Understanding the Within-Individual Variability of Forced Vital Capacity: An Exploitation of the NHANES III Spirometry Data

Yves Guiard

LISN, CNRS & Université Paris-Saclay
LTCI, Télécom Paris, Institut Polytechnique de Paris
yves.guiard@telecom-paris.fr

Abstract

While there is an abundant literature on the distribution of spirometry statistics in various subsets of the human population, apparently little is known of the structure of the typically very small sample of measures that can be gathered during a spirometry session. This paper starts with a theoretical analysis of the relation linking the measure of forced vital capacity (*FVC*) to the parameter of total lungs capacity (*TLC*). Since the maximization effort exerted on *FVC* measures by the testees is opposed by the resistance of *TLC*, their impassable personal upper limit, a ceiling effect must take place on the continuum of *FVC* measurement. Two predictions follow concerning the within-subject distribution of *FVC.* One is that the distribution should be negatively skewed, the other is that its first and second moments should correlate negatively across sessions. These predictions were tested with the publicly available large-scale spirometry data collected by the Third National Health and Nutrition Examination Survey. Using original data processing techniques especially devised to unveil the shape of small session samples of *FVC* measures, the paper reports highly consistent confirmatory evidence, based on the analysis of thousands of individual test sessions, that a typical session sample of *FVC* is indeed strongly skewed negatively and that the session mean and the session standard deviation of *FVC* do indeed bear a strong negative correlation. Several implications of these results are discussed, some of which cut across the frontiers of respirology. It is suggested that the procedural rigor and simplicity of spirometry testing make it a privileged paradigm for understanding quantitative performance measurement in general.

1. Introduction

A spirometry measure is a meaningless number unless it can be situated in the frame of reference of its natural variation among humans. Thus an abundant statistical literature has accumulated providing reference curves aimed to inform practitioners about the admittedly normal range of variation of spirometry statistics in different human subpopulations, taking into account such factors as age and stature, gender, and ethnicity (e.g., Hankinson et al., 1999; Stanojevic et al., 2008; Quanjer et al., 2012; Rochat et al., 2013; Coates et al., 2016).

Statistical sampling theory distinguishes three sorts of numerical entities: (1) *basic measures*, which come in finite samples; (2) *summary statistics*, such as the sample's mean and the sample's standard deviation, which serve to compress the empirical information contained in a sample of measures; and (3) *parent population parameters*, such as the population's mean and standard deviation, which we often want to estimate inductively from summary statistics. Statistical sampling takes a special form in spirometry. While their basic observation is the measure of forced vital capacity (*FVC*) given by the spirometer,[1] practitioners use a single summary statistic, which is not an average but an extremum, namely the maximum of the session's sample of *FVC* measures (FVC_{max}). The relevant parent population parameter here is total lungs capacity (*TLC*), the upper limit on which FVC_{max} would gradually converge were it possible to obtain, in an unending session, an infinite sample of measures from the same subject.

While the statistical literature on spirometry revolves about the between-individual sort of variability, this paper, in contrast, is mainly concerned with the *within-individual* variability of *FVC* measures, the variability observable across the successive maneuvers of a spirometry session. Although obviously different, the between-individual and the within-individual variability problems are, in one regard, equivalent: both raise a statistical sampling problem, that of inferring inductively properties of a parent distribution from a limited sample of observations. In our treatment of the within-individual distribution of *FVC* below we will have in mind a parent population of maneuvers rather than a parent population of human individuals.

Comparatively little attention seems to have been paid to the problem of the within-subject, within-session variability evident in every single spirometry session. That problem looks intractable at first sight as the sample of measures that can be actually gathered in a session is typically so small as to defy any statistical description—one may wonder what could be learned from a frequency histogram constructed with only three or four measures. Nevertheless the very fact of asking, as we will do, about the mechanisms that explain the within-subject variability of *FVC* across successive maneuvers implies the assumption that, however small the session sample, there exists a parent population of within-session *FVC* measures.

A word of terminology is in order. Statistically speaking the problem of spirometry is remarkably simple, a session delivering just one sample of *FVC* measures from just one subject. Therefore the adjectival expressions "within-session" and within-subject", henceforth noted "W-S", will be used here as synonyms and so will the expressions "between-session" and "between-subject", jointly noted "B-S".

[1] For simplicity, this paper focuses on *volumetric* spirometry, flow measurements being essentially left aside.

2. Respiratory Capacity, Effort, and Performance

Forced vital capacity (*FVC*),[2] a widely used measure of spirometry, is defined as "the maximal volume of air exhaled with maximally forced effort from a maximal inspiration" (Brusasco, Crapo, & Viegi, 2005, p. 321; Quanjer et al., 1993, p. 11).

While standardized instructions insistently ask the testee for a *maximal* inspiration and expiration effort, practitioners encounter the difficulty that the testee's effort is never strictly maximal, varying erratically from maneuver to maneuver. Let us assume that the magnitude of this effort, noted *E*, ranges from 0% to 100%. Were the maximum effort requirement perfectly met, with *E* invariably equal to 100%, the maneuvers of a session would all deliver the same value, one that would each time coincide exactly with what respirologists call the *total lungs capacity* (*TLC*)—the volume of air that the testee's lungs and airways can physically contain. However, the testee's effort is never exactly total and so practitioners must content themselves with the fact that the measure they record is almost surely less than *TLC*:

$$FVC \leq TLC. \tag{1}$$

Reflecting the size and the functional state of the pulmonary apparatus, *TLC* is a testee-specific parameter. In these pages it will be considered an anthropometric parameter whose value is fixed during a spirometry session, just like, say, body weight. Quite unlike body weight, however, *TLC* cannot be *measured* by the practitioner. It is an unknown constant whose inductive estimation from a samples of *FVC* measures constitutes the main goal of volumetric spirometry. Inequality 1 says that *TLC* constitutes the *upper bound* of the *FVC* measure.

Obviously the measure *FVC* depends on the capacity *TLC*. The most plausible model of this dependency is a linear function whose slope is given by the magnitude of the testee's effort:

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

The multiplication of a constant by a random variable yields a random variable: the source of the haphazard variability of *FVC* is indeed the haphazard variability of *E*. At this point it must be recalled that the maximal expiratory maneuver requires two consecutive efforts, an inspiration effort E_{insp} followed by an expiration effort E_{exp}, neither of which can be strictly maximal. Since the two percentages combine multiplicatively

$$E = E_{insp} \times E_{exp}, \tag{3}$$

occasionally the value of *E* in Equation 2 may be problematically low. Suppose that in a maneuver the testee makes two decent efforts, for example E_{in} = 80% and then E_{ex} = 90%. The product of these two efforts will be *E* = 80% × 90% = 72%, yielding a non-negligible mismatch between the measured value of *FVC* and *TLC*.

[2] Incidentally, the traditional terminology of spirometry is slightly misleading. If the expression "total lungs capacity" seems quite appropriate to designate what may 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, the term "capacity" is somewhat unfortunate in the expression "forced vital capacity" or "forced capacity" because the latter quantity, consisting of a performance measure which varies from maneuver to maneuver depending on the strength of the testee's effort, is a capacity in neither sense.

Letting ε denote that underestimation error,

$$\varepsilon = FVC - TLC, \tag{4}$$

we can see from Equations 2 and 4 that it varies as an affine function of the testee's effort

$$\varepsilon = (E \times TLC) - TLC, \tag{5}$$

whose slope and intercept are both given by the unknown constant *TLC*. The estimation error increases linearly from 0% (0 ml) to 100% (–*TLC*) as the effort declines from 100% to 0% (Figure 1).

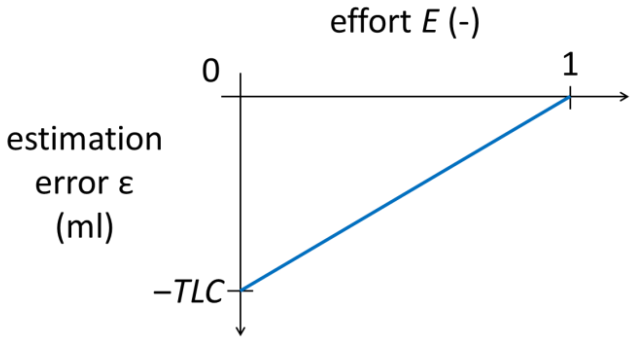

Figure 1. The error made in estimating *TLC* from *FVC* as a function of the magnitude of the testee's effort.

Thus, the reason why practitioners, following the recommendations of established standards (Graham et al., 2019), so insistently urge their patients to maximize their respiratory efforts seems clear: the stronger the effort, the closer the value of *FVC* to that of *TLC* and hence the smaller the practitioner's error in estimating *TLC*. By the same token, Equation 5 explains why the standards of spirometry also ask practitioners to summarize the various measures of a session with the session maximum: FVC_{max} is indeed the session's best value, the one which estimates *TLC* with the smallest error.

