GV measurements, dynamical measurements and classical psychometric measurements.

By Wal Torres, MSc., Ph.D.

Gendercare Gender Clinic

Abstract:

The main goal of our work is not the research of GV- gender variance and GI – gender identity formation etiology. On the contrary, the main goal of our work is to measure GI formation dynamics and compare subjects with typical states – to develop a method of evaluation – of quantification – of GV – not considering any special etiology model and mainly never a simplistic one cause-one effect model. The dynamic quantification of GV is possible considering typical signatures – a method already considered in physics and neurobiology for example – with the background that we have a many causes – one effect system. Experimentally we define typical signatures for typical SOC 6th (WPATH standards of care 6th) states – and we may compare signatures. When someone shows a possible real mental disorder we classify as GIDNOS – as someone that need more special research. They are rare. Here we consider mainly also some aspects of the dynamic measurements and classical psychometric measurements – to try to explain why the new methods may not be compared with Gaussian or "classical" ones. A subsidiary consideration of our research considers the possible qualification of GI formation as a self-organized criticality considering that phenomenon follow a 1/f power law.

Preface

When I was in Chicago, presenting my paper at the 20th Biennial HBIGDA/WPATH Symposium about the background of our method for dynamic quantification of gender variance states – which enables the possibility to develop web-based gender variance evaluations, really I felt as if I landed in Chicago from Mars, Venus or perhaps Alfa-Centaur.

My communication there was really difficult. Not only due to a foreign language for me – but mainly a foreign accent – so variable accents - and a foreign culture – mainly a foreign scientific background and culture.

I am very engaged in the new many cause – many effect complex systems culture, far-from-equilibrium emergent states and variability/diversity is my home.

Their home was the old styled science – the one cause one effect linear point of view – something each day is more far from me – and my point of view.

So, here I put that quotation from Prof. Solé and Prof. Goodwin – I extrated from their preface, to their important book published in 2000*:*

"A remarkable burst of creativity in science is transforming traditional disciplines at an extraordinary rate, catalyzing movements whereby old boundaries are dissolving and newly integrated territories are being defined. The new vision comes from the world of complexity, chaos and emergent order. This started in physics and mathematics but is now moving rapidly into the life sciences, where it is revealing new signatures of the creative process that underlie the evolution of organisms. A distinctive sign of life is the emergence of new order out of the complexities of its material foundations. The concept of emergence, once regarded by many biologists as a vague and mystical concept with dangerous vitalist connotations, is now the central focus of the sciences of complexity. Here the question is: How can systems made up of components whose properties we understand well give rise to phenomena that are quite unexpected?....

It will become evident that the new understanding of complex processes takes us beyond the traditional scientific perspective of prediction and control of nature, to a relation of participation in natural processes that are unpredictable, though still intelligible."

(Ricard Solé & Brian Goodwin – Signs of Life – How Complexity Pervades Biology, Basic Books, 2000)

I hope these words will explain a little more, and better – what we are intending to do – and doing.

The main goal of our work **is not** the research of GV- gender variances and/or GI – gender identity **formation etiology.**

On the contrary, the main goal of our method is to measure GI formation dynamics and compare subjects with typical states – **not being considered which was the etiology that triggered that state.**

It is a method of evaluation – of **objective quantification** – to help local face-to-face – and web-therapists - to be sure if someone is a transsexual - when would probably need be referred to hormone therapy and surgeries including SRS –sex reassignment; or if someone has a transgender state and needs some surgeries, some hormone therapy and no SRS, or even if someone is a simple cross-dresser – or intergender – and may need some help.

With our method we are never limiting which are the causes – the complex causes of gender identity development, mainly when absolutely unexpected.

Our goal – that quantification – mainly through the web – looked like a fake procedure for some. As if we were doing a fake “tele-psychiatry”. Later, some still ask for validation and reliability data, considering concepts very different from our concepts. So we need that Introduction – and the Preface, to try to explain better our methods and goals.

A proposed pilot study – we accepted it immediately – was never improved.

But now we believe a pilot study (with a sponsor and not a free study from our part) would be a good idea if we could develop it with a serious partner – perhaps an American or European Institution as a partner – and sponsor.

Introduction

We will try to show here, the main mathematical background differences between classical psychometric measurements and dynamic measurements.

