dependency of the two parallel compression phenomena. The results confirm that the balance of the minimization effort, which is manipulated systematically in the Fitts paradigm, determines the balance of downward distributional compression. As the effort is made to increase on one front, more data compression is observable on that front, and less on the other.

10. General Discussion

This study aims to clarify an old and obviously important concept of psychology, the quantitative concept of performance, to which surprisingly little attention has been paid. In the context of laboratory research the meaning of the highly frequent word “performance” is in general taken for granted. If there indeed exists a wide and stable consensus on the meaning of that word, we have realized that this consensus suffers a blind spot. Even though psychologists never omit to ask their participants to do their best to extremize their performance measures, at the stage of data processing it looks like they systematically forget that they recorded these measures under highly special, heavily pressurized conditions.

The simple definition formulated at the start of this paper brings to light what might be considered on second thoughts a truism, namely that performances measures are deliberately and strongly strained by an extremization effort. Much of the paper was aimed at understanding in what precise sense the measures of human performance are made special by this characteristic. We have seen that the effort-strained, extremized measures we obtain in performance testing situations confront us with an axiological issue. We found that measures can be classified in terms of the curvature of the function linking the axiological to the numerical. While many measures are axiologically undefined, the function is strictly concave for regulated measures and strictly convex for performance measures.

An attempt was then presented to elaborate a provisional understanding of the mechanism by which the effort strains the performance. The simple one-dimensional instance of spirometry helped us to see that the performer's effort is best conceptualized as a force exerted against a resistance, and that the resistance results from the existence of a capacity limit—in the case of spirometry an obvious mechanical upper bound. It was proposed that any measure of performance is determined by the interplay of two variables neither of which can be directly measured: the effort, which we must suppose to vary haphazardly from trial to trial, and the performer's capacity, a personal parameter best thought of as a constant, at least at the time scale of a performance-testing session.

On this specific issue the paper has just outlined a gross approach. The capacity limit was conceptualized tentatively as an impassable bound on the measurement continuum, relying on the metaphor of a hard wall, but more realistic models seem desirable. One promising perspective seems to be offered by the stress-strain relationship of mechanical engineering.

We then turned to the case of two-dimensional performance, frequently encountered in the psychology laboratory. We focused on data from an experiment on speeded aimed movement where the participants had to minimize two measures known to trade with each other, *MT* and |*E*|. A distributional analysis revealed that the sharing of effort between the two incompatible demands of speed and accuracy was faithfully reflected in the shapes of the distributions. The more effort dedicated to one dimension of the task, the more downward compression observed in the corresponding measure, and the less in the other. Only in conditions where the minimization effort was rather weak, because it had to be pretty strong on the concurrent task dimension, did the distributions of performance measures resemble Gaussians.

Apparently four necessary conditions must be satisfied in practice for the distributional compression effect to show up in a performance measurement situation.

1. *The measure considered must be subject to an extremization pressure*. In studies of aimed movement, for example, no compression effect will be observed if the accuracy of the movements is estimated by the signed value of the endpoint error *E*. The quantity that is actually minimized in aimed-movement experiments is |*E*|, not *E*. As we have seen, |*E*| qualifies and behaves as a performance measure whereas *E* is a classic random variable.

2. *The extremization pressure must be sufficient.* This condition is typically not met in psychological studies of response time, most of which which impose on participants an unconditionally high level of accuracy, hence precluding from the outset the possibility of a strong time-minimization effort. As a matter of fact, most studies of *RT* record their time measures in conditions that resemble the minimum-tolerance condition of the Ferrand (1997) experiment shown in the upper panel of Figure 11. In that condition the distributions of *MT* indeed look quasi-random for lack of a sufficiently strong minimization effort.
3. *The measurement continuum considered must have a lower or upper bound to resist the extremization effort.* As noted, latency measures provide an obvious counter-example. An *RT* measure is just not bounded. The only constraint preventing an *RT* from approaching or even crossing the 0-ms value is the concurrent accuracy requirement: the faster the guess, the higher the probability of an erroneous response. The quantity that is bounded in the response time paradigm, because of the shortage of the effort resource (Norman & Bobrow, 1975; Guiard & Rioul, 2015), is the combined speed/accuracy performance. To obtain a directional compression effect in this paradigm and more generally in two-dimensional paradigms, one would need to consider some combined index of performance.[28]
4. *The location of the capacity limit must not be blurred by data pooling or averaging across participants.* Since the location of the performance limit on the measurement continuum is a personal parameter whose between-individual distribution is likely to be Gaussian, between-individual data pooling will tend to produce quasi-random distributions. It is to avoid this caveat that only within-participant distributions were considered in Section 9.