What has just been proposed is an idealized and schematic model of the relationships linking the three important quantities of volumetric spirometry, the effort, the capacity and the measure. In particular there is little doubt that the physical capacity of the lung does not result in a *strictly* fixed upper bound on the continuum of *FVC* measurement, if only because of the elasticity of the various tissues involved in the spirometry maneuver. Nevertheless, this simplified, heuristic conceptual framework will help us formulate the statistical problem at hand and then guide our exploration of empirical data.

3. A Ceiling Effect in FVC Measurement

If it is assumed that (1) testees do try their best upon each maneuver to produce as close a value of *FVC* as possible to their personal *TLC* limit, that (2) the magnitude of their maximization effort varies randomly from maneuver to maneuver and that (3) throughout the

test session *TLC* represents a fixed upper bound on the continuum of *FVC* measurement, then one must expect a *ceiling effect* in W-S distributions of *FVC.*[3]

The expected effect is reminiscent of that examined by psychologist George Miller (1956) in his famous paper on the limited capacity of humans for transmitting information. Miller discussed the widely replicated finding that as the information content of a stimulus is gradually increased in absolute identification tasks, the volume of information per judgment effectively transmitted by experimental participants levels off at about 2.5 bits (7±2 items). Such a ceiling effect, Miller explained, reflects the existence of an impassable upper limit in the human information-transmission capacity, best modeled mathematically by what Shannon (1948) called the capacity of an information transmission channel.

Below we will focus on two tightly related, yet independently testable predictions concerning the W-S distribution of *FVC,* which follow from the above assumptions and constitute two different expressions of the same ceiling effect. The most obvious prediction is that the W-S distribution of *FVC* should be *negatively skewed*. In general the testees are willing to comply with the instructions they receive, and thus they should tend to accumulate their small samples of *FVC* values not far from their personal upper limits. In other words there should be an abrupt, non convex *front* on the right-hand side of the W-S distribution, constrained by the hard wall of an impassable upper bound, and an evanescent, convex *tail* on the left-hand side, not constrained by a lower bound. Since the skewness of a distribution measures the relative extensions of its two tails,[4] obviously a distribution whose values tend to cluster in the vicinity of a fixed upper limit will be left or negatively skewed.

The justification of the metaphor of a *front* observable on the bounded side of performance distributions (Guiard, 2020; Guiard, Olafsdottir, & Perrault, 2011; Guiard & Rioul, 2015) is rather straightforward: to accumulate their *FVC* values as close as they can to their *TLC* limit is, after all, precisely what testees are explicitly instructed to do. If the maximization effort of spirometry is conceptualized as a physical force oriented upward, then the front can be defined as the region of the continuum of *FVC* measurement where that force meets the resistance of the capacity limit. For the *FVC* scores of a session the *TLC* parameter plays a dual role: it is, by definition, a *global attractor*, testees being supposed to push each of their *FVC* measures as close as possible to that limit; but at the same time it is a *local repeller* in the sense that pushing one's *FVC* value closer and closer to one' limit means experiencing a harder and harder repelling reaction.

The second prediction we will investigate below is that the mean and the standard deviation of *FVC* measures should tend to *correlate negatively* across sessions. That correlation (henceforth referred to as the MS correlation) should exist and be negative because the stronger the maximization effort made in a session, the higher the mean of *FVC* but at the same time, due to the ceiling effect, the more nearly deterministic the *FVC* value. At the limit, were the subject's effort an invariable 100%, the mean of *FVC* would reach its absolute maximum of *TLC* while the variance of *FVC* would reach its absolute minimum of zero.

[3] The sort of ceiling effect considered here should not be confused with the technological artefact reported in situations where a measurement device fails to completely cover the relevant range of measurement. This is not the case of properly calibrated spirometers (Madsen, 2012).

[4] See for example https://mathworld.wolfram.com/search/?query=skewness&x=12&y=12

The prediction can also be explained in terms of respiratory capacity rather than effort. Forming groups of subjects with more and more homogeneous capacities is pretty much like deblurring the upper limit of *FVC*. While there should be little or no MS correlation on *FVC* across sessions run by subjects having a whole diversity of respiratory capacities, the expected negative correlation should become observable in sufficiently homogeneous groups of capacities.

4. The NHANES III Spirometry Data

The results to be presented below exploit the very large set of spirometry data made publicly available by the US Center for Disease Control and Prevention.[5] The data, collected in 1988-94 by trained technicians during the Third National Health and Nutrition Examination Survey (NHANES III), come from about 20,000 subjects of both genders, aged 8 years and over, selected from households across the United States (Hankinson et al., 1999).

Two files were of special interest for the present purposes. The file named SH3SPIRO.csv, released in June 2001, contains detailed quantitative spirometry data on each maneuver of each testee required to perform at least five technically satisfactory maneuvers. The other file, named GROWTHCH.xpt, released in November 2012, contains rich anthrometric information on each testee. The survey having assigned a unique identification number to each individual testee, it was possible to merge the two files. In the analyses below every single value of *FVC* came from a testee whose age, gender, weight, and standing height were known.[6]

Table 1. Age and Gender Composition of the Data Set

COUNT OF SUBJECTS

	Age (years)																			
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	8-17 years	18-25 years	All
Males	233	267	291	281	201	184	187	184	195	189	170	164	143	158	155	170	179	166	2 212	1 305	3 517
Females	226	267	252	269	232	218	219	197	214	210	185	188	173	181	180	195	197	148	2 304	1 447	3 751
Both genders	459	534	543	550	433	402	406	381	409	399	355	352	316	339	335	365	376	314	4 516	2 752	7 268

COUNT OF MANEUVERS

	Age (years)																			
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	8-17 years	18-25 years	All
Males	1 652	1 902	2 065	1 899	1 340	1 178	1 192	1 181	1 241	1 137	1 077	996	857	966	923	1 034	1 080	1 047	14 787	7 980	22 767
Females	1 574	1 886	1 801	1 844	1 546	1 507	1 478	1 300	1 449	1 379	1 189	1 156	1 094	1 178	1 118	1 242	1 222	887	15 764	9 086	24 850
Both genders	3 226	3 788	3 866	3 743	2 886	2 685	2 670	2 481	2 690	2 516	2 266	2 152	1 951	2 144	2 041	2 276	2 302	1 934	30 551	17 066	47 617

MEAN NUMBER OF MANEUVERS PER SUBJECT

	Age (years)																			
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	8-17 years	18-25 years	All
Males	7.09	7.12	7.10	6.76	6.67	6.40	6.37	6.42	6.36	6.02	6.34	6.07	5.99	6.11	5.95	6.08	6.03	6.31	6.63	6.11	6.47
Females	6.96	7.06	7.15	6.86	6.66	6.91	6.75	6.60	6.77	6.57	6.43	6.15	6.32	6.51	6.21	6.37	6.20	5.99	6.83	6.27	6.62
Both genders	7.03	7.09	7.12	6.81	6.67	6.68	6.58	6.51	6.58	6.31	6.38	6.11	6.17	6.32	6.09	6.24	6.12	6.16	6.74	6.20	6.55

The merged csv file prepared for this study includes a total of 47,617 measures of *FVC* collected in 7,268 male and female subjects aged 8-25 years (see Table 1). More often than

[5] https://wwwn.cdc.gov/nchs/nhanes/nhanes3/default.aspx (Series 11 No. 9A).

[6] Thanks are due to Francisco Grisanti, who carried out the merging of the two files in partial fulfilment of a Master in computer science of the University of Houston during a 4-month stay at the University of Paris-Saclay under the supervision of this writer: Grisanti, F. (2018). *Development of a User Interface for Access to Biometric and Spirometry Data from the NHANES III Survey*. Unpublished Master thesis.

not below the data from just adults will be more than enough to settle the empirical facts of interest.

The samples of data gathered in the spirometry sessions of the NHANES III survey were slightly larger than those typically gathered by clinicians, the technicians being instructed to obtain at least five satisfactory maneuvers, and so the survey reports an average of 6.5 successful maneuvers per session. This sample size remains rather small for an analysis of the W-S variability of *FVC*. Pooling many individual samples is not a solution because the distribution we are curious about will be drowned in a large amount of B-S variability. One solution that was devised for the present study capitalizes on the idea that different individuals with the same respiratory capacity are, from the viewpoint of volumetric spirometry, like *clones*. A sample of *FVC* measures from many individuals with the same respiratory capacity is essentially equivalent to a sample of *FVC* measures from many sessions run by one and the same individual. We will see that the capacity-cloning technique makes it possible to unveil some highly consistent patterns of *FVC* readily interpretable as resulting from the interplay of the subject's effort and capacity.

5. Results

We will examine first the skewness of the W-S distribution of *FVC* (Sub-section 5.1) and then the correlation linking the means and standard deviations of session samples of *FVC* (Sub-section 5.2).

