SJS The Self Journal of Science

Research article

Assessing Mood Change With Visual Analogue Scales: Composite Versus Vectorial Approaches Version 1

Stéphane Vautier [1] and Mohammad Hassan Afzali [1]

1. Unité de Recherche Interdisciplinaire Octogone (EA4156) - Université Toulouse Jean Jaurès

Made public on Oct, 19th 2015 under Creative Commons 4.0 Attribution License
Reviewed and discussed at http://www.sjscience.org/article?id=425

Abstract We examined the assessment of mood change using multiple visual analogue scales from the perspectives of computing a composite difference score and verifying the vectorial comparability of test and retest ratings. The composite approach raises the question of whether the true score can measure mood and whether valid conclusions can be derived from true-score differences within individual data. The vectorial approach allows clinicians to use the data to test the causal and functional conception of ordinal measurability of patients' mood. This falsificationist approach may lead clinicians to recognize that, in some cases, ratings are not measurements and should be treated as speech acts in a conversational setting.

A psychologist wants to monitor a patient's mood by collecting her/his self-ratings on a series of visual analogue scales (VASs ; [4, 25]). But what should be done with these self-ratings? Everything depends on how they are thought to measure patients' mood. Classical Test Theory (CTT) recommends computing a composite score, based on the rationale that summing the observed scores asymptotically reduces measurement error (if there were an infinite series of experimentally independent scores, their mean would be the true score). We begin by describing the conceptual and epistemological issues raised by what we call the *composite* approach. We then describe an alternative approach that consists in viewing vectors of self-ratings as outputs of a multivariate measurement function – the *vectorial* approach. We conclude that psychologists should be ready to use self-ratings not as measurements but as speech acts, when faced with falsifying evidence.

Let us start by taking a concrete example of a set of six VASs designed to operationalize *tense arousal*. Patients were instructed to rate their mood on a 102-mm straight line where the end anchors were the extremes of a given feeling (VAS 1: tense; VAS 2: strained; VAS 3: stressed; VAS 4: unstrained; VAS 5: mellow; VAS 6: relaxed; [19]). Respondent 1 rated his mood twice, marking the lines to indicate the intensity of his internal stimuli (observed scores plotted in Fig. 1).

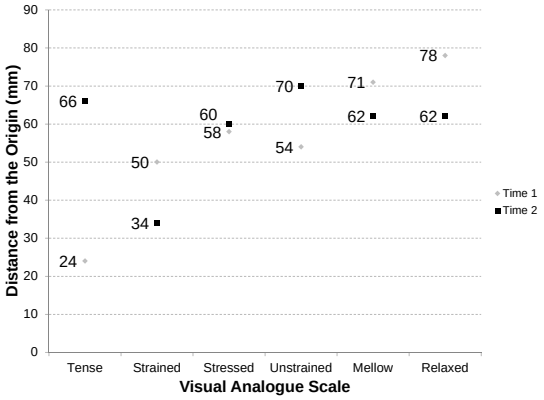

Figure 1. Self-ratings of Respondent 1.

THE COMPOSITE APPROACH

A curious structural feature of previous tense arousal test-retest data [19] is *imperfect dynamic bipolarity* (for previous developments of this concept, see [22, 23]): the modelling of the difference scores revealed that the correlation between the true-score differences measured by the *negative* VASs (i.e., tense, strained, and stressed) and the reversed true-score differences measured by the *positive* VASs (i.e., unstrained, mellow, and relaxed; see [19], Fig. 1, right panel) was .82 (.03). Consequently, the negative and positive VASs seemingly measured intraindividual variations of distinct kinds, suggesting either that tense arousal is not a single dimension of mood, or, if we argue that tense arousal is a single dimension of mood, that the positive VASs lacked construct validity with respect to the negative VASs. In either event, psychologists would be well advised to compute two difference scores, one from the negative VASs:

$$[(66 + 34 + 60) - (24 + 50 + 58)]/3 \simeq 9.3 \text{ mm}, \quad (1)$$

and one from the positive VASs:

$$[(70 + 62 + 62) - (54 + 71 + 78)]/3 = -3 \text{ mm}. \quad (2)$$

But what conclusion could be drawn from these difference scores? Previous structural equation modelling (SEM) analyses [19] indicated that the composite difference variables were quite reliable (Cronbach's alpha: .95 and .94, respectively). However, as the error components were unknown, all psychologists could do in this case would be to use the observed difference scores as better estimates of the respective true-score differences. Thus, they could say that *it seems* that tense arousal is slightly increased – +9.3 mm –, but they would still have to decide how to interpret the second difference score – −3 mm. Then again, were they to consider the maximum composite variation (less than 10 mm) with respect to the VASs' total range (102 mm), they could argue that as they had no evidence that Respondent 1's mood changed markedly, his mood change must have been negligible.

A Conceptual Issue: Are True Scores Measurements of Tense Arousal?

Unfortunately, the psychometric rhetoric above is akin to obscurantism (for a broad criticism of academic obscurantism, see [2, 13, 14]; in the field of psychometrics, see [8–12]). The true score does not refer to an amount of tense arousal that the observed score measures more or less reliably (hoping that the measurement error is *small*). According to [6], the true score is the expectation of a *counterfactual* random variable that takes its values from an empirical set of numerical values. In the VAS context, the test score is the distance between the mark on the straight line and its origin. Thus, the true score is the concept of an expected *distance* measured in millimetres. CTT does not assert that this expected distance measures an amount of tense arousal because it is a *statistical* theory that can be applied to marks on VASs.

Tense arousal cannot be observed, and the expected distance is also impossible to observe, so psychologists interested in the subject can easily use the term *true* score

to refer to an amount of tense arousal. It is not, however, valid to deduce that two things sharing the same feature – non observability – are one and the same thing.

Conceptual confusion is also maintained by the pervasive use of the term *measurement error* in the specialist literature. In CTT, this term denotes the difference between the observed score and the expected score, while in assessment literature, it suggests that the observed score results from a *measurement* process, that is, from a process that relates an amount of something (that cannot be observed) to an observable event like a mark on a VAS, such that the observed event is a causal effect of that amount. After all, if no measurement process occurs, why name the difference between the observed score and the true score a *measurement* error? Although psychologists understandably wish to measure tense arousal rather than mere distances on straight lines, CTT tells us nothing about how observed or expected marks on VASs actually measure tense arousal.

An Epistemological Issue: Invalid Conclusions From Observations

Structural properties can be tested statistically through SEM. It can be looked at whether the three VASs of the same sign (i.e., all negative or all positive) measured the same expected score differences [19]. This property is not derived from the tautological CTT decomposition.

For example, coming back to Respondent 1's scores, this hypothesis states that the true-score differences associated with the observed differences $66-24 = 42$ mm (tense), $34-50 = -16$ mm (strained), and $60-58 = 2$ mm (stressed) are one and the same expected value δ_n. Accordingly, it assumes that $9.3 = \delta_n + \varepsilon_n$, where ε_n is the mean of the measurement errors.

However, neither true-score theory nor SEM analyses based on relevant latent variables restricts the range of measurement error (from -102 mm to $+102$ mm). Thus, we cannot deduce from the premise of CTT decomposition, the nice feature of essential tau-equivalence of same-sign difference score variables, or the data at hand, that the true-score difference δ_n is strictly positive, as we cannot exclude the possibility that $\varepsilon_n < -10$ mm, for example. The argument of logical invalidity due to unrestricted measurement error is detailed in [21] and [24]. SEM testing of the structural properties of statistical moments (i.e., means, variances, and covariances) tells us nothing about individual measurement errors. Consequently, even if psychologists realize that true scores are not measurements of tense arousal, and focus their reasoning on true distances and the differences between these true distances, they cannot exclude the possibility that 9.3 is compatible with a negative associated true-distance difference. In practice, they ignore the sign of the associated true-distance difference.

