Test Validation without Measurement: A Plea for Disentangling Scientific Explanation of Item Responses and Justification of Focused Assessment Policies Based on Test Data

Emilie Lacot, Mohammad Hassan Afzali, Stéphane Vautier

Test Validation without Measurement: A Plea for Disentangling Scientific Explanation of Item Responses and Justification of Focused Assessment Policies Based on Test Data

Emilie LACOT [a, b*], Mohammad H. AFZALI [b],

and Stéphane VAUTIER [b]

[a] Université de Picardie Jules Verne, UPJV, Centre de Recherche en Psychologie - Cognition, Psychisme, Organisation - EA 7273, Amiens, France.

[b] Université de Toulouse, UTM, Laboratoire OCTOGONE - EA 4156, Toulouse, France.

* Correspondence concerning this article should be addressed to Emilie Lacot, PhD candidate, Université de Picardie Jules Verne, CRP-CPO, Bâtiment E. Chemin du Thil, 80025 Amiens Cedex 1, France.

E-mail: emilie@lacot.net

Abstract

Test validation based on usual statistical analyses is paradoxical, as, from a falsificationist perspective, they do not test that test data are ordinal measurements, and, from the ethical perspective, they do not justify the use of test scores. This paper (i) proposes some basic definitions, where measurement is a special case of scientific explanation; starting from the examples of memory accuracy and suicidality as scored by two widely used clinical tests/questionnaires. Moreover it shows (ii) how to elicit the logic of the observable test events underlying the test scores, and (iii) how the measurability of the target theoretical quantities—memory accuracy and suicidality—can and should be tested at the respondent scale as opposed to the scale of aggregates of respondents. (iv) Criterion-related validity is revisited to stress that invoking the explanative power of test data should draw attention on counter-examples instead of statistical summarization. (v) Finally, it is argued that the justification of the use of test scores in specific settings should be part of the test validation task, because, as tests specialists, psychologists are responsible for proposing their tests for social uses.

Keywords: Assessment; Criterion-related Validity; Explanation; Falsifiability; Measurement; Prediction; Decision-making; Validity.

> *There is the greatest difference between presuming an opinion to be true, because, with every opportunity for contesting it, it has not been refuted, and assuming its truth for the purpose of not permitting its refutation.*
>
> John S. Mill

1. Introduction

Ziegler and Vautier (2014) have initiated a debate about test validation and psychological assessment in this Journal. The present article draws attention on the necessity to bypass the current status quo in the practice of test validation, by splitting up the test validation task into two kinds of tasks. The scientific task, aimed at *explanation* of test data, entails that if measurement turns to be a false explanation, test scores are not measurements, and comparability of testees' response patterns is not warranted. The sociotechnical task, aimed at justifying focused comparison policies based on test data, raises the issue of the final purposes comparison techniques are supposed to serve.

In current test validation practice, the justification of comparability of people's psychological traits is grounded on traits' measurability instead of social necessity to compare. Consequently, psychologists are committed to defend that test scores—composite scores or psychometric estimates of latent real numbers—are measurements, even by distorting the concept of measurement (see Michell, 1990, 1997, 1999, 2000, 2001, 2003, 2008b). Such an ideological imperative entails pervasive consequences on the epistemological and methodological norms that govern psychological scientific research. If what makes a psychological research a scientific contribution is its social efficiency or utility, and if, in test validation research, social efficiency is comparability—the universality of the more-than, or comparative language—, test validation researchers are committed not to highlight their state of *ignorance* with respect to the phenomena they can observe through their descriptive windows, which impinges on potential progress in the explanation of these phenomena. Psychometric analysis of test data is based on the probabilistic modeling of theoretical ordinal

measurement anomalies (Heene, 2013; Michell, 2008a), which enhances the researcher's statistical skills to the detriment of the theoretical and experimental skills that would be required to explain these anomalies (Krause, 2010; Vautier, Lacot, & Veldhuis, 2014) or to invent alternative explanations. The probabilistic approach is so deeply entrenched in mainstream methodology that the scientific value of a theoretical anomaly seems nearly irrelevant: a theoretical anomaly supposes a theoretical expectation that is formulated as an empirical impossibility (Vautier, 2011). But statisticism (Lamiell, 2013) formats us to be satisfied with statistical regularities, and prevent us to ask for falsifiability of psychological theories, as counter-examples are normal (see also Popper, 1992, p. 288). Consequently, we are accustomed not to ask for logical validity of our inferences, since it is normal that a proposition of form "if *a* then *b*" may be acceptable even if some non-*b*s are known. In some decision-making settings, statements of form "if *a* then *b most of time*", or of form "$\Pr(b|a) > \Pr(b|\neg a)$" may be useful, but this kind of usefulness has not to bound scientific research in psychology. As if *a* then *b* most of time, *a* is not a sufficient condition for *b*, and hence it does not explain *b*.

Section 2 proposes definitions of the key concepts of scientific explanation of observable phenomena, observable phenomena that are incomparable, *ordinal* measurement of theoretical, quantitative variations, which is a special case of the scientific explanation of observable comparable phenomena, by taking a falsificationist perspective (Popper, 1959). These definitions are simple and, although they do not preclude alternative definitions, they enable anyone to ask for the state of knowledge/ignorance in any well-defined testing field. As ordinal measurement is a necessary condition for quantitative measurement, it is useless to focus on quantitative measurement while ordinal measurement may easily be falsified. Section 3 applies these definitions to the description of phenomena that can be observed with two clinical tests, the Free and Cues Selective Reminding Test with Immediate Recall

(FCSRT-IR, Buschke, 1984; Grober & Buschke, 1987), and the version 5.00 of the Mini International Neuropsychiatric Interview's suicidality module (Lecrubier et al., 1997; Sheehan et al., 1997, 1998). The choice of these examples results from the substantive interests of the two first authors. What is at stake is to acknowledge that whatever the test we use, we do not observe scores but *m*-tuples of responses to *m* test items. Section 4 details the measurement issue and shows that the relevant test observations are not ordinal measurements if they are not comparable. Section 5 uses the notion of criterion-related validity to elaborate on the concept of scientific explanation and to stress that despite it is the intrinsic goal of scientific research, it offers no epistemic security since its rests on the invocation of universal truths that are not surely true. Section 6 sketches the issue of making socially acceptable or good decisions, when most of time no scientific explanation is available or relevant.

2. Definitions

We borrow the definitions from Vautier's academic blog[1] where working notes are open to public criticism.

2.1. Scientific Explanation

A scientific explanation of the phenomenon *p* consists in invoking either a sufficient condition *a*, such as if *a* then *p*, or a necessary condition *b*, such as if *p* then *b*. If one can establish *a*, one is able to cause *p*. If one can suppress *b*, one is able to prevent *p*. If *p* is not observable, explanative claims are not testable, and hence, according to the falsifiability criterion of demarcation (Popper, 1959), they are not scientific claims.