The above list of necessary conditions may help explain the curious blind spot of psychology and other fields about the truism of performance extremization. However obvious

[28] For example Shannon's (1948) information theoretic framework makes it possible to compute such a combined index of performance in the classic *RT* paradigm. If we take transmitted, or mutual information, expressed in bits, to measure the accuracy of responses, and average response time, expressed in s, to measure transmission speed, then the ratio of these two quantities, expressed in bits/s, has the desired characteristic. This kind of measure is necessarily upper bounded (for a presentation of similar ideas, see Attneave, 1959, or Miller, 1955). Unfortunately the cost of such a computation is very high, as each data point of the distribution must have been estimated from a sample of pairs of measures.

this characteristic, we have seen that its main empirical effect, namely the clustering of measures near the performer's limit, are not that easy to observe.

Finally, we must ask about the suitability of the classic toolbox we have at our disposal for describing and summarizing samples of extremized measures. We have noted that practitioners of spirometry have been dispensing altogether, for more than a century, with any central-trend estimate of respiratory performance, contenting themselves with session maxima, and we have seen that this practice is perfectly reasonable. We have found essentially the same picture in athletics, where it would seem rather pointless to try to summarize Bolt's performances, say over a season, with an arithmetic mean or a median. Analysis of one-dimensional performance has helped us see that what counts in empirical samples of extremized measures are not averages and distributions bulks, but rather best values and distribution fronts. By the same token, reasons for skepticism have arisen about the relevance of spread statistics such as the standard deviation or the inter-quartile interval centered about the mean and the median, respectively.

One simple and intuitive solution to this statistical description problem might be considered. The idea is simply to assign the role of a *location indicator* to the sample's best value, well in line with the tradition of spirometry, and the role of a *spread indicator* to the distance separating the median from the best value. In this sort of approach best values would serve to estimate capacities of performance while median deviations from bests would serve to estimate effort strengths.

References

Attneave, F. (1959). *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. New York: Holt.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, *7*(4), 145-147.

Donders, (1868). Die Schnelligkeit psychischer Prozesse. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medizin*, Leipzig, Veit, 657–681.

Ebbinghaus H (1880) Urmanuskript "*Ueber das Gedächtniß*". Passau: Passavia Universitätsverlag.

Fechner, G.T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf and Hartel.

Gori, J., Rioul, O., & Guiard, Y. (2018). Speed-accuracy tradeoff: A formal information-theoretic transmission scheme (Fitts). *ACM Transactions on Computer-Human Interaction*, *25*, 1-33.

Gori, J., Rioul, O., Guiard, Y., & Beaudouin-Lafon, M. (2018). The perils of confounding factors: How Fitts' law experiments can lead to false conclusions. *Proceedings of the 2018 ACM CHI Conference on Human Factors in Computing Systems*, 1-10.

Gori, J., & Rioul, O. (2019, September). Regression to a linear lower bound with outliers: An exponentially modified Gaussian noise model. *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, 1-5. IEEE.

Guiard, Y. (1997). Fitts' law in the discrete vs. cyclical paradigm. *Human Movement Science*, *16*, 97-131.

Guiard, Y. (2009). The problem of consistency in the design of Fitts' law experiments: Consider either target distance and width or movement form and scale. *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*, 1809-1818.

Guiard, Y., & Ferrand, T. (1998). Effets de gamme et optimum de difficulté spatiale dans une tâche de pointage de Fitts. *Science et motricité, 34*, 19-25.

