The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars

Stéphane Vautier, Michiel Veldhuis, Emilie Lacot, Nadine Matton

HAL Id: hal-00667697
https://hal.science/hal-00667697

The Ambiguous Utility of Psychometrics for the Interpretative Foundation of Socially Relevant

Avatars

Stéphane Vautier[1], Michiel Veldhuis[1], Emilie Lacot[1], and Nadine Matton[2]

[1]Université de Toulouse

[2]Ecole Nationale de l'Aviation Civile

Author note

Stéphane Vautier, OCTOGONE, Université de Toulouse; Michiel Veldhuis, CLLE-LTC, Université

de Toulouse, Emilie Lacot, OCTOGONE, Université de Toulouse, and Nadine Matton, Ecole

Nationale de l'Aviation Civile, Toulouse. Correspondence should be addressed to Stéphane Vautier,

OCTOGONE, Pavillon de la Recherche, Université de Toulouse-Le Mirail, 5 allées A. Machado, F-

31058 cedex 9, France. Electronic mail may be sent to vautier@univ-tlse2.fr.

Abstract

The persisting debates that measurement in psychology elicits can be explained by the conflict between two aspiration types. One, the epistemologic aspiration, resting on the search for scientific truth, and two, the social aspiration, resting on the demonstration of a capacity to contribute to psychological assessment problems in particular. Psychometrics answer essentially to psychology's demand for social utility, leading to the quasi-exclusive attribution of importance to quantitative interpretation. For psychology to be considered an empirical science, it has to establish its capacity for the measurement of psychological phenomena, even if this means that it recognizes that these phenomena are essentially qualitative.

The Ambiguous Utility of Psychometrics for the Interpretative Foundation of Socially Relevant

Avatars

The persistence of critical publications about measurement in psychology bears witness to a profound crisis at the heart of the discipline (e.g., Borsboom, 2003; Cliff, 1992; Dooremalen & Borsboom, 2010; Essex & Smythe, 1999; Haig & Borsboom, 2008; Hood, 2009; Johnson, 1936; Lumsden, 1976; Michell, 1990; 1997; 2000; 2003a; 2005; 2008; Porter, 2003; Trendler, 2009; Zumbo & Rupp, 2004). We interpret this crisis as the result of a conflict between two types of aspirations. The first aspiration type is epistemologic: psychology needs to be edified as an empirical science, implying that it respects the empirical facts it observes. This includes the facts that imply a definitive absence of empirical laws in a given descriptive reference system (Vautier, 2011a; 2011b), as for example the absence of a conjoint order relatively to a set of test items (Bertrand, El Ahmadi, & Heuchenne, 2008; Guttman, 1944). The second aspiration type is political: psychology hopes to become an academic discipline that is capable of not only satisfying but also creating a social demand, implying that it respects the way social problems, particularly psychological assessment problems, are proposed by the employers of psychologists. It becomes ever more difficult to ignore the respective divergent implications of these two aspiration types. Whereas the scientific aspiration leads to the abandonment of the dream of discovery of authentic quantitative psychological properties (Michell, 2003b; 2004; Trendler, 2009), the social utility aspiration leads to the defense of the evaluation of these properties as the basis of psychologists' work (Brown, 1992; Danziger, 1990; Martin, 1997; Porter, 2003). Effectively the social demand rests on anthropological simplifications based on value scales. This in order (i) to be able to think that human beings can be ordered according to their psychological properties (e.g. the benefit of a course or sexual offenders' recidivism), and (ii) to be able to objectify their evolution (e.g. therapeutic progress). These practical exigencies create the necessity of psychological anthropometrics, i.e. a science of construct measurement (Cronbach & Meehl, 1955; Maraun,

1998). The promotion of this science has, from the start of the 20th century, been assured by psychotechnic engineering (Brown, 1992; Martin, 1997).

A lot of effort has been expended to develop psychometrics. We think it useful to distinguish psychotechnic engineering from psychometric engineering, the development of the latter following chronologically on that of the former. Psychotechnic engineering responded to a need of aggregating a set of qualitative observations into a score, and the optimisation of this score (Cronbach & Meehl, 1955; Messick, 1995). The purpose of psychometric engineering, as a discipline of psychotechnic data modeling, is less clear (Thissen, 2001). In this article we support the point of view that psychometric engineering is interesting for construct psychologists, because of the mere conceptual possibility of these constructs, and not because these statistical models based on continuous latent variables confirm some empirical truth.