Classical psychometric measurement starts considering as a point of principle, that we will statically compare characteristics of someone with a group. So we define a population of “normals” and we compare the individual with that population, considering the statistical values (means, standard deviations, etc.) of the population and the values of the testing subject.

That approach considers the population distribution of results follows a “normal” Gaussian distribution, and tests of Reliability and Validity are performed considering these points of principle (we show at the end of that note the Wikipedia.org description of the classical psychometric statistic point of view).

Other more sophisticated statistical methods are used in psychometric measurement, as IRT- item response theory and Rasch models, for example, but the background is almost the same: comparison of answers/characteristics of one individual with a family of individuals, considering Gaussian distributions.

As we will see, these methods always consider static points of view – and not dynamic points of view. Also they always consider linear points of view – and not nonlinear ones.

When in psychometric practice someone would try to study “time-series”, always linear analysis were considered – auto-correlation, variance analysis, and so on. All these linear methods always consider Gaussian distributions as a point of principle.

Surely, these methods, more or less sophisticated, are absolutelly good and correct, the most adequate to evaluate static comparisons between individuals and Gaussian populations – if we consider persons or personality characteristics in psychology, or prices and values in econometric measurements or grain size distributions in a beach in chemistry – for example.

Really we know that most part of limited populations in Nature follow a Gaussian distribution (in stationary populations, most of times the distribution is Gaussian – for example among typical transsexuals we may find among them a Gaussian distribution) – the differences between individuals inside a population is most of the times – not always necessarily – Gaussian.

That way, if we would like to consider the necessity to compare someone with a Gaussian population, and the subject of the comparison was something psychological, with no doubth a classical psychometric test would be required as the best solution.

But we are not considering here any comparison between a subject and a Gaussian group of “normal” subjects.

So, we really believe we are not intending to develop – nor to use - “classical psychometric tests” methods.

Really what we would like to measure?

- We would like to measure the self-perception of being a man, a woman, or both, or none.

- To fulfill our main goal we would like to measure the evolution of that perception in time – defining time-series data that could be evaluated considering nonlinear dynamic methods.

- To fulfill a possible subsidiary goal we would like to have a measure of it as an intensity data to compare with published incidence data. That subsidiary goal may be important for the future, to understand better the mechanisms that may govern the variables that may be important for gender identity formation.

- We know:

1. that perception is a function of a lot of known and unknown factors;

2. it is surely unpredictable “a priori”, due to the factual existence of transgenders, crossdressers, intergenders and transsexuals, with and without any atypical sexual development;

3. we know, when from an almost same beginning, unpredictable states (low and high probability states) may happen – that is an unbiased signature of a chaotic system.

4. we also know a chaotic system to be analysed need nonlinear methodology (Sprott 2003);

- Considering:

1. We know not all variables. We know some… genes, hormones, brain tissues and groups of neurons … we know the mother`s stress affect it… also perhaps culture, family, rearing are important… and perhaps other factors may also be important. Surely nature is the background and nurture is a big influence - Nature and nurture in co-operation.

2. Surely that way we need to consider that many variables are generating many effects – which situation defines a complex nonlinear system.

3. We know if we study a many cause to many effect system at discrete space we may consider Taken’s theorem and reduce the problem to a many cause to one observable discrete and iterative dynamic system.

4. Is the structural unpredictability of the gender system something pathologic? All far-from-equilibrium state is really a pathologic state? That was, and is nowadays for medicine and psychology, most of the time the most considered point of view. But after the perception of the emergence of self-organized systems by Prigogine and others, that point of view need to be reviewed.

5. Is the structural unpredictability of gender system a derivation of natural variety and diversity? How may we establish when a system with far-from-equilibrium states, these low frequency states may be triggered by diversity and when there is a sign of "pathology"?

6. We know it is controversial, but we consider If a phenomenon follows an incidence vs intensity (an intensity spectrum) 1/f power law, that is a necessary (perhaps not necessarily sufficient) condition for diversity (about the incidence of 1/f power spectrum in human answers and in psychology research it is interesting to know the controversy between Gilden 2001; Van Orsen et al (2003, 2005); Thornton & Gilden 2005; Wagenmakers et al (2004, 2005) – for example);

7. In these cases, we may study dynamically that system.

8. We consider also if something is really "pathological", it necessary need to be disordered – and may not be deterministic.