Guiard, Y., & Olafsdottir, H. B. (2011). On the measurement of movement difficulty in the standard approach to Fitts' law. *PLoS one*, *6*(10), e24389.

Guiard, Y., Olafsdottir, H. B., & Perrault, S. T. (2011). Fitt's law as an explicit time/error trade-off. *Proceedings of the 2011 ACM CHI Conference on Human Factors in Computing Systems*, 1619-1628.

Guiard, Y., & Rioul, O. (2015). A mathematical description of the speed/accuracy trade-off of aimed movement. *Proceedings of the 2015 British HCI Conference*, 91-100.

Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'Institut Henri Poincaré, 5*, 115-158.

Hart, S. L. (1971). Axiology : Theory of value. *Philosophy and Phenomenological Research, 32*, 29-41.

Hölder, O. (1901). Die Axiome der Quantitat und die Lehre vom Mass. *Berichte iiber die Verhandlungen der Koniglich Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, 53, 1-46.

Hollnagel, E. (2009). *The ETTO Principle: Efficiency-Thoroughness Trade-Off: Why Things that Go Right Sometimes Go Wrong*. Ashgate Publishing.

Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge (MA): MIT press.

Kornblum, S. (1969). Sequential determinants of information processing in serial and discrete choice reaction time. *Psychological Review, 76*, 113.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement* (Vol. 1). New York: Academic Press.

Kugler, P. N., & Turvey, M. T. (2015). *Information, Natural Law, and the Self-Assembly of Rhythmic Movement*. Routledge.

Luce, R. D. (1985). *Response Times and their Role in Inferring Elementary Mental Organization.* New York: Oxford University Press.

Luce, R. D. (1986). Response time distributions in memory search: A caution. In: F. Klix and H. Hagendorf (Eds), *Mechanisms and Performances*, pp. 109-121. Amsterdam: North-Holland.

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., & Keith Smith, J. E. (1988). Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological review, 95*, 340-370.

Michell, J. (1997). Bertrand Russell's 1897 critique of the traditional theory of measurement. *Synthese, 110,* 257–276.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement, 38*, 285-294.

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7-24.

Michell, J. (2014). *An Introduction to the Logic of Psychological Measurement*. Psychology Press.

Michell, J., & Ernst, C. (1996). The Axioms of Quantity and the Theory of Measurement: Translated from Part I of Otto Hölder's German Text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology, 40*, 235-252.

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

Miller, M. R., Hankinson, J. A. T. S., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., ... & Jensen, R. (2005). Standardisation of spirometry. *European Respiratory Journal, 26*, 319-338.

Mottet, D., Guiard, Y., Ferrand, T., & Bootsma, R. J. (2001). Two-handed performance of a rhythmical Fitts task by individuals and dyads. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 1275.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*, 44-64.

Pachella, R. G. (1973). *The Interpretation of Reaction Time in Information Processing Research*. Technical report No. TR-45. Michigan University Ann Arbor Human Performance Center.

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn Jr, J. T. (1979). Motor-output variability: a theory for the accuracy of rapid motor acts. *Psychological Review, 86*, 415.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379–423, 623–656.

Slifkin, A. B., & Eder, J. R. (2017). Degree of target utilization influences the location of movement endpoint distributions. *Acta psychologica*, *174*, 89-100.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta psychologica*, *30*, 276-315.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

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

Wobbrock, J. O., Cutrell, E., Harada, S., & MacKenzie, I. S. (2008). An error model for pointing based on Fitts' law. *Proceedings of the 2008 ACM CHI Conference on Human Factors in Computing Systems*, 1613-1622.

Annex 1. Use of the Performance Concept in Experimental Psychology: A Quantitative Content Analysis of One Issue of JEP: HPP

This quantitative content analysis was undertaken in 2016 with the goal of inventorying the various senses in which the *performance* notion is used in experimental psychology. Of special interest was the incidence of the strict quantitative sense defined at the start of this paper, namely a measure subject to a deliberate extremization effort exerted by a human agent against the resistance of a limit. The word is polysemous, as can be checked in any dictionary, and we will see that if psychologists have recourse to its various senses, they do use the specific quantitative sense of interest here, which general dictionaries of English, notably the Collins[29] and the Merriam-Webster[30], fail to spell out.