To respond to the social demand of the comparison of human beings relatively to constructs, the testing psychologist needs his scores to have a significance as construct measures (see Cliff & Keats, 2003, p. 16). However, tests in themselves do not allow one to conclude about the existence of constructs as continuous quantities. A psychometric model associated to a test rests on the explicit imagination of a causal role of the construct that the test is supposed to measure (Borsboom, 2008; Borsboom, Mellenbergh, & van Heerden, 2003). As such the psychometric model constitutes the probabilistic pole of the construct measurement paradigm.

However its probabilistic nature renders it nomothetically sterile - inept at the discovery of laws - as it is tautological, and epistemologically counter-productive. This because psychometric estimations require empirical conditions - inter-individual aggregation - that deny the individual nature of events occurring in the space where these probabilities have their foundations. Which leads to the conclusion that (i) scientific psychology has to put itself at distance from the social demand of the evaluation of human beings, to be able to fulfill its quest for the truth and that (ii) the psychotechnic engineering has to admit that its lack of scientific foundations calls for a political foundation (Dagonet, 1993).

In this article, (i) the production process of psychotechnic data as well as the conceptual problems that the discrete scores cause will formally be discussed. Then (ii) an analysis of how the classical test theory solves these problems is performed. Also (iii) the solution that item response modeling for dichotomic items proposes will be discussed. Then (iv) the unfalsifiable character of the probability of a response given a certain true score or latent trait comes next. The application of a psychometric model demands a leap from the individual level to the aggregate level to permit a latent score estimation; then (v) it will be shown that the estimation of a model is not equivalent to a validity test of the generic interpretation. Finally, (iv) the epistemologic and ethical consequences of this change of empirical phenomenon apprehension level will be analyzed.

From standardized observation to objective quantitative assessment: Psychotechnics

The production of psychotechnical scores can be described as a composite function that allows the description of people via an empirical process of standardized observation (testing) followed by a symbolic process of quantification of this description (numerical coding). These two applications are described in the following (for a more technical analysis, see Vautier, Hubert, & Veldhuis, 2011).

A function is a relation defined from a set, called its domain, on a set, called its codomain, such as any element of the domain is related to one and only one element in the codomain (e.g., Selby & Sweet, 1963). We note Ω the domain of the former, i.e., the application of standardized description, and ω an element of Ω. Every element of Ω is a couple (ordered pair) (person, date of observation), that is noted $\omega = (u, t)$. The observation at a moment in time of a person ensures the unicity of the empirical objects ω what leads the description to be a function - someone that has been evaluated at two different dates can be described differently. We note M^k the codomain of the application, comprising the possible response patterns. M^k is a finite set, most of the time the Cartesian product of several (k) descriptive sets $\{x_{i1}, x_{i2}, \dots, x_{im(i)}\}$ (i = 1, …, k), where the $m(i)$

modalities x_i are nominal or ordinal values. For example, in a test of 50 items with scores in (0, 1) every ω can be attributed an 50-tuple in $M^{50} = (0, 1)^{50}$.

The standardized description application, that we also call *psychometric description*, accomplishes a double objectivation, whence the epistemological importance is fundamental for the empirical constitution (and thus scientific) of psychological facts. Firstly, every object ω has necessarily a state in the set M^k of possible descriptions. Would this not be the case, the description of ω could not be universal. Secondly, the determination of ω's state does not depend on specific characteristics that identify its observer. The double objectivity of the psychotechnic description permits to attribute to every proposition that specifies the state of ω a trueness value (true or false) that is epistemologically *accessible*. These are the facts that can be observed in M^k that permit to falsify a theory of what is happening - this theory will be called general *and* falsifiable if it excludes at least the possibility of one state (Popper, 1959; Roberts & Pashler, 2000; Vautier, 2011a; 2011b). The second application is of a symbolical nature in the way that it does not depend on what is happening in the set Ω: it is about the attribution of one and only one score to every state of the set M^k of possible descriptions. We call this application a *numerical*, *metaphorical coding*.