2.2. Observable and (In)comparable Test Phenomena

Psychological tests allow one to observe *m*-tuples of responses, where *m* is the number of items. For the sake of concision, we will restrict our analysis to tests comprised of

[1] Retrieved from http://epistemo.hypotheses.org/.

items associated with responses defined in a finite, discrete, and simply ordered set (a set is simply ordered by ≥ if any pair of two elements a and b is such that $a \geq b$ or $b \geq a$).

Notations. Let x_i: $\Omega \rightarrow D_i$ denote the descriptive function associated to the item i, which assigns to any observation unit ω from the set Ω one and only one value in the set D_i of descriptive values. Let $\mathbf{x} = x_1x_2\ldots x_m$: $\Omega \rightarrow \mathbf{D} = D_1 \times D_2 \times \ldots \times D_m$ denote the descriptive, conjoint test function—where the time laps needed to treat the items is neglected.

Comparability of m-tuples. $\mathbf{x}(\omega_1) \geq \mathbf{x}(\omega_2)$ if and only if for all is, $x_i(\omega_1) \geq x_i(\omega_2)$. If there are at least two items i and j such as $x_i(\omega_1) > x_i(\omega_2)$ and $x_j(\omega_1) < x_j(\omega_2)$, the m-tuples $\mathbf{x}(\omega_1)$ and $\mathbf{x}(\omega_2)$ are incomparable, as neither $\mathbf{x}(\omega_1) \geq \mathbf{x}(\omega_2)$ nor $\mathbf{x}(\omega_2) \geq \mathbf{x}(\omega_1)$. In this case the descriptive image $\mathbf{x}(\Omega)$ is not simply ordered but merely partially ordered by ≥.

2.3. Ordinal measurement of a theoretical, quantitative variation

Measurement. Observations (including responses to test items) *measure* a quantity if and only if their variability depends on the variation of the quantity, in such a way that it is possible to deduce from the observed variation that the quantity has increased or decreased.

Remark. If one admits that the observed variation depends on other factors, and if one ignores the extent to which these factors determine the observed variation, one can deduce from the observations neither an increase nor a decrease of the quantity, because one cannot exclude that this variation results from undesired perturbations when the quantity amount did not change. This is the fatal flaw of the psychometric approach to psychological measurement, since measurement error is unrestricted (Vautier et al., 2014; Vautier, Veldhuis, Lacot, & Matton, 2012).

The quantitative hypothesis. Measurement rests on the hypothesis that a quantity does exist—either one measures a quantity or one does not measure; saying that a process is measured is an example of misleading speaking. The quantitative hypothesis applied to a set Ω can be stated as a hypothetical function q: $\Omega \rightarrow [0, \max]$, which assigns to any observation

unit from Ω one and only one amount from the segment [0, max], where "max" denotes the highest possible amount of the quantity.

The ordinal measurement function. $\mathbf{f} = f_1 f_2 \ldots f_m$: [0, max] → $\mathbf{D} = D_1 \times D_2 \times \ldots \times D_m$ is the conjoint ordinal function, where for all *i*s, f_i: [0, max] → D_i is an increasing step function—if $x_1 > x_2$ then $f(x_1) \geq f(x_2)$, where $f(x)$ is an observed value in D_i and D_i is simply ordered. This definition entails that **f** is increasing (see comparability, 2.2).

The psychotechnical (not psychometric) measurement hypothesis. By definition of ordinal measurement, if a test allows ordinal measurement of amounts $q(\Omega)$ from [0, max], a function **f** exists such that the test observations $\mathbf{x}(\omega)$ are the images $\mathbf{f}[q(\omega)]$ of the observation units ωs. Let ω_1 and ω_2 be two observation units. Their ordinal measurements are respectively the observations $\mathbf{f}[q(\omega_1)]$ and $\mathbf{f}[q(\omega_2)]$. Consequently, as **f** is increasing, psychotechnical ordinal measurements are *comparable* multivariate descriptions; the image $\mathbf{f}[q(\Omega)]$ is simply ordered. Reciprocally, incomparable test descriptions (*m*-tuples) falsify the existence of a measurement function **f** or, in other words, falsify ordinal measurability of the amounts $q(\Omega)$ by the test descriptions.

3. The Descriptive Language of the Test: Two examples

3.1. The MINI's Suicidality Module

The MINI's suicidality module is composed of 12 items with a response format of *yes* or *no*. A typical item is "Think about suicide?" (C4). Figure 1 shows that the questionnaire contains the nested items C1a, C1b, and C4a. A nested item is an item that is ruled out of the test procedure in case of a negative answer to its preceding, parent item. For example, the negative answer to the parent item C1a: "In the past month, plan or intend to hurt yourself in that accident either passively or actively?" rules out the item C1b: "Did you intend to die as a result of this plan?".

A test response is a 12-tuple of item responses. The test score associated with the test

response is calculated on the basis of the items' weights (indicated in Figure 1). For example, endorsing the 12 items yields the 12-tuple (0, 0, 0, 1, 2, 6, 8, 8, 9, 4, 10, 4), which is associated with the test score of 52 points; rejecting the items C1, C1a and C4 (i.e., the parent items) but endorsing the others yields the 12-tuple (0, *m*, *m*, 1, 2, 0, *m*, 8, 9, 4, 10, 4), where '*m'* codes for the non-attendance of the nested items, and the associated test score is 38 points.

3.2. The FCSRT

The FCSRT encompasses two test tasks—free, and cued recall—following the encoding task of 16 items belonging to various semantic categories (e.g., fruit for grapes, flower for daisy, etc.). In the free recall task, the patient is asked to retrieve the names of the 16 items without cues ("In this phase, you must recall as many items belonging to the learning list as possible"). For a specific item *i*, the set of the observable events *at the free recall task* is {0, 1}, where 0 and 1 code for a retrieval failure and a retrieval success, respectively.

In the cued recall task, the patient is asked to retrieve the name of those items that were failed at the free recall task. The clinician takes the first failed item and provides the patient with the corresponding semantic cue (e.g., "what is the name of the flower?"), and so on until the last item that was not retrieved at the free recall task. The set of the observable events associated with a *remaining* item is {0, 1}, where 0 and 1 code for a retrieval failure, and a retrieval, respectively. A crucial feature of the *cued* recall task is that the opportunity to observe a failure at a given item depends on the previous observation of a failure at this item during the *free* recall task. In the case of success at the free recall probe, the cued recall cannot be probed.