THE VECTORIAL APPROACH

Turning to the proposed alternative, we adopt a falsificationist methodology in order to learn hypothetico-deductively from respondents' longitudinal data, and we rely on a simple definition of scientific measurement. First of all, we reject the definition of measurement as mere rule-based numerical encoding, which obfuscates the concepts of natural law and social norm: "The making of a decision, the adoption of a norm or of a standard, is a fact. But the norm or the standard which has been adopted, is not a fact. That most people agree with the norm 'Thou shalt not steal' is a sociological fact. But the norm 'Thou shalt not steal' is not a fact, and can never be inferred from sentences describing facts." ([16], p. 61) Measurability is not a matter of convention but a matter of fact. Furthermore, if, like Michell and others (e.g., [7, 15]), we prefer to restrict the definition of scientific measurement to metrical measurement, that is, to the idea that measurement as a process results in a real-valued proportion of a measurement unit (or, more accurately, in an interval surrounding that proportion), then clearly nobody has yet discovered a measurement process that would enable us to find a ratio of a tense arousal unit. In short, we cannot scientifically assert that marks on VASs measure tense arousal.

There is room, nonetheless, for a scientific view of ordinal measurement – not to be confused with Stevens' 1946 notion of an ordinal scale [18] –, based on the idea that a quantity can be measured ordinally by some observable events if – and only if – there is a monotone function that is defined in the domain of the quantity and takes its values from the codomain of these observable events (a necessary but not sufficient condition for metrical measurement). [17] distinguishes between thermoscopes (ordinal measurement) and thermometers (metrical measurement; see also [1]), and, as we will see, the existence of a multivariate function can be an empirical issue: some auxiliary hypotheses enable us to test the notion that marks on two VASs are ordinal measurements of the same quantity (for a similar approach to item responses expressed in a discrete format, see [5, 20]). The first subsection below contains the necessary definitions and notations. The second subsection sets out the overall hypothesis that marks on a series of VASs designed to measure the same amount of tense arousal are ordinal measurements. The third subsection discusses the testing of this hypothesis.

The Measurement Function of Respondents' Ratings on a Given VAS

Existence of an amount of tense arousal at any time in a respondent's life. First of all, for a mark on a VAS to be a measurement of tense arousal, it must be the measurement of an amount of something, that is, we must assume that an

amount of tense arousal does exist. This hypothesis can be stated formally as the hypothesis of a function Q that links any time point in the respondent's life to one and only one amount of tense arousal, which can vary from zero tense arousal to a maximum amount of tense arousal.

$$\begin{aligned} Q : \Omega &\to [0,\ \max] \\ t &\mapsto Q(t) = q \end{aligned} \tag{3}$$

We assume that [0, max] is a continuous segment of amounts of tense arousal.

Ordinal measurability. In order to assume that an amount of tense arousal in a given individual is ordinally measurable by marks made by that person on a VAS, say VAS i, we must assume the existence of the measurement function

$$\begin{aligned} F_i : [0,\ \max] &\to [0\text{ mm},\ 102\text{ mm}] \\ q &\mapsto F_i(q) = y_i. \end{aligned} \tag{4}$$

Let us assume that a *negative* VAS (i.e., $i = 1,\ 2,\ 3,\ F_i$) broadly increases, that is:

$$\forall (q,\ q') \in [0,\ \max]^2,\ q > q' \Rightarrow F_i(q) \geq F_i(q'). \tag{5}$$

Let us also assume that a *positive* VAS (i.e., $i = 4,\ 5,\ 6,\ F_i$) broadly decreases, that is:

$$\forall (q,\ q') \in [0,\ \max]^2,\ q > q' \Rightarrow F_i(q) \leq F_i(q'). \tag{6}$$