The composite function, of the standardized description and the numerical coding applications, permits to attribute a single score to every ω. Projected on the scale of test scores, every ω_i can be compared to every ω_j with the help of the algebraic operator $\leq$. The composite function of these applications equips the psychologists with the tools to cope with the social problems that demand an objective comparison - i.e., impersonal and consensual (see Porter, 2003) - of people.

The semantic register of this comparison's interpretation is not exactly linked to that of the set M^k of possible responses on the test. The interpretation of the comparison mobilizes the semantics of the *construct* as a total order. Consider a construct as numerical ability for example. The semantics of numerical ability are such that a sentence like "Paul is (whatever the date we

observe him) numerically more able than Jean (whatever the date we observe him)" has a "social" significance.

But the psychotechnical procedure of projecting a person on a scoring scale reveals a logical problem. Suppose that the scoring scale of an ability test is the following sequence (0, 1, …, 50). The semantics of numerical ability do not specify how it is possible that the ability variations are organized in 51 steps, because the concept of numerical ability is intuitively construed. By default, the non-specification of the construct leads numerical ability to vary on a continuous scale. As such it would be impossible to measure a score of 30.32 points with this test for example. If, as a *measurement instrument*, the test would measure a continuum, it should *operationalize* the principle of continuity that it measures. Thus, it is embarrassing to assert that the test *measures* numerical ability, while the issue at hand is to understand how the test *could* measure a continuous quantity, whereas it fabrication principle does not permit this continuity to be established (Vautier, Gillie, & Veldhuis, 2011).

The metaphysical meaning of test scores (1): Classical Test Theory

Classical test theory proposes the following solution: if the observable scores vary in a ascending sequence $(y_0, y_1, …, y_z)$, then the *true score* varies in the continuum $[y_0, y_z]$. But, as we will show, the true score concept does not have an empirical significance.

(1) Write down the time (t). Let someone (u) make the test and save his score. (2) Go back in time to a moment before moment t. (3) At moment t let u make the test again and save his score again. (4) Forget about mortality and repeat operations (2) and (3) until infinity. (5) At the end of infinity calculate the mean score: now you have discovered the true score of (u, t). This score, which is also called a mathematical expectation (Lord & Novick, 1968; Steyer, 2001), is an empirical impossibility, nothing but a thought experiment (Borsboom, Mellenbergh, & van Heerden, 2002). As such, the construct "numerical ability, measured with the numerical ability test" has a pseudo-operative significance.

For now, the true score has been construed from an imaginary series of scores. As such, the true score depends *logically* on the imaginary series of scores. But the construct measurement psychologist looks to comprehend how a score on a test can be interpreted as the *result* of a measurement process. In this perspective, forget about observed scores and consider that *there is* a true score: existing independently of the series of test scores. By doing this, a metaphysical postulate is called upon, that permits to surpass the paradox of the non-operational status of something that is only defined by a thought experiment. From now on, the true score is ontologically primary, instead of the mean score of several measures.

Then there is the concrete measure (i.e., the test administration and the calculated score as a function of made observations), that leads to biases. These biases are added to the true score, which leads to the observed score. This has nothing to do with realism, neither has the consideration of a construct as a continuous quantity for that matter. Let us write this in a mathematical fashion: let τ (tau) be the true score, ε (epsilon) the bias, and y the observed score.

$$y = \tau + \varepsilon. \tag{1}$$

Now it has to be shown how τ varies in the continuum $[y_0, y_z]$. For this, it suffices to take a step further in the definition of τ. This step will allow us to show how a second spiritual point, corollary to the first, is mobilized: the probability of observing a score knowing τ. Let there be $z+1$ probabilities: the probability, knowing τ, to obtain the score y_0, is noted p_0; the probability, knowing τ, to obtain the score y_1, is noted p_1; etc. until the probability, knowing τ, to obtain the score y_z, is noted p_z. These probabilities are nothing but the frequency of every score, when this frequency is calculated with the infinity of measure points that the initial thought experiment provided. We also admit that we will always get a score when the test is administered (we neglect the situation where a participant refuses to respond). As such, the sum of the probabilities of every score equals 1:

$$p_0 + p_1 + \cdots + p_z = 1. \tag{2}$$

With these probabilities, another definition of the true score can be given:

$$\tau = p_0 \cdot y_0 + p_1 \cdot y_1 + \cdots + p_z \cdot y_z. \tag{3}$$

The appendix shows formally that τ varies in $[y_0, y_z]$. Here we will stick with an intuitive approach. Does τ have a minimum and a maximum? With some possible scores and some probabilities, one can do whatever one wants as long as their sum remains equal to 1. To minimize τ, the probability to have the minimal score has to be maximized, meaning that p_0 has to be 1, implying that all the remaining probabilities equal 0. Which leads to $\min(\tau) = y_0 \cdot 1 = y_0$. Analogous reasoning leads to $\max(\tau) = y_z \cdot 1 = y_z$. In addition to this, τ can take every real value as the $z+1$ probabilities vary in $[0, 1]$.

The construction of the true score shows two philosophical problems. The first problem derives from the volte-face that we have to do in order to call a series of scores on a test the true score: after all, nothing proves that this true score is identical to the true score that the psychologist wants to see as the metaphysical cause of the observed score. The second problem is due to the fact that the true score in classical test theory is contingent on the test. As such, the metaphysical psychologist has to elaborate another solution to show how it is possible that true scores linked to different tests can be caused by the same construct.

The metaphysical meaning of test scores (2): Item Response Theory

Item response modeling tackles the problem of the relation between the set M^k of possible descriptions and the continuum on the item level instead of the test level. For simplicity reasons we only consider dichotomic items (i.e. items varying in $\{0, 1\}$).

The basic idea is very simple, analogous to the true score, albeit a bit more sophisticated. Let there be a continuum and suppose that a person that responds to an item has a position on this continuum, noted θ (theta). No thought experiment is needed, the continuum exists by definition and every object $\omega = (u, t)$ has a position on it. Knowing this position on item I_i, we admit that the probability p_{i1} to find event '1' exists (the probability $p_{i0} = 1 - p_{i1}$ to find event '0' as well). This leads to the simple interpretation of the observed response as a Bernoulli trial (e.g., Feller, 1966). The expected value of the score is thus p_{i1} (indeed: $\tau = 0 \cdot p_{i0} + 1 \cdot p_{i1}$). Formulated differently, the

true score on item I_i depends on θ. But to the metaphysical psychologist θ is the quantity of interest, not the true score.

The generalization of this conceptualization to any item I_i leads to the postulation of the existence of probability p_{j1} knowing ω, observing event '1' on item I_j. An item response model stipulates that the true score (i.e., p_{i1}) depends in an ascending manner on θ: the bigger the value on the latent trait, the more the true score approaches 1. The shell game, that needs to be played in classical test theory, consisting of defining the true score as an expected value and then considering that it exists independently of the scores on which it is based, is of no need here. The postulate of the continuum forms the foundation, because there is only one continuum for as many items as are supposed indicating it.

The two approaches (classical test theory and item response modeling) permit one to conceptualize how a discrete event (the score on a test) depends on the latent trait from a continuum. The advantage of item response modeling is that it clarifies the hypostatic state of the continuum at the same time that it makes the true score logically dependent on its indicator. We will now show why propositions from these two approaches can never be falsified.

Unfalsifiability of probabilistic statements about single random experiments

A proposition that can be falsified is a proposition that can be found to be false. In order for a proposition to be able to be falsified, its negation has to be able to be verified. A theoretical proposition founded on an empirical science can be falsified if it permits the *possibility* to verify an empirical fact which it excludes theoretically (Popper, 1959). In all the preceding theory, can a proposition that can be falsified be found? No, not if the theory *saturates* the observation reference system, which is the case in all of the above - a theory saturates the reference system if it proposes no theoretical impossibility (cf. Vautier, 2011a; 2011b).

Let us consider the proposition (1): $y = \tau + \varepsilon$. Let y be an observed score in $(y_0, y_1, \ldots, y_z)$. A number τ exists in $[y_0, y_z]$ and a real number ε so that their sum equals y. This does not exclude any event in $(y_0, y_1, \ldots, y_z)$. Thus this proposition cannot be falsified.