The 25 articles that compose the October 2015 issue of the *Journal of Experimental Psychology : Human Perception and Performance 41*(5) were chosen as the test sample. Each article was downloaded from a bibliographic data base and its body, including the title but not the list of references, was pasted in a word processor, where a search was made for occurrences of the seven-character string "PERFORM". All the detected occurrences were subsequently pasted, together with their surrounding sentence context, in a spreadsheet to allow automated processing, using search and text functions.

In English the root PERFORM appears with different endings, yielding two nouns, *performance* and *performer*[31], and one verb appearing in four forms, *perform*, *performs*, *performing*, and *performed*. Table A shows the incidence of these six words in each article of the issue, leaving aside the verb *outperform*, found to occur only three times overall.

Only one of the 25 articles was found to offer zero occurrence of the target root (Tan & Yeh, pp. 1325-1335). In total the root PERFORM wad detected 467 times in the 296 pages of the journal issue, meaning an average incidence of 1.6 occurrences per page, and thus confirming that the performance concept is indeed used routinely in experimental psychology. Table A also shows that 54% of all occurrences of the target string of characters corresponded to the noun *performance*, and 38% to some conjugated form of the (un-prefixed) verb

[29] https://www.collinsdictionary.com/dictionary/english (August 7, 2020).
[30] https://www.merriam-webster.com/dictionary (August 7, 2020).
[31] In experimental psychology the popularity of the noun *performer* is low. 23 (92%) of the 25 articles of the sample issue dispense with it. 36 of the 37 occurrences found in the issue came from one and the same article (Ramenzoni et al., pp. 1209-1222).

perform. The meaning of these two most frequent words (92% of all occurrences) is our focus below.

Insert Table A about here

From a careful inspection of the contexts surrounding the verb PERFORM and the noun PERFORMANCE a clear-cut pattern arose concerning the semantics. It turned out that one may distinguish the quantitative sense of the verb or noun from other, non-quantitative senses using the criterion of word *transitivity*. The rule of thumb is that the concept of performance is taken in a non-quantitative sense when used transitively; it is when used *in*transitively that the concept takes, almost unmistakably, the strict quantitative sense that was defined at the start of this paper.

Regardless of the mode (active vs. passive) and regardless of whether one faces its verb or noun version, it happens to be the case that the performance concept can be used either transitively or intransitively. The concept is used transitively when an agent is said to perform some specified task—or, which amounts to the same, when a specified task is said to be performed by some agent. For example, in an expression like *"performing task X"* (verbal phrase) and likewise in an expression like "*the performance of task X"* (noun phrase), it is easy to see that the task is the direct, accusative object of the action under consideration. However, the agent's performance need not take any direct, accusative object. For example, in expressions like "*participants performed better in condition A than condition B*" (verbal phrase) or "*performance improved from condition A to condition B*" (noun phrase), the performing or performance action is clearly intransitive.

The next two tables illustrate this rule of thumb by offering a few concrete quotes from the reviewed issue. The transitive case considered in Table B is quite straightforward: as explained by the Collins dictionary "*to perform a task or action, especially a complicated one, is to do it*" (sense #1 of the verb PERFORM). What we call the transitive use of the *noun* is defined by the Collins accordingly: "*the performance of a task is the fact or action of doing it*" (sense #6 of the noun PERFORMANCE).

Table C, however, reveals that the intransitive notion of performance, whether expressed verbally or nominally, can refer to two markedly different senses. Fortunately, however, these are fairly easy to tell apart based on context. One intransitive sense is the entertainment of an audience in such contexts as drama, music, and dance performance—the Collins, along with its peers, does not omit to mention this sense.[32] The other intransitive sense is the strictly quantitative one that dictionaries fail to identify.