5.1. Skewness in the W-S Distribution of FVC scores

One simple way to estimate the skewness of the W-S distribution of *FVC* is to compute the skewness coefficient[7] for each single session and to then examine the distribution of that statistic across sessions. Since on average 6.5 maneuvers were performed per session in the NHANES survey and 99% of all sessions contained more than three maneuvers, it was possible to estimate sample skewness in two comfortably large samples of independent sessions. The skewness statistic was computed for 1,302 and 1,430 sessions with male and female adults, respectively.

Figure 2 plots the distribution of that sample statistic, confirming that the NHANES data contains many more negatively-skewed than positively-skewed samples of *FVC* scores. The ratio is 3/1 in males and 4/1 in females, the median value of sample skewness being –0.73 and –0.79, respectively.

[7] The usual parametric formula was used: $\gamma = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$.

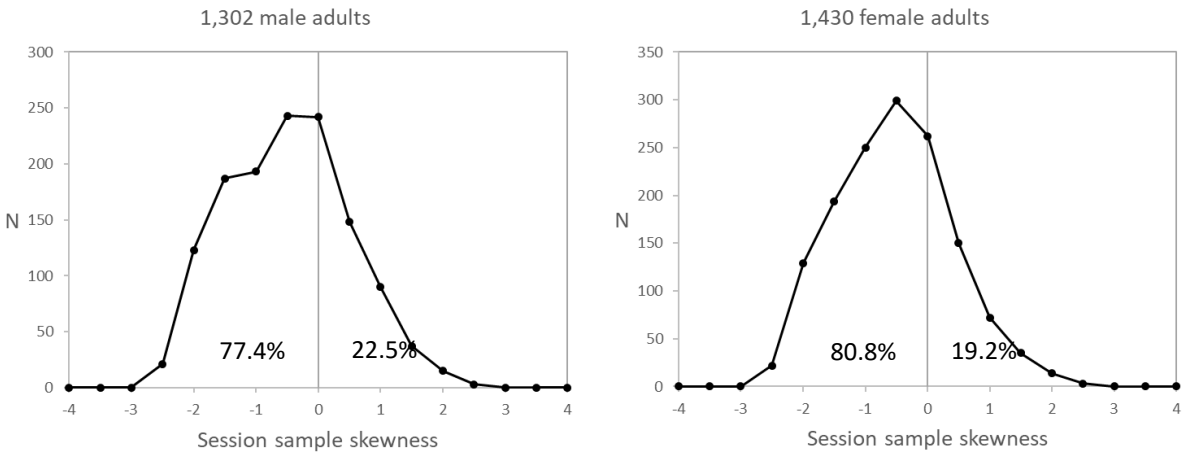

Figure 2. Distribution of the skewness coefficient over all sessions, separately for male and female adults.

Finer evidence is given in Figure 3, which plots the mean value of the session skewness statistic separately for each of the 18 age groups available for each gender, thus providing 36 statistically independent estimations. The negative skewness hypothesis is massively corroborated, all group averages of the skewness statistic falling well below zero, in the range from – 0.50 to – 0.74 for males and from – 0.61 and – 0.83 for females.

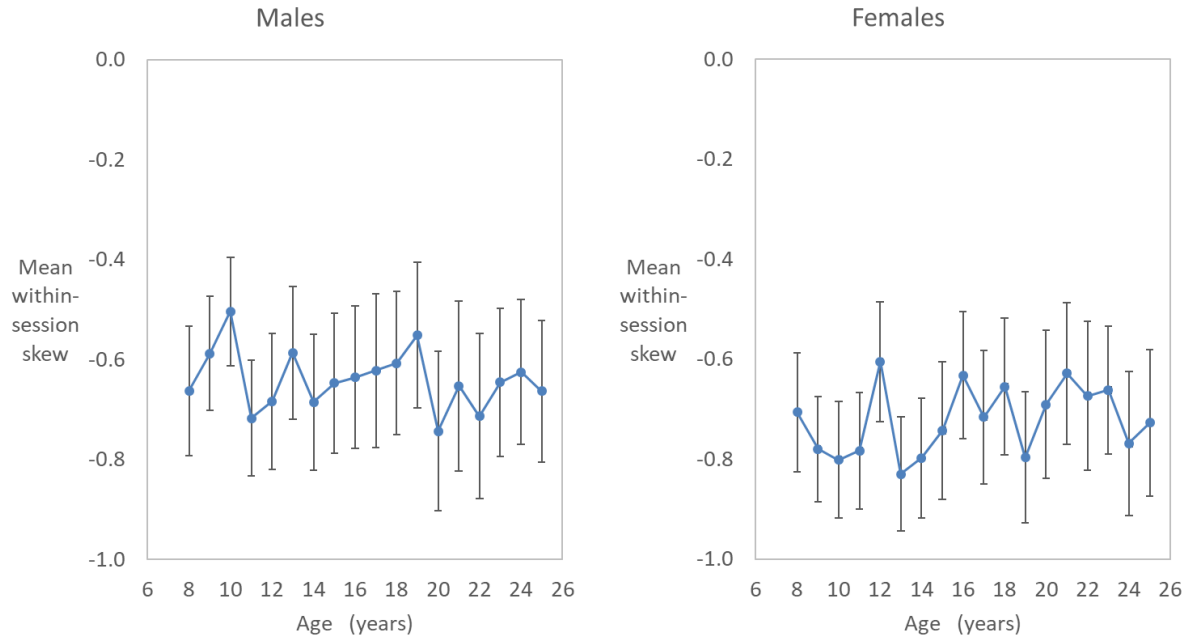

Figure 3. Mean session skewness computed for each age group within each gender. Error bars represent 95% confidence limits of the means.

Thus Figures 2 and 3 provide strong evidence that spirometry sessions do produce negatively skewed samples of *FVC*. We may now ask about the relation between skewness and effort. A fact familiar to practitioners is that willingness to spend a physical effort like that required in a spirometry test is not guaranteed, some testees accepting less whole heartedly than others the maximal effort instructions (e.g., NHANES, 2011). One simple statistic to characterize the

general level of effort spent in a session is the distance from the session's median (FVC_{med}) to the session's highest value of *FVC*.

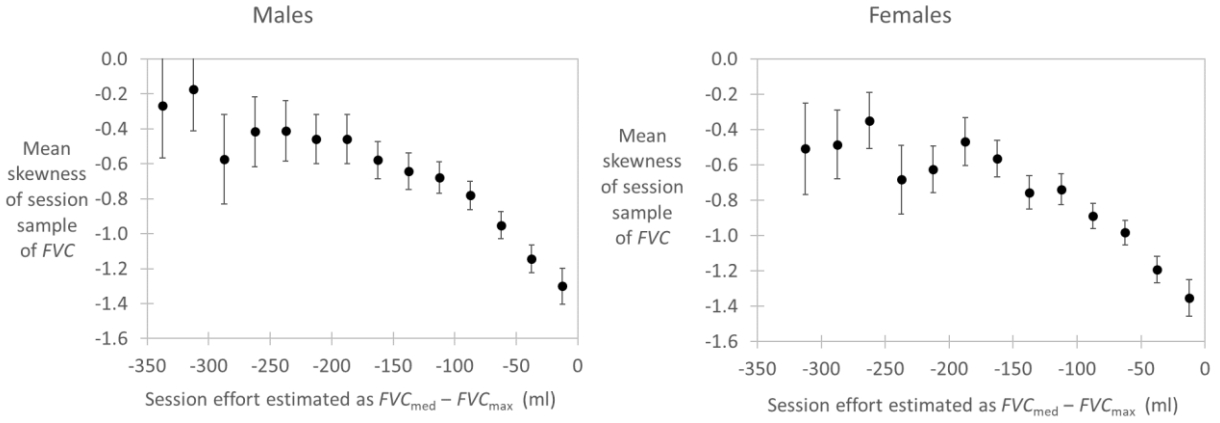

Figure 4. Session skewness of *FVC* as a function of session effort, computed as $FVC_{med} - FVC_{max}$. Each data point corresponds to one specific effort group where that difference, represented on the horizontal axis, falls within an interval of 25 ml (e.g., from 100 to 125 ml). Error bars represent 95% confidence limits.

Figure 4 shows the relation between the mean skewness of session samples of *FVC* and the session effort. On its horizontal axis the figure distinguishes narrow 25-ml bins on the continuum of $FVC_{med} - FVC_{max}$, defining non-overlapping effort groups each including many subjects (on average 239 and 271 subjects for males and females, respectively). The figure eloquently confirms that the skewness of the session samples of *FVC* increases monotonically with the general level of effort during the session.

The results illustrated thus far in Figures 2-4 all rest on the blind computation of session skewness, however small the sample of *FVC* measures. We now turn to an alternative, complementary data-processing approach aimed to *visualize* the W-S distribution of *FVC*.