If we go a little bit further and examine the proposition

$$p_0 = p_1 = \cdots = p_z = 1/(z+1) \tag{4}$$

(and $\tau = \frac{1}{z+1} \cdot (y_0 + y_1 + \cdots + y_z) = (y_0 + y_z)/2$. Can this proposition be falsified? For a person and an experimental condition, every observable score is compatible with these probabilities because the theory does not limit the domain of definition of ε.

In order for a proposition of this kind to be falsified it should at least stipulate that one of the probabilities equals zero, in which case it would preclude the scores of which the probability of their observation is zero. Another possibility would be the restriction of the domain of definition of ε (for example, if we predict that the probabilities equal $1/(z+1)$ and we restrict ε's domain to [-3, 3], then we predict that $y \in \left[\frac{y_0+y_z}{2} - 3, \frac{y_0+y_z}{2} + 3\right)$). The classical test theory does not specify any case of this kind.

The probabilities that classical test theory proposes are entities that do not have a Popperian scientific significance. However it does propose a theory of what is measured, based on (1) the disguised postulate of a continuum and (2) on the postulate of the conditional probabilities of observed scores. It is an interpretative theory that can be of use if the sole goal is to obtain an interpretation of a score as an intuitive quantity measure.

Now consider the probabilistic modeling of the response on item I_i of a test. By definition, every object $\omega = (u, t)$ has a score θ (the continuum postulate is not disguised anymore). While the theory does not give a zero (or certain) probability to an event '1', every proposition that specifies the values of θ and p_{i1} cannot be falsified (the case where θ tends to infinity does not have intuitive significance). As in classical test theory, item response modeling uses probabilities whereof the scientific significance remains uncertain. This scientific significance is doomed to remain uncertain,

as one would have to study the distribution of '0' and '1' while controlling θ, and a measurement instrument that measures θ does not exist (see Trendler, 2009). In other words to consider the unique response of a person at a certain moment in time as the result of a Bernoulli trial is a thought experiment as well. This is why psychometric models are considered to lack a scientific vocation (in a Popperian way), leaving only an interpretative vocation. Remarkable in this regard is that specialists of latent variable models as Muthén and Muthén (2006) declare in the introduction of their program's user manual that "The purpose of modeling data is to describe the structure of data in a simple way so that it is understandable and interpretable" (p. 1).

From the individual level to the collective level: The psychometric model as a pre-scientific fiction

The impossibility to study the probability of observed scores of an object ω empirically leads to the abandonment of the individual in favor of the collective scale to fit a model to the data. Here one needs to pose the question "why fit a model to the data"? Up to this point it has been admitted that a psychometric model permits one to explain how scores can be (probabilistic) measures of a continuous latent quantity. Consequently, it suffices that the model be recognized as a conceptual *possibility* so that the conceptual reconciliation of the observation reference system (the score scale or M^k) with the construct is finished.

The estimation of a psychometric model consists of the specification of a null hypothesis (meaning a vector v of υ numerical values belonging to a set M^{υ}) in a way to minimize a measure of the distance between the data and the modeled image of the data. In itself the psychometric model is not a falsifiable theory. It is a model without a falsifiable theory. The model could be useful to falsify a psychological theory if the theory restricted the space of theoretically possible parameters to a strict subset of M^{υ}, in a way that it would always be possible to find, on an empirical basis, a unique solution outside this subset (in which case the theory would be falsified, see Roberts & Pashler, 2000).

In the psychometric data modeling practice, the psychometric model is taken to be mispecified (e.g., Drasgow & Kanfer, 1985). Consequently, the idea according to which the psychometric model is a theory of a stochastic process of observed data generation is not seriously considered and the null hypothesis is almost unanimously disqualified (e.g., Hambleton, Swaminathan, & Rogers, 1991; van der Linden & Hambleton, 1997).

Ethics of psychological evaluation: social and instrumental objectivity and relativism

In practice, the people that use construct measurement in order to characterize individuals at a moment in time consider that observed scores are the best possible measures of a construct, knowing what we call (abusively, see Messick, 1995; Thompson & Vachaa-Haase, 2000) the tests' psychometric properties.