To define the test events precisely, let us call an '*i*-probe' the test procedure corresponding to the item *i*. Figure 2 depicts the probe corresponding to the item *i*. The patient is instructed to freely retrieve *i*. If *i* is retrieved, the probe stops; the resulting empirical event

is coded '1*m*' ('*m*' for missing). If the patient does not retrieve *i*, cued recall is asked. As the cued recall can be succeeded or failed, the resulting empirical test events are '01' or '00', respectively. Hence, each *i*-probe is valued in the set of three possible elementary test events, which is {00, 01, 1*m*}.

The FCSRT comprises 48 probes as each *i*-probe is repeated three times, and $i = 1, 2, ..., 16$. In current practice, three types of test scores are used: the total free recall score counts the number of '1*m*'s from the three trials, the total recall score counts the number of '1*m*'s or '01's from the three trials, and the "cue efficiency" is defined as the ratio of cued recall successes (number of '01's) to the number of cued recall attempts (48 minus the number of '1*m*'s), and it is more rarely employed (e.g., Buschke, 1984; Grober, Buschke, Crystal, Bang, & Dresner, 1988; Grober, Lipton, Hall, & Crystal, 2000; Grober, Merling, Heimlich, & Lipton, 1997; Grober, Sanders, Hall, & Lipton, 2010; Sarazin et al., 2007).

4. Step Functions for Ordinal Measurement

The *Standards* (1999) focus on test *scores* instead of the underling test events: "… test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test" (p. 174). Thus, the FCSRT's total score of a patient is to be interpreted as indicating her/his standing on *memory accuracy*, the theoretical quantity to be measured. In the same vein, the suicidality module's score is to be interpreted as indicating the test taker's standing on *suicidality*, the theoretical quantity. If 'standing' refers to a point on the quantity, the test score is to be interpreted as an ordinal measurement of the quantity. The question is to explain the associated measurement theory that links the quantity to the possible test scores.

The first subsection demonstrates that suicidality and memory accuracy test scores can be viewed as ordinal measurements but that such theorizing is not testable. The second and third subsections examine the problem of incomparability of test responses (or events).

4.1. Linking Test Scores to Suicidality and Memory Accuracy, Respectively

Both kinds of quantities, suicidality and memory accuracy, refer to a theoretical segment if change of the quantity is continuous or to a union of theoretical points or segments if change is discontinuous. For the sake of simplicity it suffices to consider that the theoretical variations of interest are continuous. To determine the test taker's standing on the quantity consists in inferring her/his position on [0, max].

Suicidality and FCSRT total scores vary in a simply ordered scale ranging from 0 to 52, and 0 to 48 points, respectively. Ordinal information about the test taker's position on the segment [0, max] can be inferred validly from (i) the premise of the test score and (ii) a function that links the segment to the test score's scale. This function is a step function. Figure 3 illustrates the step functions for the ordinal measurement of suicidality (left side) and of memory accuracy (right side) through the possible test scores. The *x*-axis represents the theoretical quantity, viz., suicidality and memory accuracy, respectively, varying from 0 to a maximum, and the *y*-axis represents the test scores.

Let us detail how the step function works for the ordinal measurement of suicidality (Figure 3, left side). If suicidality fluctuates from 0 to threshold *A*, the score will be 0. If suicidality fluctuates from *A* to *B*, the score will be 1. If suicidality fluctuates from *B* to *C*, the score will be 2, and so on until to the 52rd threshold (*BA*). If suicidality fluctuates from *BA* to max, then the score will be 52. The thresholds' values are unknown, but they are ordered *by hypothesis*. Using the step function, one may infer *validly* that an empirical increase (the test score increased) means that suicidality increased; reciprocally, an empirical decrease means that suicidality decreased. In case of no empirical change one cannot deduce that suicidality did not change. Although the step function enables valid inference, it is tautological, i.e., not testable.

The same reasoning works for the FCSRT's step function (Figure 3, right side). If memory accuracy fluctuates from 0 to threshold *A*, the score will be 0 and so on until the last segment defined from *AV* to max, for which the score will be 48. Such a step function is also tautological. However, generalizing this conception to the test responses allows testability. We derive below the relevant consequences for couples of items from the suicidality module, and from the FCSRT.

4.2. Linking Suicidality with Responses to Items from the Suicidality Module

Let us consider the items C2 and C3 of the MINI's suicidality module. The possible test responses are 2-tuples defined in the set {00, 10, 02, 12}—which is partially but not simply ordered—, where '0' means rejection, and '1' or '2' mean endorsement (the differences refer to the item's weights of MINI's suicidality module). It is assumed that the two items measure the same quantity, and that no change occurs within the short time period of the test taking. Each item requires one specific threshold, denoted by *A* and *A'*, respectively. Consequently, there are three possible orderings of *A* and *A'*: (i) $A = A'$, (ii) $A < A'$ and (iii) $A > A'$. As depicted in Figure 4, if $A = A'$, 10 and 02 are precluded; if $A < A'$, 02 is precluded; if $A > A'$, 10 is precluded. Any precluded 2-tuple is, according to Popper's word, a *falsifier* of the relevant step function. The same reasoning will be applied to the FCSRT's items.

4.3. Linking Memory Accuracy with Responses to Items from the FCSRT

Any FCSRT's item is associated with the possible responses 00, 01 or 1*m*. By definition, memory accuracy is higher when the item is retrieved at the free recall task (1*m*) than when the item is retrieved at the cued recall task (01), and '01' denotes higher memory accuracy than '00' (no retrieval). Figure 5 displays the two thresholds *A* and *B* that are needed to define the item's step function. If memory accuracy fluctuates from 0 to *A*, the response will be '00'; if memory accuracy fluctuates from *A* to *B*, the response will be '01'; and if

memory accuracy varies in the interval [*B*, max], the response will be '1*m*'. Such a step function is tautological.

Now, let us consider two FCSRT's items. The possible test responses are the nine 4-tuples 0000, 0001, 001*m*, 0100, 0101, 011*m*, 1*m*00, 1*m*01, and 1*m*1*m*, which are partially but not simply ordered. It is assumed that all items measure the same quantity and that the amount of the quantity does not change during the test taking. As each item requires two specific thresholds (i.e., *A* and *B* for the first one, *A'* and *B'* for the second one), there are 13 possible orderings of *A*, *B*, *A'* and *B'*. As depicted in Figure 6, if $A < B < A' < B'$ (left side), the responses 0001, 0101, 001*m*, 011*m* are precluded; if $A' < B' < A < B$ (right side), 0100, 0101, 1*m*00, 1*m*01 are precluded. The reader can verify that each of the 13 orderings is testable.