To begin with, let us consider the overall distribution of the 17,000 measures of *FVC* gathered in all maneuvers performed by all adults of both genders (Figure 5).

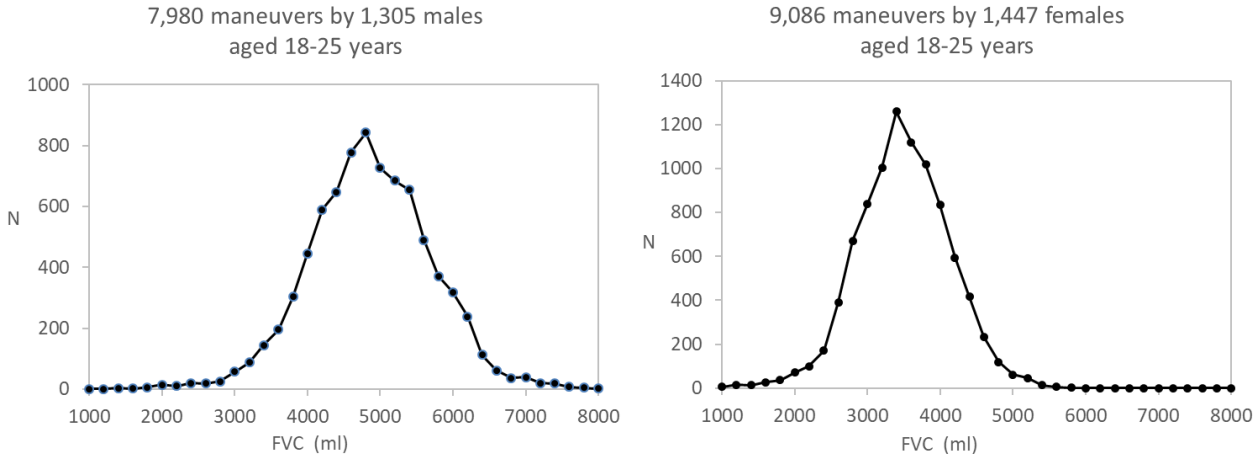

Figure 5. Distribution of raw *FVC* from all maneuvers of all male and female adults.

For both genders the distribution of *FVC* is bell shaped, with some negative skew ($\gamma = -0.103$ in males and -0.485 in females).

Note that if Figure 5 does visualize *FVC* distributions, the offered picture is corrupted by a great deal of B-S variability, all subjects being pooled together. Obviously we want to disentangle the W-S variability of *FVC* from the B-S variability of *TLC*.

Figure 6 isolates the B-S variability by showing the distribution of FVC_{max}, the estimate of *TLC*, over all adults of our data set. If *TLC* is an anthropometric parameter pretty much like body weight or standing height, then the distribution of FVC_{max} across subjects should be Gaussian. This indeed appears to be the case in the data.

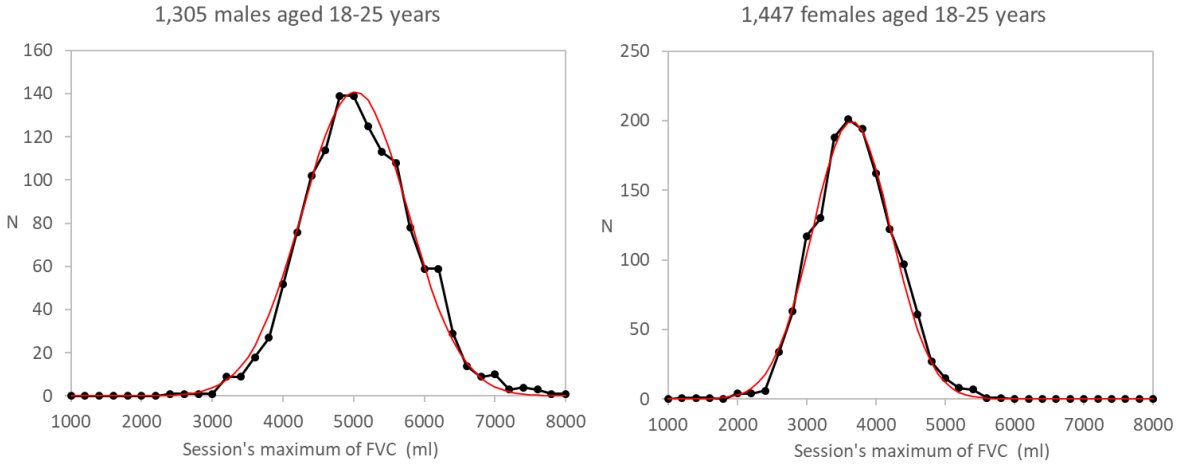

Figure 6. Distribution of FVC_{max}, the estimator of *TLC*, in adults of both genders with fitted Gaussians of parameters μ = 5,050 ml and σ = 740 ml for males, and μ = 3,650 ml and σ = 570 ml for females.

Let us now inquire into the W-S distribution of *FVC*. To rid the *FVC* measures of the B-S variability, the session's maximum was subtracted from all the measures of *FVC* gathered in that session, thus adjusting the origin of the continuum of *FVC* measurement so that every session sample of *FVC* now has its maximum at 0 ml. The result is a recalibrated *FVC* score which measures the distance to the personal capacity limit of the subject who produced the score. The advantage of the recalibration is that the measure can now be pooled from many different subjects with no more interference from B-S variability. Its distribution in adults of both genders is visualized in Figure 7.

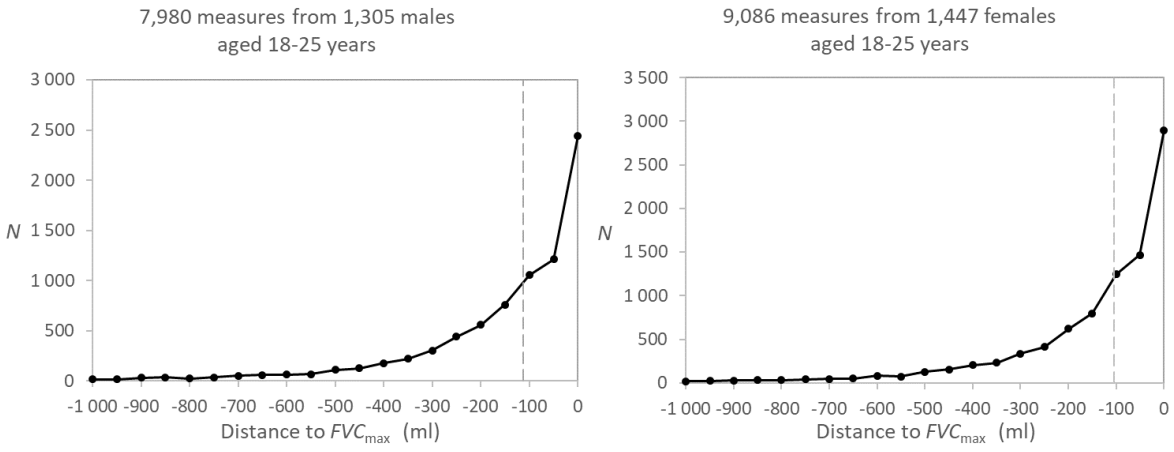

Figure 7. Distribution of recalibrated *FVC* in male and female adults. All measures from all maneuvers of the data set are pooled. The dashed line shows the median recalibrated score.

The picture is markedly different from that shown in Figure 5. We now face a very strong clustering of *FVC* measures against their respective upper bounds.[8] Computed over the 17,000 maneuvers of our adult data set, the skewness coefficient is now–5.02 in males and –4.28 in females.[9]

Figure 7 makes it quite visible that the shape of the W-S distribution of respiratory performances is qualitatively different from that of the B-S distribution of respiratory capacities (Figure 6). While the latter is nearly Gaussian, the W-S distribution of recalibrated *FVC* is monotonically increasing with positive acceleration throughout, reminiscent of an exponential distribution.

However dissimilar their shapes, the W-S distribution of Figure 7 and the B-S distribution of Figure 6 can be compared in terms of their total range of variation. It is interesting to notice that the total range of recalibrated *FVC*, on the order of 1 liter, amounts to about one fourth of the total range of FVC_{max}, from 3 to 7 liters in male adults and from 2 to 6 liters in female adults, as this observation explains the failure of Figure 5 to unveil the true shape of the W-S distribution. Taken in the absolute, the total amount of W-S variability of recalibrated *FVC* is impressively small, the median of Figure 7 hardly exceeding 100 ml (105 ml and 116 ml in male and female adults, respectively), a result compatible with Becklake and Permutt's (1979) report of a range of 90-200 ml for the standard deviation of *FVC* across maneuvers. This result means that about 50% of the *FVC* measures recorded in a typical spirometry session fall at a distance of 100 ml or less from the session maximum.