Our main and subsidiary goal

The main goal of our method **is not the research of GV- gender variances and GI – gender identity formation etiology.**

On the contrary, the main goal of our method is to measure GI formation dynamics and compare subjects with typical states – to develop a method of evaluation – of quantification – to be sure if someone is a transsexual and need

hormone therapy and surgeries including SRS –sex reassignment ones with no mental problem – or if someone has a transgender state and need some surgeries, some hormone therapy and no SRS, or if someone is a simple crossdresser – or intergender – **not being considered which was the etiology that triggered that state.**

With our method we hope we will contribute for etiology studies – because we are never limiting which are the causes.

Considering we are studying a complex nonlinear system = a many causes – many effects system – we would like to make also a contribution to GV etiology studies – not considering causes – but considering what could be a dynamical qualification of the way the causes interact to trigger these effects.

Considering the inner unpredictability of the system, and knowing typical manifestations of its high probable states – male or female states – and its low probability states – GV states – we know that system is surely a chaotic system.

On the other hand we know that system of low probability states has a **gradation** of states.

We know that gradation, in a power spectrum diagram shows a near 1/f distribution – so we suggest a probable self-organized criticality as the most probable **etiology dynamic mechanism**.

What that means?

We know not all variables, but **probably** some early variables – as genes for example – may live a critical state, that could trigger other biological critical states in hormones, brain formation… and so on. Later life – rearing, culture, etc.. would also be important. Too much energy, too much critical states as a domino sequence… and finally a higher intense – and less frequent – situation may arise, as many higher frequencies and lower intensity states also could happen.

Incidence and Intensity measurement – a subsidiary goal.

How we measure incidence and intensity in Gender Space?

Incidence, Conway(2002, 2007) and others are developing their measurements – we use their work, we will not measure GV incidence in Brazil, nor in the world. We use the published data we have.

And intensity?

We developed some years ago – more precisely in 2001, a questionary we called MF9 for male assigneds and FM1 for female assigneds. We used our results with

these scales – simplistic psychometric scales – to look for intensity data. We used also other scales – for example Cogiati – among other possibilities.

We performed reliability and validity verifications for these questionaries – so they are reliable and valid – but we never published it due to difficulties to find an international publisher – and also due to the fact that they were not developed to identify people – and with no clinical value.

It was used only to see if the gender space is a 1/f power space or not – and it probably is.

I hope you understand, considering intensity scales we are not evaluating people, mainly we are not evaluating psychologically people, but we are only measuring the intensity to relate to incidence to see if that gender system of possible states is a fractal – a natural – phenomenon – measuring it`s incidence versus intensity power spectrum.

So there is never simplistic comparison with Gaussians – or “normals” – but only the measurement of the intensity of a phenomenon and its comparison with incidence data.

Dynamic Testing – our Main Goal

To understand the dynamic characteristics of something we need to study its movement.

Movement, as Newton showed in the past, is related to positions and velocity. With time. With space.

That way we defined a Gender Space – the space where gender identity develops – so moves.

When we have simplistic phenomena, we may have differential equations to define movement. Poincare discovered even with simple differential formulas, most of the time we may not precisely calculate them, when the number of variables increase.

That way, and from Poincare ideas, nowadays – and with the development of computers and numeric calculation of difference equations - we have all development of Chaos theory, and more recently far-from-equilibrium systems theory, emergence and self-organized theories and complex system theories.

First of all, to work in a continuum space, considering differential calculations is difficult – so we work in iteractive space – singular space – considering more simple difference equations and not differential equations – considering computers and iterative numerical techniques.

That way we may draw “maps”. Maps are the relation of the actual state of a phenomenon with its past states – in a discrete space. Computers are excellent instruments to calculate and plot maps.

Also we may define a phase space – where we may recognize typical movement conditions, transient states, stationary states. Phase space is a space of positions and gradients (momentum, velocities, etc).

Through the phase space analysis and returm map analysis we may qualify a development – as simplistic, periodic, chaotic, emergent, biotic, random – all these expressing order and determinism – or stochasticity and desorder.

Considering later a family of typical persons we may discover typicalities as “signatures” of groups of persons – never Gaussian necessarily and never necessarily “normals” or “abnormals”.

Read in the next pages some simple data about psychometry and later see our paper in Chicago. That introduction surely will clarify some points – I hope – as most of the people is not too much familiar with complex and dynamic nonlinear systems.