Insert Tables B and C about here

In Table D the classification is applied to each of the 25 articles of the reviewed issue. We can see that while the verb PERFORM is used in the transitive way almost invariably (in 93% of occurrences), the opposite is true of the noun PERFORMANCE, used most of the time in the intransitive way (in 94% of occurrences). Thus there is little doubt that it is mostly the noun (231 intransitive uses in 20 articles), and only occasionally the verb (12 intransitive uses in four articles), that psychologists take in the quantitative sense.

Insert Table D about here

The considerable variability of frequencies across the lines of Table D raises the question of a possible link between the incidence of the transitive vs. intransitive use of the performance notion and the content of the reviewed studies, all based on experiments with human volunteers. Table E reproduces the statistics of Table D but this time side to side with a rough description of the contents of the studies, with special attention paid to dependent measures.

[32] However, the list of examples offered by the Collins to illustrate the first sense of the noun PERFORMANCE mix up somewhat confusingly transitive cases like "*a performance of Bizet's Carmen*" with intransitive cases like "*her performance as the betrayed Medea*".

Insert Table E about here

Unsurprisingly, the vast majority of the articles (22 of 25) report genuine performance studies in which participants are instructed to extremize at least one quantitative score. Ten of these studies of human performance asked participants to maximize an accuracy score (e.g., a percentage of correctly recognized items) or to minimize an error score (e.g., a mean estimation error), with no time pressure. In the other 12 performance studies, a time score—more often than not a latency—had to be minimized while an error rate had to be maintained as low as possible.

The three exceptions are the studies of Romero et al. (pp. 1223-1235), Gagnon et al. (pp. 1385-1395) and Ma et al. (pp. 1409-1419). Romero et al. inquired into the structure of joint-angle variance in mutually aimed movements performed by dyads. Although the aiming task was to be carried out as fast and as accurately as possible, the time and error data were considered irrelevant and ignored in the results section. Gagnon et al. examined the influence of somebody else's body size on self-affordance judgments concerning whether or not an aperture is passable. The dependent variable of the experiment was the threshold between no and yes responses, with no extremized quantity considered. Finally Ma et al. explored the ability of people to extract information from non-fixated target words in Chinese reading. If the authors mention that the comprehension accuracy rate vas very high overall, they report no extremized quantity, their main focus being the duration of eye fixations.

Insert Table F about here

In Table F the frequency data are grouped so as to contrast, with regard to the relative frequency of the transitive s. intransitive use of the performance concept, the studies based on the explicit measurement of extremized scores (22 articles totaling 261 pages) vs. those which dispense with that kind of measurement (three articles totaling 35 pages). For use of the verb,

the two families of studies do not differ: whether or not the study is based on performance measurement, the verb PERFORM is nearly always used transitively. In contrast, the two families of studies differ markedly in the way they employ the noun PERFORMANCE. Whereas in performance-based studies the noun is used almost always (in 97% of the cases) intransitively (as, e.g., in "*performance improved*"), the trend is opposite in the other family of studies, prone to use the noun transitively (62%) more often than intransitively (38%).[33]

[33] A supporter of null hypothesis statistical significance testing may like to learn that the probability of such a departure from chance assuming no statistical link between the two binary variables of the table is $p < .0001$ (1-df Chi square test with or without the Yates correction for continuity).

Table A. Incidence of Nouns and Verbs composed with Root PERFORM in Reviewed Issue