5.2 The MS Correlation on FVC for Capacity Groups of Different Homogeneities

Figure 8, where each data point corresponds to one individual testee, shows scatter plots of the session standard deviation vs. the session mean of *FVC*. The degree of homogeneity with regard to the estimated respiratory capacity of the group of subjects whose means and standard deviations are plotted is made to increase systematically from panel to panel. Panel A starts with a tolerance interval for FVC_{max} so large (4,000 ± 4,000 ml) as to include all 3,517 males aged 8-25 years of the data set, and so the mean correlation between means and standard deviations (called the MS correlation henceforth) is visualized for a totally heterogeneous capacity group. At the other extreme, Panel F shows the relationship for one relatively small but fairly homogeneous group of 'quasi-clones', composed of 175 individuals with FVC_{max} values in the narrow range of 4,000 ± 125 ml (or 4,000 ml ± 3%).

[8] A demonstration that the monotonic increase of Figure 6 is not an artefact of the recalibration technique is provided in Annex 1.

[9] These intriguingly high values of skewness are discussed in the final section.

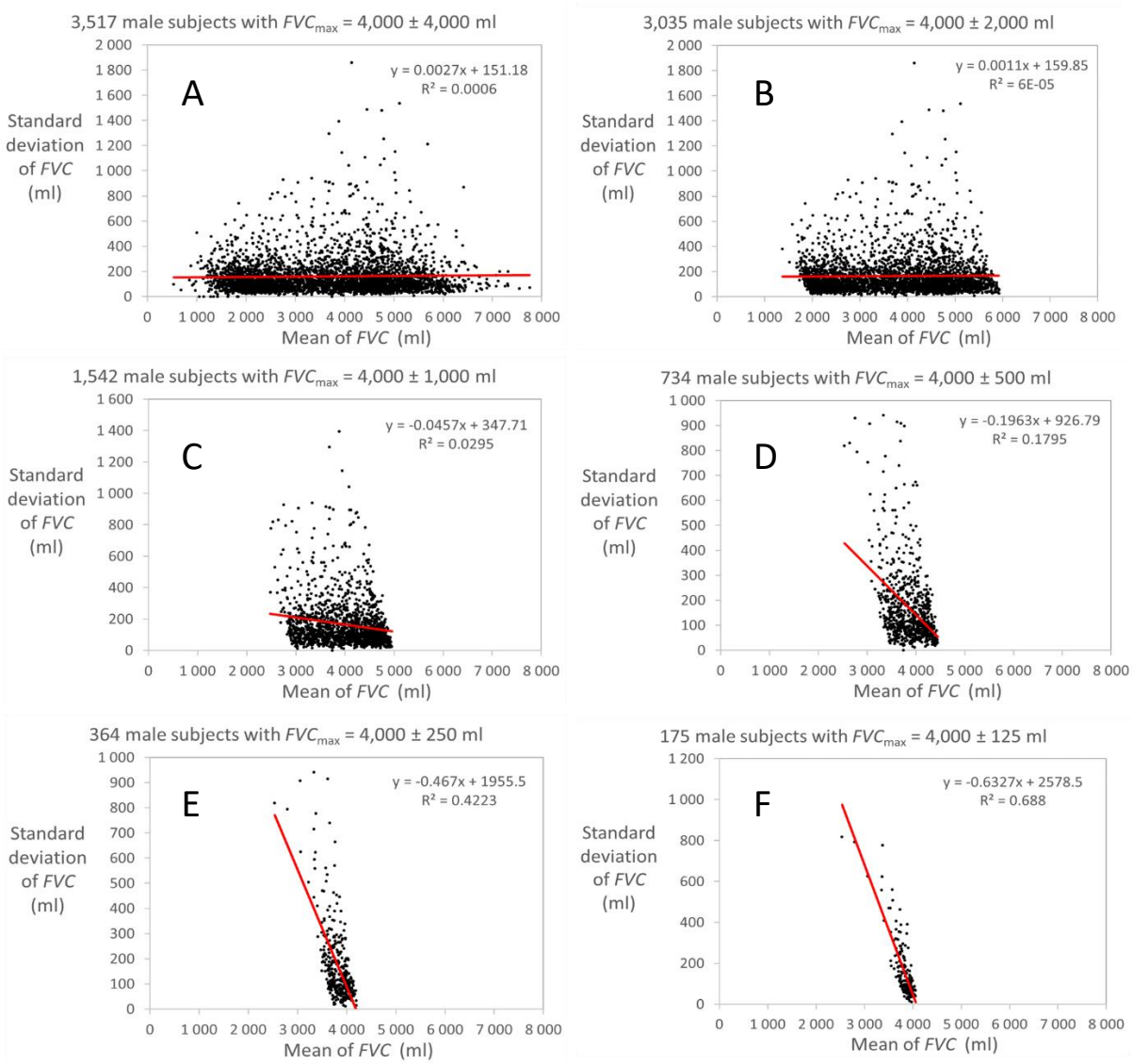

Figure 8. An example of the gradual emergence of a negative MS correlation on *FVC* as the B-S heterogeneity of respiratory capacities is gradually reduced from FVC_{max} = 4,000 ± 4,000 ml to FVC_{max} = 4,000 ± 125 ml.

As the tolerance interval for FVC_{max} is halved again and again, thus reducing the amount of B-S noise, the expected negative MS correlation gradually emerges. While at first only statistical noise is visible (r = .02 in Panel A), the pattern takes shape progressively, ending up with an r of –.83 for the group of quasi-clones of Panel F.

The result shown in Figure 8 is just one example. The method was repeated over the whole continuum of respiratory capacities, yielding the large sets of results visualized in Figure 9. The x coordinate of each data point gives the central value of FVC_{max} for the capacity group whose degree of homogeneity is specified in parameter, and the y coordinate gives the corresponding value of r.

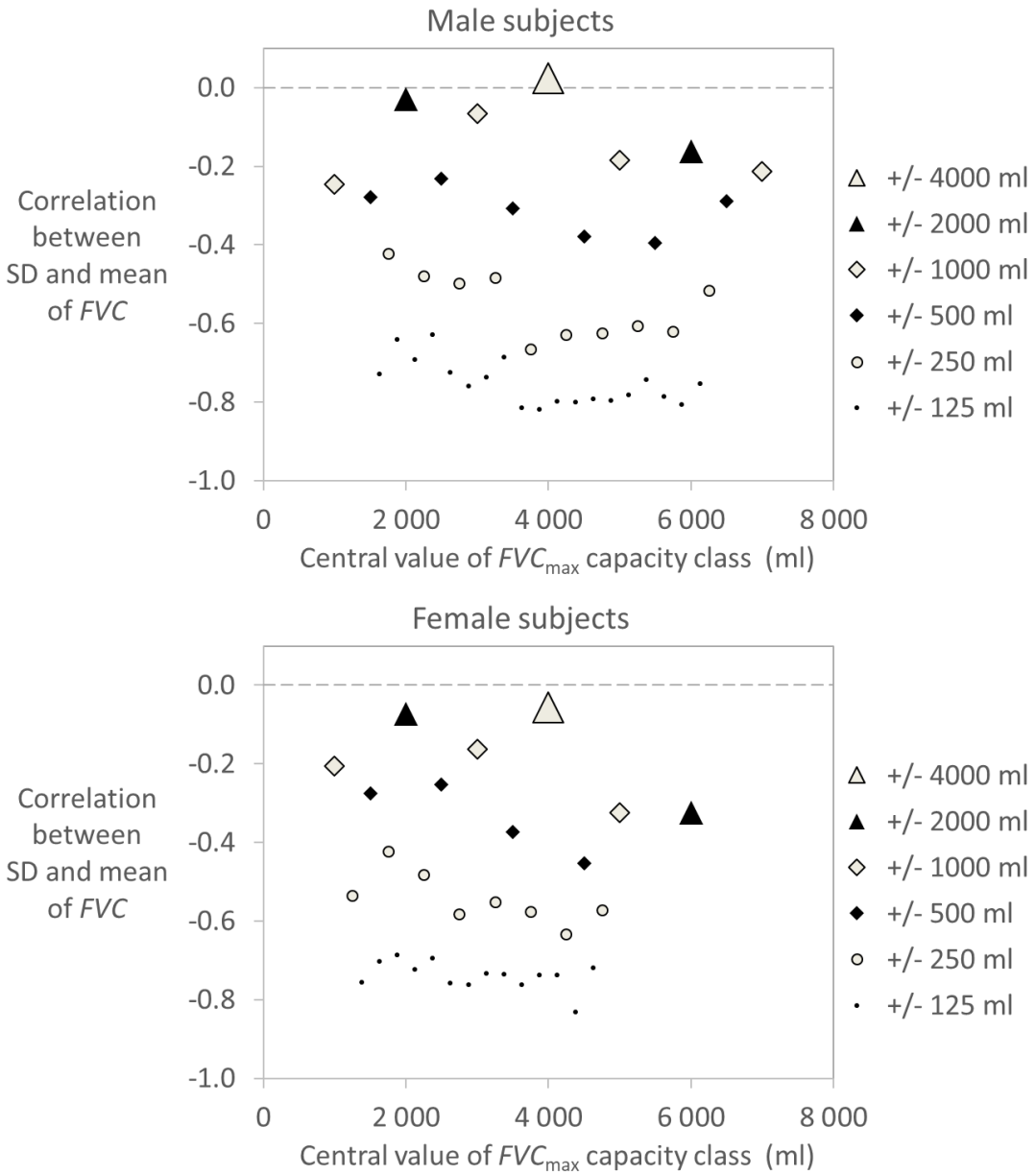

Figure 9. The MS correlation on *FVC* in groups of systematically varied homogeneities with regard to the criterion of FVC_{max}. Capacity groups including fewer than 50 subjects were left aside.

