Foundations of a Weak Measurement-Theoretic Approach to Software Measurement

Sandro Morasca

Dipartimento di Scienze Chimiche, Fisiche e Matematiche
Università degli Studi dell'Insubria, Via Valleggio 11
I-22100, Como, Italy
sandro.morasca@uninsubria.it

Abstract. Measurement Theory has been proposed as an approach for providing software measurement with a sound basis. However, Measurement Theory has hardly ever been used for building new measures. Much more often, it has been used to analyze the properties of existing measures, but there has been little agreement on the compliance of existing measures with Measurement Theory's strict requirements. This paper introduces a modified version of Measurement Theory, called Weak Measurement Theory. Because it has weaker requirements than Measurement Theory, Weak Measurement Theory may be better suited for the needs of the state of the art of software measurement. We provide an extension of the theory of the levels of measurement and we focus on ordinal scales and extensive measurement for Weak Measurement Theory. In addition, we show how Weak Measurement Theory can be used for widening the application scope of mean values.

1 Introduction

Theoretical validation is a fundamental step that should be carried out when a measure is proposed to show (or at least provide compelling evidence) that the measure really quantifies what it purports to measure. Especially in the early years of software measurement, measures have often been defined without adequate theoretical validation, if any. The measures themselves were operational definitions of the attributes to quantify. Thus, a large number of measures were defined, but little agreement existed as to which ones really captured the software attributes they intended to measure, as little agreement existed on the properties of software attributes and of their measures.

This lack of agreement is mainly due to two reasons. First, software engineering is a relatively recent discipline, if compared to more consolidated scientific and engineering disciplines. Second, software is not a physical object, so its attributes are more "elusive" than the attributes of physical objects encounteed in everyday life. Measurement Theory (MT, for short) [3, 5], originally defined for measurement in the Social Sciences, has been proposed as an approach to making the properties of software attributes and measures explicit. MT uses a full-fledged mathematical approach and thus allows software measurers to state the properties of software attributes unambi

M. Pezzè (Ed.): FASE 2003, LNCS 2621, pp. 200–215, 2003.

guously. This should facilitate (1) discussing properties that each software attribute should have, (2) reaching a consensus on the properties of each software attribute, and (3) defining software measures that are consistent with intuition.

Unfortunately, MT has hardly ever been used for defining new software measures, or even for stating the properties of software attributes. To the best of this author's knowledge, no measurement-theoretic proposals exist for important software attributes such as cohesion and coupling. It is true that MT proposals exist for the properties of software complexity and size. The proposals for software complexity have been widely discussed, but the software measurement community is far from an agreement. The proposals for software size are actually inspired by the properties of the size of physical objects. As a result, MT is often viewed as a rather theoretical approach whose benefits to software measurement are quite limited.

Also, MT's requirements may be too strict for the current state of the art of the software discipline; this is probably why MT has not produced many results in software measurement or played a more important role in it. This paper provides the foundations for modifying MT to make its requirements weaker and probably better applicable to the measurement of software attributes. Thus, we introduce a modified version of MT that we call Weak Measurement Theory (WMT, for short). In addition, we provide a weakened characterization of the levels of measurement for WMT and we focus on ordinal scales and extensive measurement. Finally, we show that the mean value of a distribution can be meaningfully used even for measures defined according to WMT and ordinal measures defined according to MT.

A weaker approach such as WMT may allow us to rigorously model several software attributes in a way that does not — at least currently — fit in the framework of MT. Thus, WMT may provide several existing, currently used measures with sounder mathematical bases than they have and show their properties. Especially for elusive attributes as the ones encountered in software engineering, the definition of a measure may require several refinement steps. Modeling the properties of a software engineering attribute is hardly ever a straightforward process, in which the "right" properties of the attribute are known from the start, but a number of iterations may be required. Starting from an initial set of properties of an attribute, one or more measures may be defined. The characteristics of these measures (e.g., the ordering of the entities that these measures provide) may be used to refine the initial properties of the attribute. In turn, this leads to the definition of new measures or the refinement of existing ones. The iterations continue until a satisfactory set of properties for the software engineering attribute and a satisfactory set of measures are obtained. For many software engineering attributes, a satisfactory set of properties has not been reached yet. Thus, we propose WMT to support this refinement process so we do not reject measures that may be useful, but need refining. In addition, for some software attributes, it is unclear whether a satisfactory set of properties will eventually be obtained. WMT may be used in those cases as well.

Weakening the requirements of MT also comes with a price. WMT is less strict than MT, so there may be a loss of accuracy in the information provided by a measure that complies with our modified version. On the other hand, we obtain a theory that is more easily and naturally applicable to software measurement.

The remainder of the paper is organized as follows. Section 2 provides a critical analysis of MT. Section 3 contains the proposal of WMT and its levels of measurement. Section 4 shows two existence theorems for ordinal and extensive measurement. Section 5 show how the notion of mean can be extended. Conclusions and future research directions are in Sect. 6.

2 An Analysis of Measurement Theory

We provide an introduction to MT in Section 2.1 and point out some of its characteristics that may have hindered its application in software measurement (Section 2.2).

2.1 An Introduction to Measurement Theory

Representational Measurement Theory [3, 5] separates the "intuitive," empirical knowledge on a specified attribute of a specified set of entities, captured via the so-called Empirical Relational System (Definition 1), and the "quantitative," numerical knowledge about the attribute, captured via the so-called Numerical Relational System (Definition 2). A scale (Definition 5) maps the Empirical Relational System into the Numerical Relational System in such a way that no inconsistencies are possible.