4.4. Measurability Should Be Tested at the Respondent Scale

For interpreting a test response as a measurement to be a logically valid argument, the test response has to be logically compatible with a measurement function that represents the measurement process. The measurement process occurs when a specific respondent treats the test's items. Consequently, one has to think about what is happening *respondent by respondent*. For example, let us suppose that respondent *u* provides her responses to the suicidality items C2 and C3 according to the ordering $A < A'$ (Figure 4, middle panel). This does not entail that respondent *v* will provide his responses to the same items according to the same function. Even if his responses would obey the same ordering, it is not sure that his thresholds' values are the same than *u*'s thresholds.

Consequently, inter-individual comparison of test data viewed as measurements of a theoretical quantity requires not only that any respondent treats the test's items according to a step function but also that this step function does not depend on who treats the items. The last assumption seems implausible because the former presupposes that any test response depends only on the theoretical quantity to be measured, while the entire community of

psychometricians acknowledges that items' responses depend on a myriad of unknown factors (see remark in 2.3). A psychometric model is the formalization of 'measurement error' in such a way that the test datum, either a test response or a test score, results from a certain amount of the theoretical quantity *and* from a random component. The psychometric modeling of the random component is not restrictive (Vautier et al., 2012, 2014). Given a certain amount of the quantity to be measured, the modeling establishes that *any* test data is logically compatible with this amount. Consequently, valid inference from the test data to the theoretical quantity cannot be achieved by psychometric modeling. This is why those who maintain that test data can be interpreted as measurements thanks to psychometric modeling have to defend that valid inference from test data to the quantity to be measured is not mandatory (see Newton, 2012).

Efforts to validate the data from the FCSRT or the MINI's suicidality module have not been based on the assumption that these test data are ordinal measurements in the sense of corroborated evidence of hypothetical step functions. Actually, we know of no published paper, which addresses the issue of how to think of these data as scientific measurements. The default epistemological stance is that despite unrestricted measurement error, professional consensus allows one to act as if test data were measurements (for a discussion of science as consensus, see also Notturno, 2009).

In the falsificationist perspective, considering that test responses are ordinal measurements is a theoretical, testable, quantitative explanation of the test responses, where a quantitative variation such that one threshold is broken through is detectable by an observed variation in a set of simply ordered (comparable) m-tuples. As observed m-tuples are not simply ordered, this is a false explanation.

5. Do Test Responses Explain other Observable Phenomena? Criterion-related Validity and Conditional Distributions

Criterion-related validation consists in conditioning the values of a criterion on the test data. A contingency table holds the empirical knowledge provided by a validation sample—the empirical state of knowledge/ignorance. Let us consider the situation where what is at stake is to 'predict' the criterion's value for a new observation unit given its test datum and previous evidence.

5.1. Preliminary Remarks on Validity of 'Predictions'

We will not discuss the chronological meaning of the term 'prediction', but its validity as a conclusion of a possibly implicit argument. Two main epistemic situations can be distinguished with respect to the logical validity of the 'prediction', i.e., the observation unit's assignment of a criterion value. The information conveyed by the contingency table consists of the conditional proportions $p_{ij} = n_{ij}/n_{i\cdot}$, where n_{ij} and $n_{i\cdot}$ denote the frequency of the cell ij and the frequency of the row i, respectively. We first have to discard a special case of *lack of evidence*. Let c denote the number of criterion's values (the number of columns); if $n_{i\cdot} < c$, the proportions p_{ij}s are undefined (if $n_{i\cdot} = 0$) or statistically inconsistent (the conditional distribution of the criterion is chaotic from one sample to another one).

Let us consider that sufficient evidence is available ($n_{i\cdot} > c$), and a new case fell in the relevant row i. The goal consists in assigning a criterion's value to this new case, given that one ignores its current value. A necessary condition for the decision to be logically valid is that it follows from the premise of an empirical law—if a rule is hypothesized for a parent population from which the observed cases have been drawn, and that has not been falsified, it is called an empirical law. A necessary condition for such a law is the existence of at least one j from 1, 2, …, c such that $p_{ij} = 0$ (see Vautier, 2011; Vautier et al., 2014). Thus, on the basis of the fact that the new case's test data is i, and that the law is invoked, the *exclusion* of the value j is valid, and its soundness depends on the truth of the law. Here is the first epistemic

situation; a valid 'prediction' (in terms of excluding at least one criterion's value) warrants nothing, because no empirical law is surely true.

The second epistemic situation occurs if the conditional proportions p_{ij}s are strictly positive. Then, no valid 'prediction' can be derived at all from the available evidence. It is a fallacy to use the conditional distribution (the criterion's distribution of line *i*) to estimate the probability of the criterion's values for the new case at time *t*, because these probabilities are either 0 or 1. The fallacy is based on the confusion between what Hacking (1975) calls epistemic vs. objective probabilities (see also Harré, 2004). The justification of the use of empirical frequencies to decide the criterion's value under uncertainty is not a scientific, but a political (or social) issue.

5.2. Criterion-related Validation of MINI's Suicidality Module

According to Roaldset, Linaker and Bjørkly (2012), the suicidality module allows "a screen for the risk of suicidal and non-suicidal self-injury behaviors within the first year after discharge from an acute psychiatric ward" (p. 297). It is unclear what does "a screen for the risk of a given outcome" mean exactly. One can consider the available evidence and look for possible valid 'predictions'.

Roaldset et al.'s (2012) validation sample is composed of 307 patients, who were followed during 12 months after discharge from a psychiatric hospital. Before discharge, each patient was assessed with the MINI's suicidality module—the authors used six items and the scores range from 0 to 33 points. During the follow-up, the patients were asked if *yes* or *no* they had made acts of suicidal behaviors or non-suicidal self-injury behaviors (leading to another hospitalization). Then, the authors determined a cut-off point with sensitivity and specificity at 0.73 and 0.62, respectively (see their Table 4, p. 296).

The corresponding contingency table is reconstituted in Figure 7 (left side). The two conditional distributions preclude valid prediction: whatever the suicidality class (score < 6 or

≥ 6), the two criterion's values are possible ($p_{ij} \neq 0$). The proportion of "self-harm acts" can be calculated for each suicidality class, viz. $p_{11} = 17/(17+171) \approx .09$ if the suicidality score ranges from 0 to 5, and $p_{21} = 46/(46+73) \approx .39$ if the suicidality score ranges from 6 to 33 points. There is no logical relationship between the risk that a new patient commits self-harm acts next year and the available evidence. It is a convention that the risk depends on the suicidality class.

From the scientific viewpoint, the issue is whether some conditions exist such that self-harm acts are impossible or necessary in a given time interval. One can doubt that such conditions exist if one believes that self-harm acts result from free will. But it is an empirical issue to check this claim for each of the $2^6 = 64$ conditions generated by the descriptive device formed by the six suicidality items. Consequently, the scientific viewpoint raises the strategic question of the price to be paid to discover (maybe unlikely) empirical laws by using suicidality data. The descriptive language of test scores is not useful to conclude that there is no such law, because it aggregates distinct conditions (except for the extreme scores). But it may be useful if a given score i allows the observation that there is a j in $\{1, 2\}$ such that $p_{ij} = 0$ (with $n_{1.} >> 2$), as the granularity of description needs not be refined.