First auxiliary hypothesis due to temporal approximation. The time it takes to collect a rating on VAS i is an interval $[t_0,\ t_1]$ such that $t_1 - t_0 > 0$ s, where t_0 and t_1 denote the beginning and end of the rating task. We will assume that changes in tense arousal within this interval can be ignored.

$$\forall\ t \in [t_0,\ t_1],\ Q(t) = q. \tag{7}$$

Hereafter, t denotes a *moment*, that is, a *short* time interval, and the auxiliary hypothesis (7) is assumed.

The respondent's rating on VAS i at t viewed as an ordinal measurement of an amount of tense arousal. To interpret the respondent's mark on a VAS i as a measurement, we need to compose the functions Q and F_i as follows:

$$y_i = (F_i \circ Q)(t) = F_i[Q(t)] = F_i(q). \tag{8}$$

Measurement Function of Respondents' Ratings on the Six VASs

Let the vector (or 6-tuple) $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_6)$ denote the results of the measurements performed with VASs 1, 2, ..., 6, through the functions F_1, F_2, ..., F_6, at t_1, t_2, ..., t_6. Thus,

$$\mathbf{y} = [(F_1 \circ Q)(t_1),\ (F_2 \circ Q)(t_2),\ \ldots,\ (F_6 \circ Q)(t_6)]. \tag{9}$$

A second auxiliary hypothesis that a testable property of vectors observed in a longitudinal individual design can be derived. The overall hypothesis is that the vectors of marks resulting from ratings made by a respondent during a given assessment can be tested if one assumes that the amount of tense arousal varies across a limited range between t_1 and t_6. Let us formulate the extreme assumption of no change, as follows (this assumption is slightly relaxed below):

$$Q(t_1) = Q(t_2) = \cdots = Q(t_6) = q. \tag{10}$$

Accordingly,

$$\mathbf{y} = [F_1(q),\ F_2(q),\ \ldots,\ F_6(q)], \tag{11}$$

and the multivariate ordinal measurement function is

$$\begin{aligned} \mathbf{F} = F_1F_2 \ldots F_6 : [0,\ \max] &\to [0\text{ mm},\ 102\text{ mm}]^6 \\ q &\mapsto \mathbf{F}(q) = \mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_6). \end{aligned} \tag{12}$$

Deducing the comparability of two vectors. Let us assume that Respondent 1 rated her/his mood twice, and that q and q' were the amounts of tense arousal at the first and second assessments . If $q = q'$, then $\mathbf{y} = \mathbf{y}'$. If $q < q'$, then $F_i(q) \leq F_i(q')$ for $i = 1,\ 2,\ 3$, and $F_i(q) \geq F_i(q')$ for $i = 4,\ 5,\ 6$. If $q > q'$, then $F_i(q) \geq F_i(q')$ for $i = 1,\ 2,\ 3$, and $F_i(q) \leq F_i(q')$ for $i = 4,\ 5,\ 6$. We will call these structural constraints on $\mathbf{y}$ and $\mathbf{y}'$ the property of *comparability*.

Conversely, the vectors $\mathbf{y}$ and $\mathbf{y}'$ will be *incomparable* if two VASs of the same sign exhibit changes in opposite directions, or if two VASs of opposite signs exhibit changes in the same direction. For example, coming back to Respondent 1's vectors $\mathbf{y} = (24,\ 50,\ 58,\ 54,\ 71,\ 78)$, and $\mathbf{y}' = (66,\ 34,\ 60,\ 70,\ 62,\ 62)$, we can see that VAS 1 exhibits an increase from 24 mm to 66 mm, while VAS 2 exhibits a decrease from 50 mm to 34 mm. Thus, $\mathbf{y}$ and $\mathbf{y}'$ are incomparable.