			Noun forms			Verb forms					
Article	First page	First author	PERFORMance	PERFORMer	Total	PERFORM_	PERFORMs	PERFORMing	PERFORMed	Total	All forms
#1	1179	Royer	9	0	9	0	0	0	2	2	11
#2	1184	Reynolds	11	0	11	0	0	0	0	0	11
#3	1190	Beck	0	0	0	0	0	1	0	1	1
#4	1197	Leiva	12	0	12	0	0	0	1	1	13
#5	1203	Ward	0	0	0	0	0	0	2	2	2
#6	1209	Ramenzoni	18	36	54	6	0	13	19	38	92
#7	1223	Romero	12	0	12	5	0	6	4	15	27
#8	1236	Brattan	15	0	15	10	0	6	12	28	43
#9	1247	Z_Sun	31	0	31	0	0	0	3	3	34
#10	1260	Maclean	5	0	5	2	0	0	2	4	9
#11	1271	De la Malla	12	0	12	0	3	1	3	7	19
#12	1281	Lupker	1	0	1	0	0	0	0	0	1
#13	1300	Mills	9	0	9	4	0	0	2	6	15
#14	1315	Schneider	6	0	6	0	0	0	3	3	9
#15	1325	Tan	0	0	0	0	0	0	0	0	0
#16	1336	Maes	38	0	38	11	0	0	16	27	65
#17	1353	Rajsic	8	1	9	0	0	0	1	1	10
#18	1365	Ma	10	0	10	0	0	0	2	2	12
#19	1376	Hung	8	0	8	0	0	0	1	1	9
#20	1385	Gagnon	1	0	1	3	0	5	6	14	15
#21	1396	Prinzmetal	16	0	16	1	0	0	0	1	17
#22	1409	Ma	0	0	0	0	0	0	4	4	4
#23	1420	Miller	7	0	7	5	0	0	6	11	18
#24	1442	Zupan	17	0	17	0	0	0	6	6	23
#25	1462	Marsh	6	0	6	1	0	0	0	1	7
		Total	252	37	289	48	3	32	95	178	467
		% of all forms	54.0%	7.9%	61.9%	10.3%	0.6%	6.9%	20.3%	38.1%	100.0%

Table B. Illustrations of the Meaning of Verb PERFORM and Noun PERFORMANCE Used Transitively

Word	Examples from Reviewed Issue	Semantic classification		
		Collins	Merriam-Webster	Deliberately extremized measure (DEM)
Verb PERFORM	*A point-wise multiplication was then performed* (Royer & Blais, p. 1180)	Def. #6 of verb	Def. #2 of verb	Not a DEM
	observing someone perform an action (Brattan et al., p. 1237)	Def. #6 of verb	Def. #2 of verb	Not a DEM
	we performed a 2x2 split-plot ANOVA (Ma et al., p. 1371)	Def. #6 of verb	Def. #2 of verb	Not a DEM
Noun PERFORMANCE	*during performance of a discrete joint action* (Romero et al., p. 1223)	Def. #6 of noun	Def. #1a of noun	Not a DEM
	the performance of arm movements between key presses (Maes et al., p. 1343)	Def. #6 of noun	Def. #1a of noun	Not a DEM
	during performance of a piano piece (Maes et al., p. 1343)	Def. #6 of noun	Def. #1a of noun	Not a DEM

Table C. Illustrations of the Meaning of Verb PERFORM and Noun PERFORMANCE Used Intransitively

Word	Context	Examples from Reviewed Issue	Semantic classification: Collins	Semantic classification: Merriam-Webster	Deliberately extremized measure (DEM)
Verb PERFORM	**Arts or drama**	*musicians* […] *often perform under heightened cognitive load* (Maes et al., p. 1337)	Def. #4 (US)	Def. #2 of intrans. verb	Not a DEM
	Measurement	*Participants performed significantly worse with crowded stimuli* (Ma et al., p. 1370)	unidentified	unidentified	DEM
		participants performed in a way that was not strategically optimal (Rajsic et al., p. 1362)	unidentified	unidentified	DEM
		In the egocentric condition, participants were able to perform reasonably well (Ramenzoni, p. 1213)	unidentified	unidentified	DEM
Noun PERFORMANCE	**Arts or drama**	*joint music and dance performance* (Ramenzoni et al., p. 1219)	Def. #1 of noun	Def. #3b of noun	Not a DEM
		in the context of cello performance (Maes et al., p. 1337)	Def. #1 of noun	Def. #3b of noun	Not a DEM
	Measurement	*Under such conditions, change detection performance is quite poor* (Reynolds & Withers, p. 1184)	unidentified	unidentified	DEM
		enhancing no-go performance when they are relevant (Leiva, p. 1197)	unidentified	unidentified	DEM
		there was no decrement in memory performance (Sun et al., p. 1247)	unidentified	unidentified	DEM

Table D. Transitive vs. Intransitive Use of Verb PERFORM and Noun PERFORMANCE in Reviewed Issue