5.3. Criterion-related Validation of the FCSRT

Grober and Buschke's (1987) validation sample is interesting because it exhibits a case of $p_{ij} = 0$. The sample consists of 25 elderly persons judged to be demented (D) and 25 elderly judged to be non-demented ($\neg D$) respondents. The cut-off point of 43 points yielded sensitivity and specificity values of .96, and 1.00, respectively (the authors employed the total recall score.)

Figure 7 (right side) displays the corresponding contingency table, where $p_{12} = 0$, and $n_{1.} = 24 >> 2$. Thus, given a new case with a FCSRT score smaller than 43 points, is the 'prediction' that this case is a D valid? To be valid, the 'prediction' has to be derived from the

premise that any person whose score is less than 43 is a *D*. The law can be falsified by a single case, and it seems that the relevant scientific community does not claim such a law. Consequently, no 'prediction' based on this evidence is valid.

Grober and Buschke's (1987) criterion (i.e., *D* vs. ¬*D*) results from a "comprehensive evaluation—including history, physical and neurological examination" (p. 16). The objectivity of such a criterion is disputable because it seems clear that the descriptive language they used is far more complex than the "demented/non-demented" dichotomy. Sarazin et al. (2010) considered the criterion of left medial temporal lobe (MTL) volume, which is valued in mm^3. They reported a correlation of .43 between the total recall FCSRT score and the measured volumes, based on a sample of 35 participants.

This empirical knowledge alone does not allow valid prediction because the default in linear regression modeling is that the residuals are non-restricted and normally distributed. Hence, any FCSRT score allows any MTL volume. Useful empirical knowledge for valid 'prediction' would be one empirical law, that is, one statement that precludes certain volume's ranges conditionally to, at least, one FCSRT score.

The critical feature of such empirical investigation in neuroscience is that the samples are small, because of the practical difficulty to get observations. Hence, it seems that even if the descriptive conditions that are available do not convey robust statistical information, what is at stake for valid 'prediction' is to discover initial conditions that preclude at least one (range of) value(s) of the criterion of interest. What is needed is conjecturing about such conditions. Empirical cases are rare but they can be used to test these conjectures. For example, a broad conjecture is that any patient who fails at least 50% of the cued probes at the FCSRT has a MTL volume of at least $x\ mm^3$. If exceptions are found, they have to be explained, that is, other empirical laws have to be conjectured (Vautier et al., 2014).

6. Test Validation for Making Good Decisions: Psychologists Are Collectively Responsible for the Use of their Tests as Specialized Citizens

Viewing test data as measurements is not tenable if one uses a classical definition of measurement (see 2.3), and test scores are essentially numerical aggregation of multivariate, not simply ordered events. Test scores are ordinal measurements by convention, and test responses, i.e., *m*-tuples, are not ordinal measurements. It suffices to ask whether the observed *m*-tuples are comparable. Moreover, a detailed examination of criterion-related validation reveals that no resulting empirical knowledge can be used to justify new 'predictions' on the basis of logical validity. Even if the explanative efforts would suggest empirical laws and hence valid 'predictions', there were no security about their soundness. Test validation consists mainly in a social consensus to use test scores based on intensive reputed scientific work, but without specific consideration with respect to the purposed aims. Thus, does the psychologist's responsibility end with a validation study? In the affirmative, the psychologist provides merely a means to score aspects of people's behavior, given that the scientific meaning of the score of any single person is thin, and "that's all". In our opinion, such an attitude would be problematic, because the tests remain a powerful means for social selection, as it enables their users to *compare* people between them. We will develop fictive and simplistic assessment settings based on the suicidality and memory accuracy tests to illustrate how the comparability power may help social selection, which occurs in a social context that in turn deserves appraisal.

6.1. A Fictive Suicidality Assessment Policy for Minimizing Hospital Costs

Suppose that the test users (i.e., a collective of practitioners) have to decide on discharge from a psychiatric hospital for a patient asking for increased length of stay. What would be good vs. bad decisions? If the good for the patient is preventing her/him to commit self-harm acts, the possibility does exist whatever her/his suicidality condition. Should the test

user decide no discharge? The test user does not face a scientific problem, but an ethical problem. And, from the viewpoint of the hospital, this is a financial problem, as far as discharge means smaller costs.

Suppose that because of the financial situation, the annual policy consists in discharging 40% of patients who ask for increased length of stay. There is a selection problem: how to decide who to discharge? And there is the technical solution: propose a screening rule based on a suicidality scale. The fact that the scale has been validated by usual statistical analyses should not mask the ethical problem. In this story, the suicidality scale is more useful to the hospital facing the financial problem than to patients suffering from depression or other problems who are discharged despite their demand. The technical attitude that consists in validating the suicidality scale without apprehending the social context of the suicidality scale's use is questionable as soon as one realizes that the suicidality scale is not a measurement instrument.

6.2. A Fictive Memory Screening Policy for Optimizing Psychologists' Workload

The FCSRT has been promoted for identifying preclinical and early dementia patients with mild cognitive impairment (e.g., Ferris et al., 2006; Petersen, Smith, Ivnik, Kokmen, & Tangalos, 1994; Sarazin et al., 2007). However, the decision to be made in clinical practice is *yes* or *no*, the patient requires further investigations. What are good vs. bad decisions? If the good for the patient is detecting some troubles related to an elderly dementia, the possibility of elderly dementia exists whatever her/his FCSRT condition, as $n_{21} = 1$ (Figure 7, right side). This is not a scientific but a social issue: how to optimize the good for the patient?

Let us suppose a simplistic setting to illustrate the ethical side of psychological assessment. Considering that patients, whose scores are above a cut-off point, do not require further investigation; this enables clinicians to save time during their working day. Suppose that the number of patients increases as the clinician's available time stagnates: it will be

tempting to adapt the cut-off point in order to adapt the investigation policy to the available human resources. The scoring of the test performances offers a technical solution to a selection problem: given that some patients with non-optimal memory performances cannot be examined further altogether, who to select for further investigation? The comparability the scoring scale achieves allows one to find selection criteria in a flexible way to face workload's change. Does psychological assessment research and development serve to provide evaluative techniques the use of which is out of psychologists' collective scope of responsibility?