If psychometrics do not permit to give the status of scientific objects (Granger, 1995) to constructs, the testing community has to recognize their constructivist nature (Hacking, 1999). The fact that scientific psychology has not discovered a psychological quantity does not preclude that these quantities are imaginary, particularly if they have social significance. But engineering dedicated to the determination of positions that people take on these quantities at a moment in time has to be recognized as an engineering of avatar fabrication, i.e. representations of people in a virtual world.

This analysis leads us to compare the psychological evaluation to Internet gaming where players are represented by their avatars. The actors of the evaluation game subscribe implicitly to a social contract, in which the evaluated people accept to be represented by their avatars in a simplified universe that is adapted for objective decision making (i.e., impersonal and consensual). The rationality of the representation in such a universe is not psychological, but institutional, as the decision is optimized on a collective level.

The collective validity of test scores leads to the fact that the psychotechnical determination of individuals in total orders is a process that is relative to (i) a sample and (ii) to the

psychotechnical techniques (as a composition of applications of standardized description and numerical coding, Vautier et al., 2011). From an individual point of view, accepting this psychometric evaluation supposes the acceptation of the abandonment of the idea of psychological autonomy for a heteronomic identity, i.e. based on the relationship between the self and others.

Conclusion

In this article, we have shown that psychometric data modeling is not part of a nomothetical research program, as it does not contribute to the discovery of empirical laws that govern the psychotechnically observable phenomena. The contribution of psychometric modeling is fundamentally a political one, as it permits the assimilation of the reality of phenomena that are described in a qualitative way and can at best be partially ordered to a intuitively totally ordered reality, where the social utility rests on the need for comparison of human beings. If psychological research can find its legitimacy more in knowledge ethics than in social utilitarism (or practicalism, Michell, 1997), then psychometric modeling appears to be useless.

References

Bertrand, D., El Ahmadi, A., & Heuchenne, C. (2008). D'une échelle ordinale de Guttman à une échelle de rapports de Rasch. *Mathématiques et Sciences Humaines, 4,* 25-46.

Borsboom, D. (2003). *Conceptual issues in psychological measurement*. Amsterdam: Author.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinarity Research and Perspectives, 6,* 25-53.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2002). Functional thought experiments. *Synthese, 130,* 379-387.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203-219.

Brown, J. (1992). *The definition of a profession: The authority of metaphor in the History of intelligence testing, 1890-1930*. Princeton, NJ: Princeton University Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186-190.

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ: Lawrance Erlbaum Associates.

Cronbach, L. J., & Meehl, P. H. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Dagonet, F. (1993). *Réflexions sur la mesure*. Fougères, France: Encre marine.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. New York: Cambridge University Press.

Dooremalen, H., & Borsboom, D. (2010). Metaphors in psychological conceptualization and explanation. In A. Toomela & J. Valsiner (Eds.), *Methodological thinking in psychology: 60 years gone astray?* (pp. 121-144). Charlotte: Information Age Publishers.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662-680.

Essex, C., & Smythe, W. E. (1999). Between numbers and notions: A critique of psychological measurement. *Theory & Psychology, 9,* 739-767.

Feller, W. (1966). *An introduction to probability theory and its applications. Vol. I* (3d ed.). New York: Wiley.

Granger, G.-G. (1995). *La science et les sciences* (2 ed.). Paris: Presses Universitaires de France.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Hacking, I. (1999). *The social construction of what?* Cambridge, MA: Harvard University Press.

Haig, B. D., & Borsboom, D. (2008). On the conceptual foundations of psychological measurement. *Measurement: Interdisciplinarity Research and Perspectives, 6,* 1-6.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19,* 451-473.

Johnson, H. M. (1936). Pseudo-mathematics in the mental and social sciences. *American Journal of Psychology, 48,* 342-351.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA.: Addison-Wesley.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27,* 251-280.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8,* 435-462.

Martin, O. (1997). *La mesure de l'esprit : origines et développements de la psychométrie, 1900-1950*. Paris: L'Harmattan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355-383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10,* 639-667.

Michell, J. (2003a). Epistemology of measurement: The relevance of its history for quantification in the social sciences. *Social Science Information, 42,* 515-534.