That way, there is possibility to “validadte” or “measure reliability” from these new methods – easily – comparing signatures graphically (Szücs et al. 2003; Amon & Lefranc, 2004 – they considered a similar dynamic approach, in other science fields, in neurobiology and in physics respectively – see the bibliography).

That dynamic and complex systems theory procedures and methods are used in a lot of different branches of science – from astrophysics to biology, from geology to artificial intelligence, from environmental sciences to social sciences, and also in psychiatry (Sabelli 2005) and life sciences.

Only we are **since 5 years** considering these new methods to measure and study gender identity formation – trying to understand the implication of our results for GV measurement, developing an objective method to know gender variability and diversity, helping people to know the self-development and comparing people with groups – some simple statistics here!!! – topologically and not simplistically – through signatures and not only means and standard deviations.

We know we have a whole universe of knowledge to learn and develop. We are aware we did some few steps only, and we are intending to do some few next steps as soon as we can.

Classical Test Theory (Wikipedia.org)

True and error scores

Classical test theory is based on the decomposition of observed scores into true and error scores. The theory views the observed score x of person i, denoted as x_i, as a realization of a random variable X. The person is characterized by a probability distribution over the possible realizations of this random variable. This distribution is called a "propensity distribution". Person i's true score, t_i, is axiomatically defined as the expectation of this propensity distribution. This definition is formally stated as

(Eq. 1) $$\varepsilon(X_i) = t_i.$$

Secondly, the so-called error score for person i, E_i, is defined as the difference between i's observed score and his true score:

(Eq. 2) $E_i = X_i - t_i.$

Note that X_i and E_i are random variables, but t_i is a constant. Also note that it directly follows from these definitions that the error score has expectation zero:

(Eq. 3) $$\varepsilon(E_i) = \varepsilon(X_i - t_i) = \varepsilon(X_i) - \varepsilon(t_i) = t_i - t_i = 0.$$

Relation to population

The above equations represent the assumptions that classical test theory makes at the level of the individual person. However, the theory is never used to analyze individual test scores; rather, the focus of the theory is on properties of test scores relative to populations of persons. Hence, the next step is to introduce a population-sampling scheme into the structure of classical test theory. When we assume that people are randomly sampled from a population, the true score becomes a random variable too, so that we get the (in)famous equation

(Eq. 4) $X = T + E$

Classical test theory is concerned with the relations between the three variables *X*, *T*, and *E* in the population. These relations are used to say something about the quality of test scores. In this regard, the most important concept is that of *reliability*. The reliability of the observed test scores *X*, which is denoted as ρ^2_{XT}, is defined as the ratio of true score variance σ^2_T to the observed score variance σ^2_X:

(Eq. 5) $$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}.$$

Because the variance of the observed scores can be shown to equal the sum of the variance of true scores and the variance of error scores, this is equivalent to

(Eq. 6) $$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}.$$

This equation, which formulates a signal-to-noise ratio, has intuitive appeal: The reliability of test scores becomes higher as the proportion of error variance in the test scores becomes lower and vice versa. The reliability is equal to the proportion of the variance in the test scores that we could explain if we knew the true scores. The square root of the reliability is the correlation between true and observed scores.

Reliability

Note that reliability is not, as is often suggested in textbooks, a fixed property of tests, but a property of test scores that is relative to a particular population. This is because test scores will not be equally reliable in every population. For instance, as is the case for any correlation, the reliability of test scores will be lowered by restriction of range. Thus, IQ-test scores that are highly reliable in the general population will be less reliable in a population of college students. Also note that test scores are perfectly unreliable for any given individual *i*, because, as has been noted above, the true score is a constant at the level of the individual, which implies it has zero variance, so that the ratio of true score variance to observed score variance, and hence reliability, is zero. The reason for this is that, in the classical test theory model, all observed variability in *i*'s scores is random error by definition (see Eq. 2). Classical test theory is relevant only at the level of populations, not at the level of individuals.

Reliability cannot be estimated directly since that would require one to observe the true scores, which according to classical test theory is impossible. However, estimates of reliability can be obtained by various means. One way of estimating reliability is by constructing a so-called *parallel test*. A parallel test is a test that has the property that, for

every individual, it yields the same true score and the same observed score variance as the original test. If we have parallel tests x and x', then this means that

(Eq. 7) $$\varepsilon(X_i) = \varepsilon(X_i')$$

and

(Eq. 8) $$\sigma_{E_i}^2 = \sigma_{E_i'}^2.$$

Under these assumptions, it follows that the correlation between parallel test scores equals reliability (see Lord & Novick, 1968, Ch. 2, for a proof).