			Verb PERFORM			Noun PERFORMANCE			
Article #	First page	First author	Transitive action	Intransitive action	Total	Transitive action	Intransitive action	Total	Grand total
1	1179	Royer	2	0	2	0	9	9	11
2	1184	Reynolds	0	0	0	0	11	11	11
3	1190	Beck	1	0	1	0	0	0	1
4	1197	Leiva	1	0	1	0	12	12	13
5	1203	Ward	2	0	2	0	0	0	2
6	1209	Ramenzoni	30	8	38	1	15	16	54
7	1223	Romero	15	0	15	7	5	12	27
8	1236	Brattan	28	0	28	0	14	14	42
9	1247	Z_Sun	4	0	4	0	30	30	34
10	1260	Maclean	4	0	4	0	5	5	9
11	1271	De la Malla	7	0	7	0	12	12	19
12	1281	Lupker	0	0	0	0	1	1	1
13	1300	Mills	6	0	6	0	8	8	14
14	1315	Schneider	3	0	3	0	6	6	9
15	1325	Tan	0	0	0	0	0	0	0
16	1336	Maes	25	2	27	4	34	38	65
17	1353	Rajsic	0	1	1	0	8	8	9
18	1365	Ma	1	1	2	0	10	10	12
19	1376	Hung	1	0	1	0	8	8	9
20	1385	Gagnon	14	0	14	1	0	1	15
21	1396	Prinzmetal	1	0	1	1	14	15	16
22	1409	Ma	4	0	4	0	0	0	4
23	1420	Miller	11	0	11	0	6	6	17
24	1442	Zupan	6	0	6	0	17	17	23
25	1462	Marsh	1	0	1	0	6	6	7
		Total	167	12	179	14	231	245	424
			93%	7%	100%	6%	94%	100%	
					42%			58%	100%

Table E. Transitivity Statistics and Content Analysis, with Special Focus on Dependent Measures