7. Conclusion

As long as test scores are viewed as measurements, the test validation task can be apprehended as a kind of scientific activity, where the word "scientific" means that the scientist's business consists in advancing rigorously and honestly collective knowledge of specific aspects of natural world. And the social utilization of this knowledge remains a distinct field of applications, the legitimacy of which is grounded on the measurability of various, natural, psychological quantities that would have been discovered. But if the scientific community of psychological researchers recognizes that test data are not measurements, the business of psychological assessment, which heavily requires the comparative language allowed by test scores, cannot be thought of as an inheritance of the scientific knowledge that has been accumulated since one and half century. Consequently, the measurement issue in psychology, and particularly in psychological testing, is primarily ideological.

This is an objective fact that ordinal measurement has not been achieved by the testing technique. We tried to delineate a way to manage the transition towards full acknowledgment of this fact, by splitting up the test validation task into a scientific task that needs the intellectual freedom of falsifying untenable claims, and a sociotechnical task focused on

specific assessment issues. Assessment psychologists should prepare to start from explicit social problems than can be addressed in their multiple, psychological, social, financial, technical, and ethical dimensions, instead of ritualized statistical analyses like factor analyses and other latent variable modeling techniques that presuppose wrongly that measurability is achieved. The pernicious import of these analytical rituals is that the unrestricted measurement error has been normalized. Suppose you want to check that you lost weight and the doctor gives you a bathroom scale whose measurement error is unrestricted: do you will trust the doctor? As long as psychologists accept to make inferences based on wishful thinking, they impinge on their own scientific culture, specifically the practice of a falsificationist methodology (Notturno, 2000; Popper, 1959).

There is room to develop testing research based on social problems, where what is at stake is to propose various kinds of social practices instead of mere tests. Aggregation methods, including psychometric estimation, may be useful to enable specific approaches to decision-making under uncertainty, and psychologists do not need to claim falsely that aggregation equates measurement. The social challenge is that the specific policies assessment psychologists' technicality is serving should be open to criticism and to improvement beyond the specific scope of scoring.

References

Buschke, H. (1984). Cued recall in amnesia. *Journal of Clinical Neuropsychology*, *6*(4), 433–440. doi:10.1080/01688638408401233

Ferris, S. H., Aisen, P. S., Cummings, J., Galasko, D., Salmon, D. P., Schneider, L., … Thal, L. J. (2006). ADCS prevention instrument project: Overview and initial results. *Alzheimer Disease and Associated Disorders*, *20*(4), S109–23. doi:10.1097/01.wad.0000213870.40300.21

Grober, E., & Buschke, H. (1987). Genuine memory deficits in dementia. *Developmental Neurosychology*, *3*(1), 13–36.

Grober, E., Buschke, H., Crystal, H., Bang, S., & Dresner, R. (1988). Screening for dementia by memory testing. *Neurology*, *38*(6), 900–900. doi:10.1212/WNL.38.6.900

Grober, E., Lipton, R. B., Hall, C., & Crystal, H. (2000). Memory impairment on free and cued selective reminding predicts dementia. *Neurology*, *54*(4), 827–832. doi:10.1212/WNL.54.4.827

Grober, E., Merling, A., Heimlich, T., & Lipton, R. B. (1997). Free and cued selective reminding and selective reminding in the elderly. *Journal of Clinical and Experimental Neuropsychology*, *19*(5), 643–54. doi:10.1080/01688639708403750

Grober, E., Sanders, A. E., Hall, C., & Lipton, R. B. (2010). Free and cued selective reminding identifies very mild dementia in primary care. *Alzheimer Disease and Associated Disorders*, *24*(3), 284–90. doi:10.1097/WAD.0b013e3181cfc78b

Hacking, I. (1975). *The emergence of probability*. London: Cambridge University Press.

Harré, R. (2004). Staking our claim for qualitative psychology as science. *Qualitative Research in Psychology*, *1*(1), 3–14. doi:10.1191/1478088704qp002oa

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in Psychology. *Frontiers in Psychology*, *4*(May), 246. doi:10.3389/fpsyg.2013.00246

Krause, M. S. (2010). Trying to discover sufficient condition causes. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(2), 59–70. doi:10.1027/1614-2241/a000007

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology*, *31*(1), 65–71. doi:10.1016/j.newideapsych.2011.02.009

Lecrubier, Y., Sheehan, D., Weiller, E., Amorim, P., Bonora, I., Harnett Sheehan, K., … Dunbar, G. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: Reliability and validity according to the CIDI. *European Psychiatry*, *12*(5), 224–231. doi:10.1016/S0924-9338(97)83296-8

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

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. doi:10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639–667. doi:10.1177/0959354300105004

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist*, *36*(3), 211–218. doi:10.1080/00050060108259657

Michell, J. (2003). The quantitative imperative: Positivism, naive realism and the place of qualitative methods in psychology. *Theory & Psychology*, *13*(1), 5–31. doi:10.1177/0959354303013001758

Michell, J. (2008a). Conjoint measurement and the rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*(1), 119–124. doi:10.1177/0959354307086926

Michell, J. (2008b). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1-2), 7–24. doi:10.1080/15366360802035489

Mill, J. S. (1999). *On liberty*. Peterborough: Canada: Broadview Press. (Original work published 1859).

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective*, *10*(1-2), 1–29. doi:10.1080/15366367.2012.669666

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

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

Petersen, R. C., Smith, G. E., Ivnik, R. J., Kokmen, E., & Tangalos, E. G. (1994). Memory function in very early Alzheimer's disease. *Neurology*, *44*(5), 867–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8190289

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books. (Original work published 1934).

Popper, K. R. (1992). *Realism and the aim of science: From the postscript to the logic of scientific discovery*. New York: Routledge.

Roaldset, J. O., Linaker, O. M., & Bjørkly, S. (2012). Predictive validity of the MINI suicidal scale for self-harm in acute psychiatry: A prospective study of the first year after

discharge. *Archives of Suicide Research : Official Journal of the International Academy for Suicide Research*, *16*(4), 287–302. doi:10.1080/13811118.2013.722052

Sarazin, M., Berr, C., De Rotrou, J., Fabrigoule, C., Pasquier, F., Legrain, S., … Dubois, B. (2007). Amnestic syndrome of the medial temporal type identifies prodromal AD: a longitudinal study. *Neurology*, *69*, 1859–1867. doi:10.1212/01.wnl.0000279336.36610.f7

Sarazin, M., Chauviré, V., Gerardin, E., Colliot, O., Kinkingnéhun, S., de Souza, L. C., … Dubois, B. (2010). The amnestic syndrome of hippocampal type in Alzheimer's disease: An MRI study. *Journal of Alzheimer's Disease : JAD*, *22*(1), 285–94. doi:10.3233/JAD-2010-091150

Sheehan, D. V, Lecrubier, Y., Harnett Sheehan, K., Janavs, J., Weiller, E., Keskiner, A., … Dunbar, G. C. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry*, *12*(5), 232–241. doi:10.1016/S0924-9338(97)83297-X