Testing that Two Observed Vectors are Ordinal Measurements

If $\mathbf{y}$ and $\mathbf{y}'$ are incomparable, we are led to conclude that the amount of tense arousal increased *and* decreased. The auxiliary hypothesis (10) is used here to render such a conclusion absurd: either the amount remains constant, or it changes, in which case either it increases or it decreases. This hypothesis (11) is testable because it precludes the occurrence of incomparable observed vectors.

However, it can be argued that the observed vectors are not necessarily true. For example, $F_i(q)$ is defined as a point between 0 mm and 102 mm but the mark on VAS i is not a point. Consequently, there is no exact description of $F_i(q)$. We would reply that $F_i(q)$ simply needs to be surrounded, using an interval of approximation, in such a way that we can always decide what counts as a true

change between $F_i(q)$ and $F_i(q')$. We can decide not to accept the observation of a true change between y_i and y'_i if $|y_i - y'_i| < s$, where s is a convenient – conventional – threshold. The choice of s also means that we do not have to assume that the auxiliary hypothesis (10), according to which q is invariant during the rating task, holds perfectly. For example, we need to take $s \geq 17$ mm – that is, the convention of an approximation interval of at least ± 8.5 mm – to save the contention that the overall measurement hypothesis holds for Respondent 1. With this convention, only VAS 1 (tense) allows us to detect an increase in tense arousal, as the vector of the observed differences can be written as follows:

$$\mathbf{y}' - \mathbf{y} = (+42,\ 0,\ 0,\ 0,\ 0,\ 0), \tag{13}$$

where the value 0 means that the differences are neglected if the approximation intervals for each mark ($s/2 = \pm 8.5$ mm) overlap.

Were we to take $s = 10$ mm as a reasonable value for the approximation interval of any mark, the vector of the observed differences would be

$$\mathbf{y}' - \mathbf{y} = (+42,\ -16,\ 0,\ +16,\ 0,\ -16), \tag{14}$$

thus falsifying the hypothesis. Consequently, psychologists cannot use these ratings as multivariate ordinal measurements of tense arousal, and must take this scientific knowledge into account when interpreting the data.

CONCLUSION

Psychologists wanting to assess how a patient's tense arousal has changed, using her/his test-retest ratings on a series of three negative and three positive VASs, may adopt CTT to compute one composite difference score based on the negative ratings, and one composite difference score based on the positive ratings. However, if they do so, they face two problems: (1) they have to believe that the true-score difference associated with the first composite difference measures a change in tense arousal, in which case they ignore what happens to the second composite difference measure, and (2) they cannot deduce the sign of the true-score difference because the measurement error is not usefully restricted (measurement error ranging from -102 mm to $+102$ mm), even if previous studies have reported good reliability estimates based on nice structural properties (e.g., [19]). Although traditional SEM studies based on CTT may support structural hypotheses about multivariate moments, they are not designed to document the use of data in individual assessment settings.

Psychologists need to think about what *measurement* means in the practical setting that concerns them. We developed a causal view, according to which for an observed event to be a measurement, it has to result from a causal process that links the amount of something to be measured to this event through a measurement function (see also [5]). From this theoretical perspective, respondents are viewed as responding systems that have no choice in the way they behave: the marks on the six VASs depend functionally on their amount of tense arousal – just as the height of the mercury in a thermometer depends on the amount of temperature (all other relevant causal parameters being equal). One immediate objection to this view is that respondents are not systems but people: one main difference between systems and people is that the latter respond to the rating task according to their intention to put marks on the VASs. This means that psychologists must be aware of the fact that the assessment occasion is a conversational setting. A person cannot be a measurement instrument, since a measurement instrument (e.g., a thermometer) has no intentionality and no sense of behaving in a conversational setting.