The data from males and females show the same, highly consistent pattern, confirming beyond doubt that the more homogeneous, capacity wise, a group of subjects, the more strongly negative the MS correlation. Notice that Figure 9 reports the result for 33 independent groups with individual capacities in the range ± 125 ml, each composed of about 200 subjects (on average 179 for males, 264 for females). Since a sample of 200 sessions run by 200 spirometry clones is pretty much equivalent, statistically, to a sample of 200 consecutive sessions run by the same subject (while of course free of the complications of serial measurement), it is most instructive to see that for these highly-homogeneous groups all our estimates of the MS correlation on *FVC* fall in the range from –.6 to –.8. Such correlation strengths are impressive bearing in mind that there still remained a certain amount of B-S variability among quasi-clones with FVC_{max} values within an interval of ± 125 ml.

The tolerance interval used to constitute the groups of quasi-clones of Figure 9 being an absolute value rather than a percentage, obviously the higher the level of FVC_{max}, the more homogenous the group: the same interval of ± 125 ml represents ± 9% at the extreme left of the figure, but only ± 2% at the extreme right. This provides us with a possibility to check the internal consistency of the data. Notice that in Figure 9 the strength of the negative correlation tends to increase from left to right, and that this trend is replicated at every level of group homogeneity. One may speculate that the correlation would have been still stronger with perfect clones, or with many sessions with the same testee.

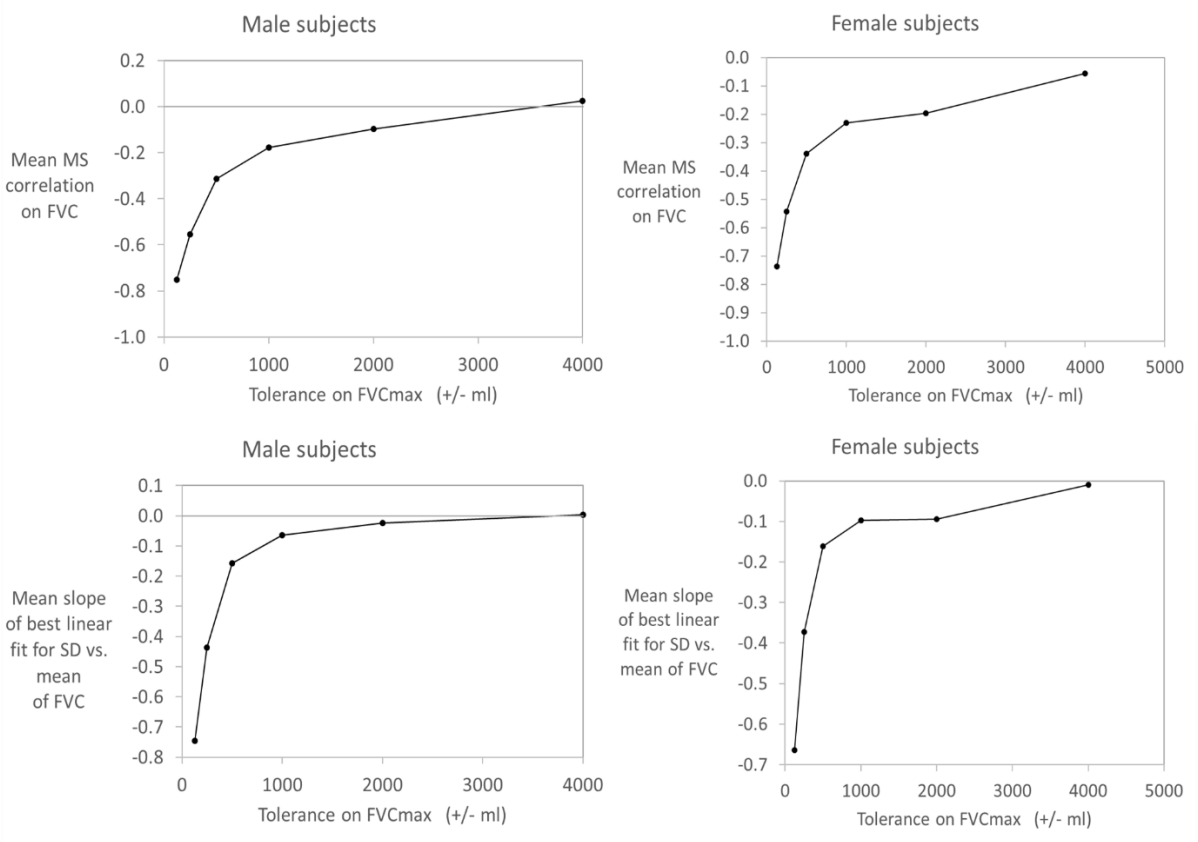

Figure 10. Strength of the MS correlation and slope of the best linear fit as functions of the homogeneity of the capacity group.

Figure 10 summarizes the results by showing how not just the strength of the MS correlation, quantified by the *r* statistic, but also the steepness of the fitted linear relationship increase with more and more homogeneous capacity groups.

Thus the NHANES III data contain strong converging evidence that there exists a strong negative correlation, across sessions, between the first and second moments of the W-S distribution of *FVC*, whose detection demands that the considerable amount of B-S variability be eliminated.

6. Conclusions and Perspectives

The above reported statistical results seem quite worthy of consideration given both their high degree of internal consistency and the size and technical quality of the NHANES III data. We may conclude with a fairly high degree of confidence that (1) session samples of *FVC* measures indeed exhibit a strong amount of negative skew, closely dependent upon the estimated strength of the testees' efforts, and that (2) there is indeed a strong negative correlation between the first and the second moments of the distribution of session samples of *FVC*. These empirical findings may be of interest to medical researchers specializing in the optimization of standards and in the statistics of spirometry. Another notable result of this research is the empirical demonstration of the workability of what was called above the 'cloning' method, to this writer's knowledge a novel method, which might be useful to statisticians of spirometry.

In this section we will zoom out to see some interesting bridges linking spirometry testing to other fields of scientific inquiry and to discuss some general implications of this work.

6.1. From Respiratory Performance to Human Quantitative Performance in General

The present study, focused on one specific measure of spirometry, is part of a wider research project aimed at developing a general understanding of quantitative human performance, where data from several different fields including experimental psychology, athletics, and gaming are analyzed. The project arose from the realization that an impressively large equivalence class is captured by a pretty strict definition of quantitative performance:

Definition. A performance score is a measure subjected to a deliberate minimization or maximization effort exerted by a human agent against the resistance of a limit, a lower- or upper bound, respectively (Guiard, 2020).

Spirometry testing unambiguously falls in that equivalence class: to explicitly ask testees to exhale maximized volumes and flows of air is to ask them for respiratory performances. In fact countless instances of performance measurement can be found in every field of science and engineering and every sector of social life and so spirometry testing is just an instance amongst many. The particular case of spirometry, however, is of very special interest as in that case the measurement of performance and the estimation of the capacity of performance happens to take on exceptionally simple guises.

The conjunction of four conspicuous features of spirometry—in particular volumetric spirometry—makes this measurement situation uniquely suitable to the study of human performance.

One is that the testee's effort is strictly one-dimensional. There is no conflict between the requirement to jointly maximize a volume and a flow of air, meaning that the subjects can—if they will—invest upon each maneuver the totality of their available effort resource. Counter-examples are countless. Thus in many psychology experiments participants are asked to minimize a time measure and an error measure concurrently, being thus confronted with a conflict that forces them to share their effort resource in various proportions between the speed and the accuracy fronts (Norman & Bobrow, 1975), a situation which complicates to a serious extent the analysis of the interplay of effort and capacity (Guiard, 2020, Section 9).

Second, the upper limit of performance is identified physically, the capacity of performance *TLC* amounting literally to the inner volume of a container. In contrast, in most performance measurement situations the performance capacity concept is just a metaphor, the real nature of the personal limit remaining elusive.