Michell, J. (2003b). The quantitative imperative: Positivism, naïve realism, and the place of qualitative methods in psychology. *Theory & Psychology, 13,* 5-31.

Michell, J. (2004). The place of qualitative research in psychology. *Qualitative Research in Psychology, 1,* 307-319.

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

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinarity Research and Perspectives, 6,* 7-24.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus: User's guide* (4th ed.). Los Angeles, CA: Author.

Popper, K. R. (1959). *The logic of scientific discovery*. Oxford England: Basic Books.

Porter, T. M. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinarity Research and Perspectives, 1,* 241-255.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358-367.

Selby, S., & Sweet, L. (1963). *Sets relations functions*. New York: McGraw-Hill.

Steyer, R. (2001). Classical (psychometric) test theory. In T. Cook & C. Ragin (Eds.), *International encyclopedia of the social and behavioral sciences. Logic of inquiry and research design* (pp. 1955-1962). Oxford: Pergamon.

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika, 66,* 473-486.

Thompson, B., & Vachaa-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot hapen. *Theory & Psychology, 19,* 579-599.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Vautier, S. (2011a). How to state general qualitative facts in pychology? *Quality and Quantity,* 1-8.

Vautier, S. (2011b). The operationalisation of general hypotheses versus the discovery of empirical laws in Psychology. *Philosophia Scientiae, 15,* 105-122.

Vautier, S., Gillie, R., & Veldhuis, M. (2011). About validity of conclusions based on multiple linear regression: A commentary on Kupelian et al. (2010). *Preventive Medicine, 52,* 465.

Vautier, S., Hubert, L., & Veldhuis, M. (2011). *Why test scores may be improper data for scientific psychology: A qualitative analysis.* Manuscript submitted for publication.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage.

Appendix

Objective

Show that the expected value of a numerical random variable admitting a limit is comprised

between the inferior and the superior limit of the variable.

Notations and definitions

Let Y be a numerical random variable admitting a limit. It is an application from the set of events Ω

to a finite set of numerical values $(y_i)_n$.

The $(y_i)_n$ notation designates an ascending sequence of n numerical values

$$(y_i)_n = (y_0, y_1, \ldots, y_n).$$

The expected value of Y is defined if a probability law governs the realisation of events of $(y_i)_n$.

Thus we define the probability law $(p_i)_n = (p_0, p_1, \ldots, p_n)$ such that:

- for every i, p_i belongs to [0, 1],

- the sum of p_i equals 1.

The expected value of Y, noted $E(Y)$, is defined by:

$$E(Y) = p_0 \cdot y_0 + p_1 \cdot y_1 + \cdots + p_n \cdot y_n.$$

Show that $E(Y) \geq y_0$

The following property will be used: if the minimal value of the variable equals 0 then all the values

are positive and the expected value is positive as well, for all p_i are positive.

Let a real number k exists such that $y_0 + k = 0$. Suppose $Y' = Y + k$ defined in $(0, y'_1, \ldots, y'_n)$. $E(Y') \geq$

0 because all values in the defintion of Y' are positive.

$$E(Y') = p_0(y_0 + k) + p_1(y_1 + k) + \ldots + p_n(y_n + k) = k + E(Y),$$

thus $k + E(Y) \geq 0,$

implying $\qquad$ $E(Y) \geq -k.$

and $\qquad$ $-k = y_0.$ $\qquad$ QED.

Show that $E(Y) \leq y_n$

The following property will be used: if the maximal value of the variable equals 0 then all the

values are negative and the expected value is negative as well, for all p_i are positive.

Let a real number l exists such that $y_n + l = 0$. Suppose $Y'' = Y + l$ defined in $(y''_0, y''_1, ..., 0)$. $E(Y'')$

≤ 0 because all sums in the definition are negative.

$$E(Y'') = p_0(y_0 + l) + p_1(y_1 + l) + ... + p_n(y_n + l) = l + E(Y),$$

thus $\qquad$ $l + E(Y) \leq 0,$

implying $\qquad$ $E(Y) \leq -l.$

and $\qquad$ $-l = y_n.$ $\qquad$ QED.