Article #	First page	First author	Paper title	Main task	Dependent Measure(s)	Verb **PERFORM**		Noun **PERFORMANCE**		Total
						Transitive action	Intransitive action	Transitive action	Intransitive action	
1	1179	Royer	When Less Is More: Impact of Face Processing Ability on Recognition of Visually Degraded Faces	face recognition, match-to-sample (same/diff)	error rate	2	0	0	9	11
2	1184	Reynolds	Understanding the Relationship Between Implicit and Explicit Change Detection: Evidence From Scan Path Data	detection of change in peripheral visual field	RT, error rate, N fixations	0	0	0	11	11
3	1190	Beck	Evidence for Negative Feature Guidance in Visual Search Is Explained by Spatial Recoding	visual search	RT, error rate	1	0	0	0	1
4	1197	Leiva	Reorienting the Mind: The Impact of Novel Sounds on Go/No-Go Performance	go/no-go speeded reaction, press one of two keys	RT, error rate (% misses, % false alarms)	1	0	0	12	13
5	1203	Ward	The Rubber Hand Illusion Depends on the Tactile Congruency of the Observed and Felt Touch	estimate finger position	subjective estimate of finger position	2	0	0	0	2
6	1209	Ramenzoni	Synchronous Imitation of Continuous Action Sequences: The Role of Spatial and Topological Mapping	get coupled with rhythmic model	tap ratio (ideal = 1), mean asynchrony, percent match	30	8	1	15	54
7	1223	Romero	Can Discrete Joint Action Be Synergistic? Studying the Stabilization of Interpersonal Hand Coordination	mutual pointing in dyads	variance per degree of freedom	15	0	7	5	27
8	1236	Brattan	Spatiotemporal Judgments of Observed Actions: Contrasts Between First- and Third-Person Perspectives After Motor Priming	judge whether video sequence after interruption restarts early, fine, or late	% late judgments	28	0	0	14	42
9	1247	Sun	How to Break the Configuration of Moving Objects? Geometric Invariance in Visual Working Memory	Detect change in motion of a configuration and memorize the motion	error rate	4	0	0	30	34
10	1260	Maclean	Does Oculomotor Readiness Mediate Exogenous Capture of Visual Attention?	Is the target character present in periph visual field? press one of two keys	RT, error rate	4	0	0	5	9
11	1271	De la Malla	Predictive Plus Online Visual Information Optimizes Temporal Precision in Interception	catch an approaching virtual tennis ball, using a data glove	length of timing error	7	0	0	12	19
12	1281	Lupker	Is There Phonologically Based Priming in the Same-Different Task? Evidence From Japanese-English Bilinguals	same/different speeded reaction, press one of two keys	RT, error rate	0	0	0	1	1
13	1300	Mills	Effects of Task and Task-Switching on Temporal Inhibition of Return, Facilitation of Return, and Saccadic Momentum During Scene Viewing	fixate the probe as quickly and accurately as possible whenever it occurs	Saccadic RT, magnitude of saccadic endpoint error	6	0	0	8	14
14	1315	Schneider	Attentional Control of Response Selection in Task Switching	speeded binary categorization (e.g. living vs. non-living), press one of two keys	RT, error rate	3	0	0	6	9
15	1325	Tan	Audiovisual Integration Facilitates Unconscious Visual Scene Processing	has the scene appeared above of below the cross? press one of two keys	RT, error rate	0	0	0	0	0
16	1336	Maes	Auditory and Motor Contributions to the Timing of Melodies Under Cognitive Load	get coupled with the model rhythm and continue	length of intertap interval	25	2	4	34	65
17	1353	Rajsic	Confirmation Bias in Visual Search	which color did the stimulus appear in? press one of two keys	RT, error rate	0	1	0	8	9
18	1365	Ma	A Deficit Perceiving Slow Motion After Brain Damage and a Parallel Deficit Induced by Crowding	main study (Exp3): indicate motion direction (left or right), press one of two keys	error rate	1	1	0	10	12
19	1376	Hung	Syntactic Processing in the Absence of Awareness and Semantics	localization and memory task, press one of two keys	error rate	1	0	0	8	9
20	1385	Gagnon	The Influence of Social Context and Body Size on Action Judgments for Self and Others	is the gap passable? verbal yes/no response	% of yes responses	14	0	1	0	15
21	1396	Prinzmetal	Spatial Attention and Environmental Information	was the stimulus an "F" or a "T", press one of two keys	RT, error rate	1	0	1	14	16
22	1409	Ma	Readers Extract Character Frequency Information From Nonfixated-Target Word at Long Pretarget Fixations During Chinese Reading	read Chinese words	gaze fixation duration	4	0	0	0	4
23	1420	Miller	A Comparison of the Psychological Refractory Period and Prioritized Processing Paradigms: Can the Response-Selection Bottleneck Model Explain Them Both?	binary choice reactions to letter identitity and stimulus color	RT, error rate	11	0	0	6	17
24	1442	Zupan	Inhibition in Time-Based Visual Selection: Strategic or by Default?	indicate the location (left/right) of the blue square, press one of two keys	RT, error rate	6	0	0	17	23
25	1462	Marsh	Dynamic Cognitive Control of Irrelevant Sound: Increased Task Engagement Attenuates Semantic Auditory Distraction	free order recall of visual material after 15 presentations	error rate	1	0	0	6	7
	Total					167	12	14	231	424

Table F. Incidence of Transitive vs. Intransitive Use of Performance Notion in Studies Measuring vs. Not Measuring Extremized Performance

		Verb PERFORM			Noun PERFORMANCE			Sum		
		Transitive action	Intransitive action	Total	Transitive action	Intransitive action	Total	Transitive action	Intransitive action	Grand total
Extremized performance measurement: 22 articles, 261 pages	*N*	134	12	146	6	226	232	140	238	378
	%	92%	8%	100%	3%	97%	100%	37%	63%	100%
No measurement of extremized performance: 3 articles, 35 pages	*N*	33	0	33	8	5	13	41	5	46
	%	100%	0%	100%	62%	38%	100%	89%	11%	100%
Total over 25 articles, 296 pages	*N*	167	12	179	14	231	245	181	243	424
	%	93%	7%	100%	6%	94%	100%	43%	57%	100%