Psychologists may therefore ask whether patients' ratings have properties that would be expected from a measurement instrument incorporating a multivariate measurement function $\mathbf{F}$ as defined above (12), thus allowing their observations to be used for assessment purposes. For example, Respondent 74 in [19] study produced the ratings $\mathbf{y} = (40,\ 36,\ 36,\ 36,\ 39,\ 28)$ and $\mathbf{y}' = (16,\ 5,\ 13,\ 68,\ 64,\ 53)$, yielding the differences

$$\mathbf{y}' - \mathbf{y} = (-24,\ -31,\ -24,\ +32,\ +25,\ +25), \tag{15}$$

thus corroborating the multivariate ordinal measurement hypothesis. Psychologists could therefore deduce from the measurement hypothesis and the data that this patient's tense arousal decreased. Importantly, the measurement hypothesis cannot be blindly generalized to any respondent, as it is in item response theory (e.g., [3]), as each respondent has to be thought of as a specific rating system with a specific function $\mathbf{F}$.

If their observations falsify the measurement hypothesis, as they did for 21.6% of [19]'s selected cases with $s = 10$ mm, psychologists must investigate the potential meaning of patients' ratings from the point of view of the patients themselves, instead of the point of view of an untenable measurement hypothesis. For example, Respondent 1 could be asked if he intended to say that he felt more tense (from 24 to 66 mm) and less strained (from 50 to 34 mm), and, if the answer was "yes", what he wanted to say.

[1] Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press, Oxford.

[2] Elster, J. (2011). Hard and soft obscurantism in the humanities and social sciences. *Diogene*, 58:159–170.

[3] Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.

[4] Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology*, 14:83–102.

[5] Lacot, E., Afzali, M. H., and Vautier, S. (2015). Test validation without measurement: Disentangling scientific explanation of item responses and justification of focused assessment policies based on test data. *European Journal of Psychological Assessment*.

[6] Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA.

[7] McGrane, J. A. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, 6:431.

[8] Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88:355–383.

[9] Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, 10:639–667.

[10] Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometricians' fallacy. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 17:414–432.

[11] Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, 6:7–24.

[12] Michell, J. (2009). The psychometrician's fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, 62:41–55.

[13] Notturno, M. A. (2000). *Science and the open society: The future of Karl Popper's philosophy*. Central European University Press, Budapest.

[14] Notturno, M. A. (2009). Three concepts of science. *Scientific Medicine*, 1:2–4.

[15] Petocz, A. and Newberry, G. (2010). On conceptual analysis as the primary qualitative approach to statistics education research in psychology. *Statistics Education Research Journal*, 9:123–146.

[16] Popper, K. R. (2013). *The Open Society and Its Ennemies*. Princeton University Press (Original work published 1945), Princeton, NJ.

[17] Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science*, 42:509–524.

[18] Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:667–680.

[19] Vautier, S. (2011). Measuring change with multiple visual analogue scales: Application to tense arousal. *European Journal of Psychological Assessment*, 27:111–120.

[20] Vautier, S. (2015). La psychotechnique des aptitudes : pour différencier une sociotechnique de l'évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance. *Pratiques Psychologiques*, 21:1–18.

[21] Vautier, S., Lacot, E., and Veldhuis, M. (2014). Puzzle-solving in psychology: The neo-galtonian vs. nomothetic research focuses. *New Ideas in Psychology*, 33:46–53.

[22] Vautier, S. and Pohl, S. (2009). Do balanced scales assess bipolar constructs? the case of the stai scales. *Psychological Assessment*, 21:187–193.

[23] Vautier, S., Steyer, R., Jmel, S., and Raufaste, E. (2005). Imperfect or perfect dynamic bipolarity? the case of antonymous affective judgments. *Structural Equation Modeling*, 12:391–410.

[24] Vautier, S., Veldhuis, M., Lacot, E., and Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative founding of socially relevant avatars. *Theory & Psychology*, 22:810–822.

[25] Zealley, A. K. and Aitken, R. C. B. (1969). Measurement of mood. *Proceedings of the Royal Society of Medicine*, 62:993–996.