(Eq. 9) $$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2.$$

The estimation of reliability by the use of parallel tests is cumbersome, because parallel tests are very hard to come by. In practice the method is rarely used. Instead, researchers use a measure of internal consistency known as Cronbach's α. Consider a test consisting of k items u_j, $j = 1, \ldots, j, \ldots, k$. The total test score is defined as the sum of the individual item scores, so that for individual i

(Eq. 10) $$X_i = \sum_{j=1}^{k} U_{ij}.$$

Then Cronbach's alpha equals

(Eq. 11) $$\alpha = \frac{k}{k-1} \frac{\sum_{j=1}^{k} \sigma_{U_i}^2}{\sigma_X^2}.$$

Cronbach's α can be shown to provide a lower bound for reliability under rather mild assumptions. Thus, the reliability of test scores in a population is always higher than the value of Cronbach's α in that population. Thus, this method is empirically feasible and, as a result, it is very popular among researchers.

As has been noted above, the entire exercise of classical test theory is done to arrive at a suitable definition of reliability. Reliability is supposed to say something about the general quality of the test scores in question. The general idea is that, the higher reliability is, the better. Classical test theory does not say how high reliability is supposed to be. In the literature a value over .80 appears to be deemed 'acceptable'; a value over .90 is 'good'. Values between .70 and .80 are seen as mediocre but still defensible; values below .70 are bad.[*citation needed*] It must be noted that these 'criteria' are not based on

reasonable arguments but the result of convention. Whether they make any sense or not is unclear.

Alternatives

Classical test theory is by far the most influential theory of test scores in the social sciences. In psychometrics, the theory has been superseded by the more sophisticated models in Item Response Theory (IRT). IRT models, however, are catching on very slowly in mainstream research. One of the main problems causing this is the lack of widely available, user-friendly software; also, IRT is not included in standard statistical packages like SPSS, whereas these packages routinely provide estimates of Cronbach's α. As long as this problem is not solved, classical test theory will probably remain the theory of choice for many researchers.

See

http://en.wikipedia.org

for details about psychometry.

Bibliography

Amon, A & Lefranc, M --- Topological signature of deterministic chaos in short nonstationary signals from an optical parametric oscillator --- *Physical Review Letters, vol92, number 9, 2004;*

Delignières, D; Fortes, M; Ninot, G --- The Fractal Dynamics of Self-Esteem and Physical Self --- *Nonlinear Dynamics in Psychology and Life Sciences, 8, 479 – 510, 2004;*

Gilden, D. L. --- Cognitive emissions of 1/f noise. --- *Psychological Review, 108, 33-56, 2001;*

Szücs, A; Pinto, RD; Rabinovich, MI; Abarbanel, HDI; Selverston, AI --- Synaptic Modulation of the Interspike Interval Signatures of Bursting Pyloric Neurons --- *Journal of Neurophisiology, 89 : 1363 – 1377, 2003;*

Thornton, TL & Gilden, DL --- Provenance of correlations in psychological data --- *Psychonomic Bulletin & Review*, **12 (3)**, 409-441*, 2005;*

Torres, W --- Web-evaluation of Gender Variances: background and methods --- Published as a free PDF download at www.gendercare.com, 2007;

Van Orden, G. C., Holden, J. G., & Turvey, M. T. --- Self-organization of cognitive performance.--- *Journal of Experimental Psychology: General*, *132, 331-350, 2003;*

Van Orden, GC; Holden, JG; Turvey, MT --- Human Cognition and 1/f Scaling --- *Journal of Experimental Psychology : General, 134, 117 – 123, 2005;*

Wagenmakers, EJ; Farrell, S; Ratcliff, R --- Human Cognition and a Pile of Sand: A Discussion on Serial Correlations and Self-Organized Criticality --- *Journal of Experimental Psychology : General, 134, 108 – 116, 2005;*

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. ---. Estimation and interpretation of 1/*ƒ* noise in human cognition.--- *Psychonomic Bulletin & Review*, **11**, 579-615*, 2004;*