Sheehan, D. V, Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59 Suppl 2*, 22–33;quiz 34–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9881538

Vautier, S. (2011). The operationalization of general hypotheses versus the discovery of empirical laws in Psychology. *Philosophia Scientiae*, *15*(2), 105–122. doi: 10.4000/philosophiascientiae.656

Vautier, S., Lacot, E., & Veldhuis, M. (2014). Puzzle-solving in psychology: The neo-Galtonian vs. nomothetic research focuses. *New Ideas in Psychology*, *33*, 46–53. doi:10.1016/j.newideapsych.2013.10.002

Vautier, S., Veldhuis, M., Lacot, E., & Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars. *Theory & Psychology*, *22*(6), 810–822. doi:10.1177/0959354312450093

Ziegler, M., & Vautier, S. (2014). A Farewell, a welcome, and an unusual exchange. *European Journal of Psychological Assessment*, *30*(2), 81–85. doi:10.1027/1015-5759/a000203

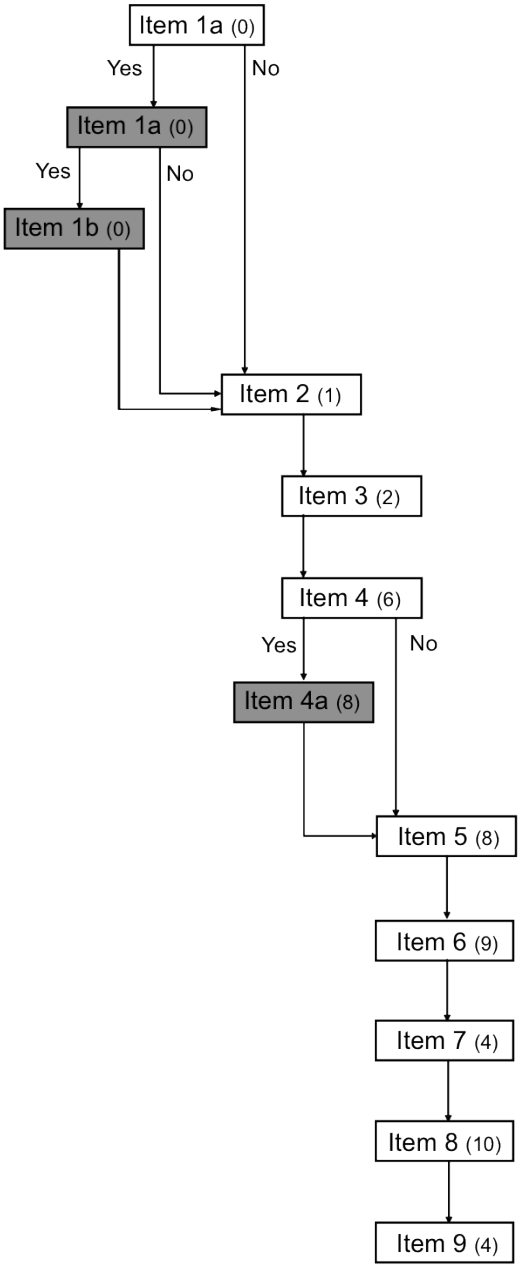

Figure 1. Structure of the MINI's suicidality module. The specific weight of each item for the score calculus is displayed in parenthesis.

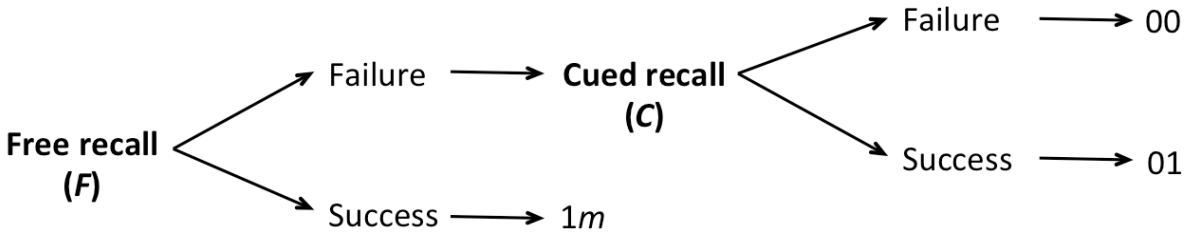

Figure 2. The empirical structure of one *i*-probe in the FCSRT comprises a free recall task (*F*) and a cued recall task (*C*). The output associated with an item *i* belongs to the set {00, 01, 1*m*}.

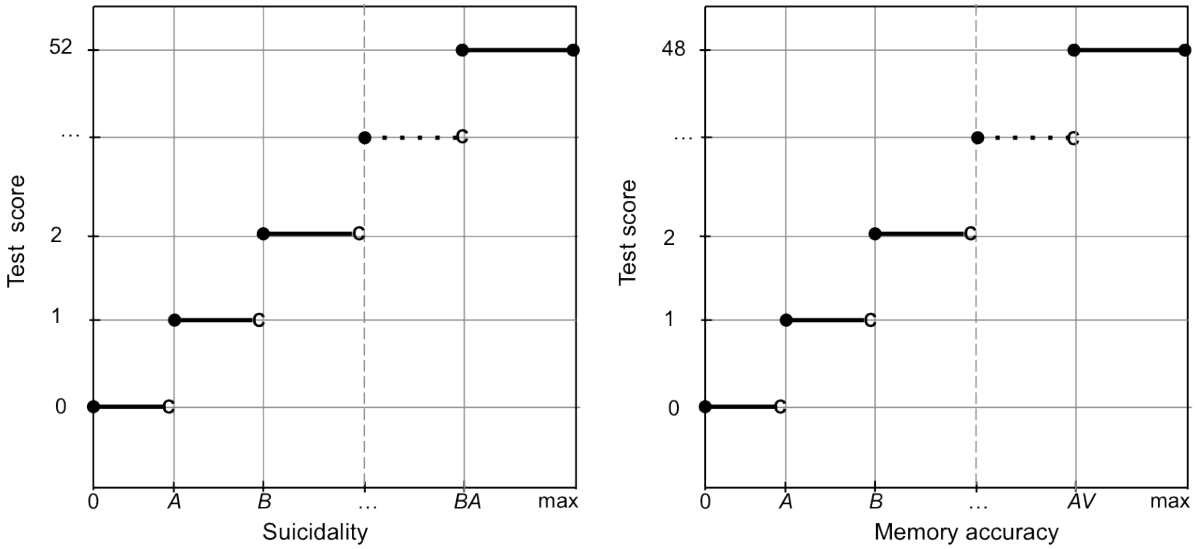

Figure 3. The step functions linking the suicidality (left side) and memory accuracy (right side) of a respondent with the possible test scores obtained at the MINI's suicidality module and at the FCSRT. The thresholds *A*, *B*, etc. follow a lexicographic order: *BA* codes for the 52rd threshold on the suicidality, *AV* code for the 48th threshold on the memory accuracy.