Third, the upper limit *TLC* which constrains the respiratory performance *FVC* is fixed at the time scale of a test session, and strictly impassable. One counter-example amongst many is the response time of psychology experiments, a measure always subjected to a minimization effort. Response time can take arbitrarily low values (and even occasionally turn negative) because it is lower bounded by just the inflation of inaccuracy (Luce, 1986; Pachella, 1973; Wickelgren, 1977). Here and in many other cases the capacity limit is soft and negotiable, quite unlike that constraining the measures of spirometry.

Fourth, spirometry testing is, from the point of view of statistical sampling, exceptionally simple, each individual testee being asked to produce in just one session just one sample of *FVC* measures. Thus the problem of the W-S distribution of *FVC* involves a single level of statistical aggregation.

For these reasons it is easier in spirometry than anywhere else to identify the basic shape of within-individual distributions of performance scores and to investigate the causal relationship linking that shape to the interplay of a randomly variable extremization effort and a fixed capacity limit. In other words, spirometry testing appears to qualify as an enlightening paradigm for the general study of quantitative human performance. The above-reported theoretical and empirical results about the simple case of spirometry have a potential to contribute to our general understanding of the mechanisms at work in performance testing situations, in spirometry and beyond.

The model of spirometry has been helpful to this writer in his reexamination, currently in progress, of speeded aimed movement, focused on the parallel distributions of movement time and error—both performance scores subjected to a minimization effort.[10] Reanalyzing several data sets in light of the present conceptual framework, and considering the speed and the accuracy dimensions in parallel, he was able to show that the above results hold for both the time and the error score whenever the experimental conditions allow the participants to allocate enough effort resource to one minimization effort at the expense of the other.

6.2 Skewed Distribution of FVC: What Do We Mean?

To characterized the shapes of our distributions above we used the conventional notion of skewness and the received formula for its calculation designed so that negative skew obtains when the left tail of a distribution is elongated relative to the right. However, there is something awkward to the statement that a shape like that of Figure 7, where the skewness statistic reaches the rather unusual value of –5, is ‘skewed’. The problem is that the frequency curve in question increases monotonically throughout and thus exhibits no tail whatsoever on its right-hand side.

[10] Guiard, Y. (in preparation). *Monotonic distributions of movement time and error in speeded aimed movement tasks.*

An analogy may help clarify the concern. One might sensibly devise some skewness index to quantify the difference, often substantial (Govind, 1989), between the left and right claws of an American lobster. That index would capture a local violation of symmetry in the context of a globally symmetrical morphology. However it would make little sense to try to quantify the degree of skewness between the lobster's front and tail because the idea of a deviation from symmetry is irrelevant, for lack of any detectable symmetry along the rostro-caudal axis. Our within-session distribution of recalibrated *FVC* measures raises a similar problem. As explained in Section 2, this distribution must have a rostro-caudal organization: if it may well show an evanescent tail on its left-hand side because on this side the measures are free, whenever the effort falters, to extend to arbitrarily low values, it must have an abrupt front on its right-hand side where lies an attractive upper limit (Guiard et al., 2011; Guiard & Rioul, 2015). And, as we have seen, that common sense argument was corroborated by the quasi-exponential shape found with a properly adjusted measure of respiratory performance.

Then, can we say that we have found a strongly skewed shape, meaning a strong departure from the familiar bell shape of statistical handbooks, if in the situation of interest the bell shape was implausible in the first place? It seems more reasonable to accept the view that the convex, increasing curve of Figure 7, verified by this writer on a variety of data sets (Guiard, 2020), describes the typical shape of a performance distribution and that that shape has a rostro-caudal morphology. But at this point some far-reaching and somewhat unsettling statistical issues arise.

The standardized practice of spirometry consisting of summarizing each sample of *FVC* scores by its maximum has been working apparently to the satisfaction of generations of practitioners, and we have seen that it is easy to justify rationally. Notice, however, that this practice is hard to reconcile with the usual recommendations from statistics textbooks. In the face of a measure that varies unpredictably across measurements, one is supposed to summarize one's empirical sample with three sample statistics. At the very least one should summarize the *location* of the sample of measures on the measurement continuum by means of some central-trend indicator like an arithmetic mean or a median. It is recommended to also measure the sample's spread or *scale* with a standard deviation or an inter-quartile interval, and its *skewness* with some parametric or non parametric index. Such recommendations rest on the fundamental assumption of conventional statistics that if the measured value is not strictly deterministic, it certainly is a random variable whose 'true' or expected value should be situated somewhere in the bulk of a bell shape.

The fact is, spirometry seriously departs from that schema. To begin with, practitioners of spirometry do not care at all about the central trend of their samples, but why should they? There is reason, both theoretical and empirical, to suspect that such a trend does not exist in a performance distribution pressurized by a strong upward or downward effort; rather than an average, they take their sample maxima to serve as their location summary. Notice that by definition an extremum cannot be representative of a sample of data—yet their option seems quite sound as FVC_{max} is the best possible estimator of *TLC*. Second, the way practitioners handle the spread or scale issue, known in the spirometry literature as the problem of the repeatability of measures, is again original. Standards of spirometry recommend to measure the distance from the best to the second best measure of respiratory performance (Graham et al. 2019, Table 7), and it is a variant of that option that was actually used in Figure 4, where the spread was measured by the difference $FVC_{med} - FVC_{max}$. Obviously the established

practice of spirometry takes us away from the conventional view that the location and spread of a distribution are best quantified by its first raw moment, or mean, and its second central moment, or variance.

Thus there is tension between the standardized practice of spirometry testing and the common understanding of statistical theory, and this tension is a source of intellectual discomfort betrayed in the spirometry literature by some conspicuous symptoms. For example, it is interesting to notice that articles offering reference values for spirometry typically omit to recall that their basic data are individual maxima, rather than averages: this is the case notably in Hankinson et al. (1999), Stanojevic et al. (2008), Quanjer et al. (2012), Rochat et al. (2013), and Coates et al. (2016). Another illustration is this curious quote from Bland and Altman (1996), two renowned specialists 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" (p. 1654). In this quote from a short note aimed to communicate some rudiments of statistical theory to a readership of non-specialists, the choice of spirometry as an illustration example was rather unfortunate as their statements, as far as spirometry is concerned, are just false.[11]

In fact the original statistical practice of spirometry discreetly conceals a profound challenge for statistics and probability theory. When it comes to the measures of spirometry—and more generally to the measures which, being deliberately extremized by a human agent, fall in the special class of performance scores—the classic concept of a random variable loses much of its relevance. The well-known law of errors, which says that the probability declines as the measure deviates more and more, whether upward or downward, from the expected value, thus yielding the familiar bell shape, does not seem to apply well in these contexts of measurement. Performance measures, anchored at a more or less solid extremum rather than centered about a probabilistic expectation, look quasi-deterministic in essence, as recently noted by this writer, who proposed explicit distributional criteria to distinguish them from the familiar random variables of probability theory (Guiard, 2020).

6.3 MS Correlation in Performance Measurement: A General Account in Perspective

The correlation we found between the mean and the standard deviation of *FVC* across samples gathered in homogeneous groups of quasi-clones is reminiscent of that known to characterize within-subject distributions of *response time* (*RT*) in psychology experiments. That correlation has been estimated by Wagenmakers and Brown (2007) in ten independent data sets from experiments with a broad diversity of memory, perception, categorization, and problem-solving tasks. These authors found that nearly three quarters of all participants had a correlation of at least .85, and they offered a convincing demonstration that a strong positive

[11] That measurers want to know the expected value of their measures is true in general, but not in spirometry. The same reservation holds for the assertion that the observed W-S variability of the measures takes its source in measurement error—in spirometry, measurement error (2-3 ml, see Hankinson et al., 1999) is a very minor concern, being more than an order of magnitude smaller than the fluctuations of the testee's effort.

MS correlation is indeed a robust and general law of *RT*, just like the old law of practice (Heathcote et al., 2000).

In our *FVC* data the strength of the correlation is of similar magnitude, but the sign is *opposite*. From the moment it is realized that *RT* is an instance of a forcefully *minimized* measure and *FVC* an instance of a forcefully *maximized* measure, such a symmetry opens an intriguing perspective. The possibility arises of Wagenmakers and Brown's law of *RT* and the above-reported patterns of spirometry data being two special instances of a more general law of human quantitative performance: ask performers to orient their effort downward, as in time minimization tasks, the mean and the standard deviation of performance will correlate positively; ask them to orient their effort upward, as in the maximization task of spirometry, the two statistics will correlate negatively. Thus, facing a task demanding the extremization of some quantitative score, not only can we safely predict that the first and second moments of distribution will correlate across within-individual samples, we can tell the sign of that correlation by just considering the direction of the required effort.

Wagenmakers and Brown (2007) showed that their positive correlation on *RT* can be satisfactorily explained by classic models of mathematical cognitive psychology.[12] It does not seem too risky to say that the sophisticated mechanisms of these models would be hard to transpose from the context of minimized *RT* to that of maximized *FVC*. In contrast, the simple idea of a ceiling effect proposed above to explain the negative correlation observed on maximized *FVC* is readily transposable to the case of a positive correlation on minimized *RT*, the ceiling effect having just to take the symmetrical form of a *floor* effect. Any variation in the performer's minimization effort (or, equivalently, any reduction of task difficulty) will tend to move the mean and the standard deviation of minimized *RT* in the same direction. For example, with a stronger effort mean *RT* will move down, but so will the standard deviation since the measures will tend to accumulate more compactly just above the lower limit below which the probability of error is likely to explode (Pachella, 1973; Wickelgren, 1977). In this account the correlation takes opposite signs simply because the capacity limit is located on opposite sides of the distribution.

This reversible ceiling/floor effect explaining simultaneously, in parsimonious terms, two empirical findings so far believed to be unrelated, the opening perspective seems worthy of a careful exploration. Preliminary evidence gathered by this writer using a variety of performance data from different fields including experimental psychology (Guiard, 2020), athletics, and gaming does suggest that the account holds in general.

References

Becklake, M. R., & Permutt, S. (1979). Evaluation of tests of lung function for "screening" for early detection of chronic obstructive lung disease. *The lung in the transition between health and disease*, M. Dekker, Editor, 345-387.

Bland, J. M., & Altman, D. G. (1996). Measurement error. *British medical journal*, 312(7047), 1654.

Brusasco, E. V., Crapo, R., Viegi, G., Wanger, J., Clausen, J. L., Coates, A., ... & Pellegrino, R. (2005). Series ''ATS/ERS task force: standardisation of lung function testing''.

[12] Ratcliff's (1978) diffusion model and Logan's (1988) instance theory of automatization.

Coates, A. L., Wong, S. L., Tremblay, C., & Hankinson, J. L. (2016). Reference equations for spirometry in the Canadian population. *Annals of the American Thoracic Society*, 13(6), 833-841.

Govind, C. K. (1989). Asymmetry in lobster claws. *American Scientist*, 77(5), 468-474.

Graham, B. L., Steenbruggen, I., Miller, M. R., Barjaktarevic, I. Z., Cooper, B. G., Hall, G. L., ... & Thompson, B. R. (2019). Standardization of spirometry 2019 update. An official American thoracic society and European respiratory society technical statement. *American journal of respiratory and critical care medicine*, 200(8), e70-e88.

Guiard, Y. (2020). On performance measurement in psychology and other fields: https://hal.archives-ouvertes.fr/hal-02943143/document.

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.

Hankinson, J. L., Odencrantz, J. R., & Fedan, K. B. (1999). Spirometric reference values from a sample of the general US population. *American journal of respiratory and critical care medicine*, 159(1), 179-187.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, 7(2), 185-207.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492–527.

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.

Madsen, F. (2012). Validation of spirometer calibration syringes. *Scandinavian journal of clinical and laboratory investigation*, 72(8), 608-613.

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

NHANES (2011). *Respiratory Health: Spirometry Procedure Manual*. https://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/spirometry_procedures_manual.pdf

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.

Quanjer, P. H., Tammeling, G. J., Cotes, J. E., Pedersen, O. F., Peslin, R., & Yernault, J. C. (1993). Lung volumes and forced ventilatory flows. *European Respiratory Journal*, 6(Suppl 16), 5-40.

Quanjer, P. H., Stanojevic, S., Stocks, J., & Cole, T. J. (2012). GLI-2012: All-age multi-ethnic reference values for spirometry. *Global Lung Initiative*.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, 9, 278 –291.

Rochat, M. K., Laubender, R. P., Kuster, D., Braendli, O., Moeller, A., Mansmann, U., ... & Wildhaber, J. (2013). Spirometry reference equations for central European populations from school age to old age. *PloS one*, 8(1), e52619.

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

Stanojevic, S., Wade, A., Stocks, J., Hankinson, J., Coates, A. L., Pan, H., ... & Cole, T. J. (2008). Reference ranges for spirometry across all ages: a new approach. *American journal of respiratory and critical care medicine*, 177(3), 253-260.

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, 114(3), 830.
Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67-85.

Appendix 1

It was shown in Figure 7 that the distribution of recalibrated *FVC*, the measure of the distance from the *FVC* score to its session maximum FVC_{max}, takes the shape of a monotonically increasing and positively accelerated curve. To control that this result is not just an artefact of our alignment method, all samples of *FVC* measures were subsequently aligned by their respective session's *minima*, obviously not constrained by any lower bound. The result is shown in the lower panel of Figure A1. Notice that bin size (50 ml) as well as the ranges shown on the vertical and horizontal axes are the same in all four panels.

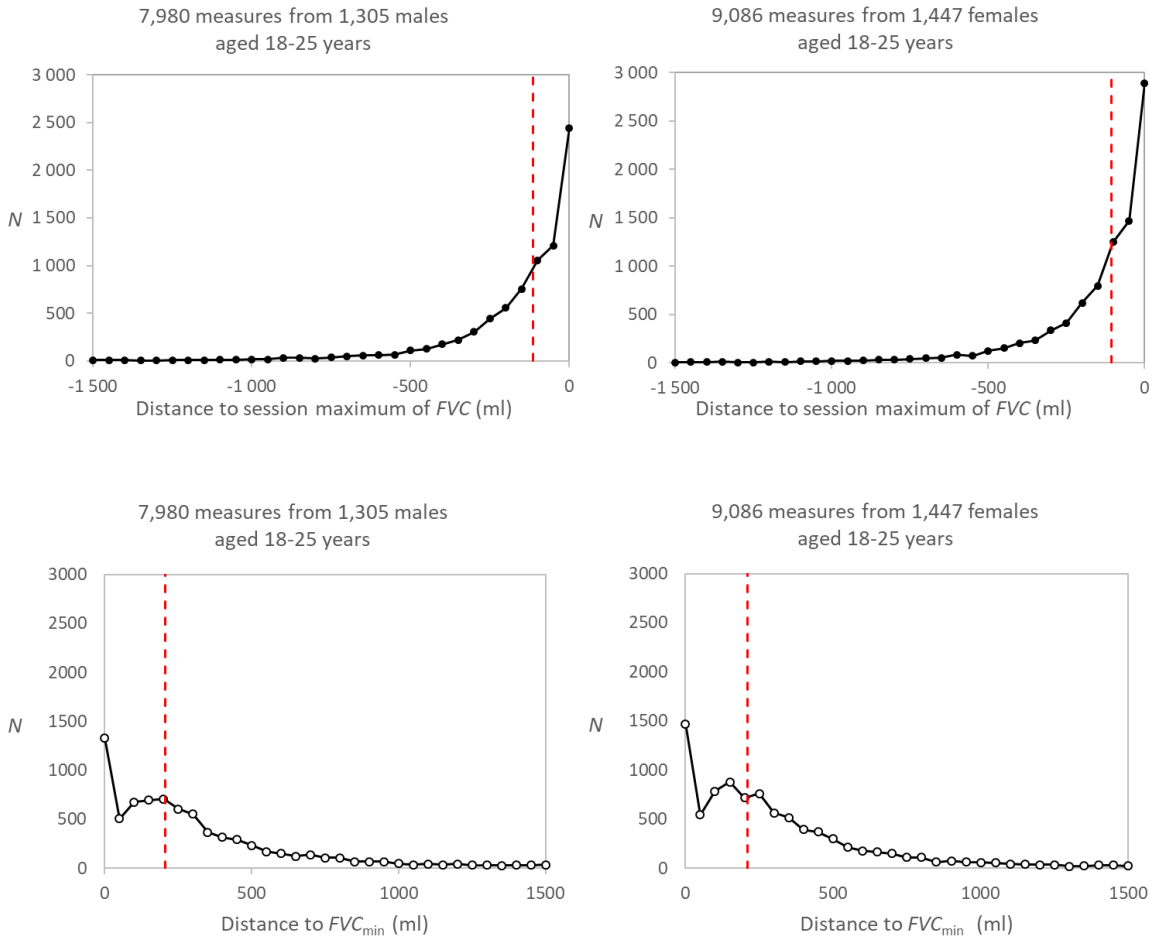

Figure A1. Distribution of the distance separating *FVC* scores from their respective session's maxima (above) and minima (below) for all maneuvers performed by all male and female adults of the data set. Dashed lines represent medians.

The median of *FVC* is twice as far from the session minimum (205.5 ml in males, 210 ml in females) as it is from the session maximum (-113 and -105 ml). The total range of variation of $FVC - FVC_{min}$ is nearly twice as large as that of $FVC - FVC_{max}$. While recalibrating the *FVC* scores by their respective session minima, rather than maxima, results by construction in a positively, rather than negatively skewed distributions, the key difference is that we obtain distinctly two-tailed distributions, with a conspicuous mode at about 200 ml above zero.