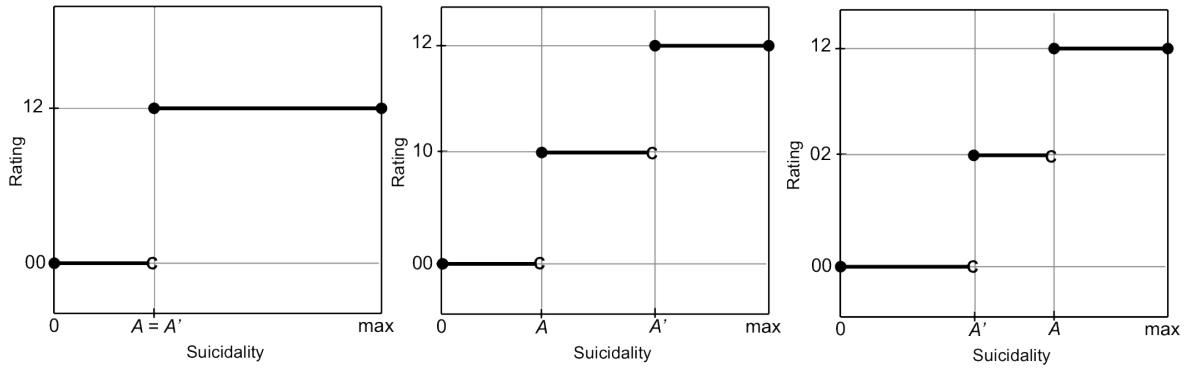

Figure 4. Three step functions linking suicidality to responses generated by the possible ordering of the thresholds associated with the items C2 and C3 of the MINI's suicidality module.

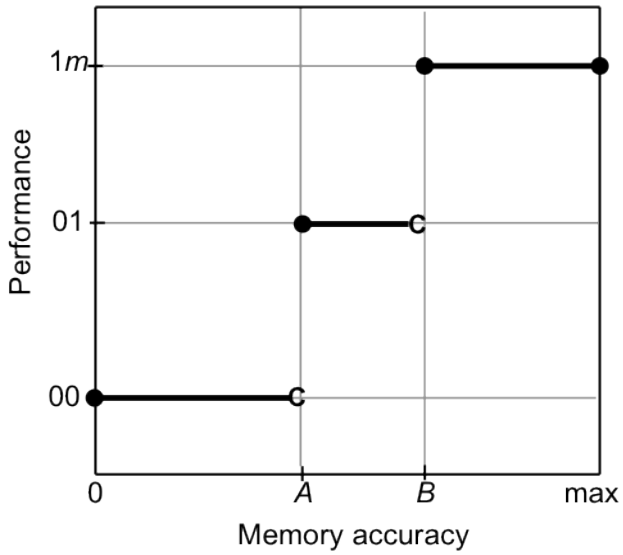

Figure 5. A step function linking memory accuracy of a respondent with the three possible responses to one FCSRT's item.

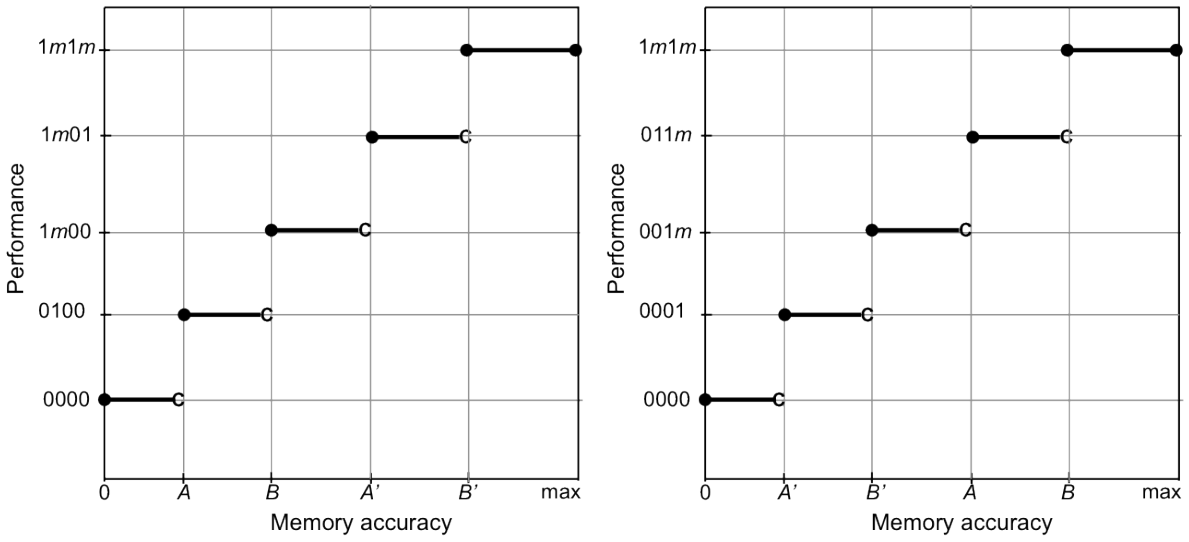

Figure 6. Two step functions linking memory accuracy to the responses to two FCSRT's items.

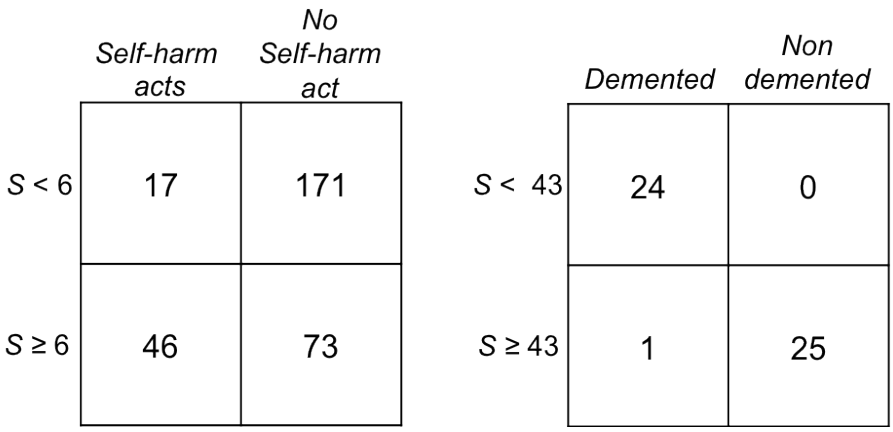

Figure 7. MINI's suicidality module (left side) and FCSRT (right side) contingency tables. Each cell displays a conjoint frequency for the validation sample.