Definition 1. Empirical Relational System. Given an attribute, let

- E denote the set of entities for which we would like to measure the attribute,
- R_1, ..., R_n denote n empirical relations capturing our intuitive knowledge on the attribute: each R_i has an *arity* n_i, so $R_i \subseteq E^{n_i}$ (for notational convenience, we write either $<e_1,..,e_{n_i}> \in R_i$ or $R_i(e_1,..,e_{n_i})$ to denote that tuple $<e_1,..,e_{n_i}>$ is in relation R_i),
- o_1, ..., o_m denote m empirical binary operations on the entities that describe how the combination of two entities yields another entity, so each o_j has the form o_j: $E \times E \rightarrow E$; these operations are represented with an infix notation, e.g., $e_3 = e_1\ o_j\ e_2$.

An Empirical Relational System is an ordered tuple $ERS=<E,R_1,...,R_n,o_1,...,o_m>$.

For example, suppose we study the control-flow complexity (attribute) of program segments (set of entities), one of the most studied attributes. We characterize it via the empirical binary relation *more_complex_than* $\subseteq E \times E$, i.e., we have $<e_1,e_2> \in$ *more_complex_than* if we believe that e_1 is more complex than e_2. The operation $\oplus$ may be the concatenation operation, i.e., $e_3 = e_1 \oplus e_2$.

The Empirical Relational System does not make use of numbers or any kind of measurement values, which are introduced by the Numerical Relational System.

Definition 2. Numerical Relational System. Let

- V denote a set of values,
- S_1, ..., S_n be n relations on the values, each S_i has the same *arity* of R_i, so $S_i \subseteq V^{n_i}$,

- $\bullet_1$, ..., $\bullet_m$ denote m numerical binary operations: each operation $\bullet_j$ is of the form $\bullet_j$: $V \times V \rightarrow V$, and is usually represented with an infix notation, e.g., $v_3 = v_1 \bullet_j v_2$.

A Numerical Relational System is an ordered tuple $NRS = <V, S_1, ..., S_n, \bullet_1, ..., \bullet_m>$.

In our program complexity example, V may be the set of nonnegative integer numbers, a binary relation for may be >, and a binary operation may be — *for instance* — the sum between two integer values. The Numerical Relational System in itself does not provide any information about the entities and the attribute.

The Empirical Relational System and the Numerical Relational System are linked via a measure (Definition 3), which associates entities and values, and a scale (Definition 5), which associates the elements of the tuple of the Empirical Relational System with elements of the Numerical Relational System.

Definition 3. Measure. A measure is a function $m: E \rightarrow V$.

A measure is only a correspondence between entities and values, but does not take into account the empirical knowledge on the attribute (expressed through the empirical relational system) and how it is translated into numerical knowledge (expressed through the numerical relational system). Thus, not all measures built according to Definition 3 are sensible ones. For instance, given three program segments e_1, e_2, e_3 such that $<e_1,e_2> \in$ *more_complex_than* and $<e_2,e_3> \in$ *more_complex_than*, a measure m may be such that $m(e_1) > m(e_2)$ and $m(e_3) > m(e_2)$. To quantify an attribute sensibly, a measure must be consistent with the empirical knowledge about the attribute, i.e., a measure must satisfy the following representation condition.

Definition 4. Representation Condition. A measure must satisfy the two conditions

$$\forall i \in 1..n \; \forall <e_1, .., e_{n_i}> \in E^{n_i} \; (<e_1, .., e_{n_i}> \in R_i \Leftrightarrow <m(e_1), .., m(e_{n_i})> \in S_i) \qquad (1)$$

$$\forall j \in 1..m \; \forall <e_1, e_2> \in E \times E \; (m(e_1 \; o_j \; e_2) = m(e_1) \bullet_j m(e_2))$$

In our complexity-related example, the Representation Condition states that, $<e_1,e_2> \in$ *more_complex_than* for any two program segments e_1, e_2 iff $m(e_1) > m(e_2)$ and $m(e_1 \oplus e_2) = m(e_1) + m(e_2)$. The above definitions lead to the concept of a scale.

Definition 5. Scale. A scale is a triple $<ERS, NRS, m>$, where *ERS* is an Empirical Relational System, *NRS* is a Numerical Relational System, and m is a measure that satisfies the Representation Condition.

In what follows, we assume that measures satisfy the Representation Condition.

Given an Empirical Relational System and a Numerical Relational System, two issues arise, i.e., existence and uniqueness of a scale that links them. We will not deal with the existence problem in this section. As for uniqueness, more than one legitimate scale may be built. This is a well known fact in everyday life, in which one may measure the weight of objects in kilograms, grams, pounds, ounces, etc. At any rate, some properties of scales are invariant, called meaningful statements.

Definition 6. Meaningful Statement. A statement is said to be meaningful if its truth value does not change when a scale is replaced by another scale. Formally, if $S(m)$ is a statement based on measure m and $S(m')$ is the same statement obtained by replacing m with m', we have $S(m) \Leftrightarrow S(m')$.

Meaningful statements provide the real information content of a scale. For instance, it makes sense to say that an object is twice as heavy as another, regardless of the scale used (kilograms, pounds, etc.). If this statement is true with one scale, it is also true with all other scales. Instead, suppose we can tell if a software failure is more severe than another and we can classify failures with a 4-value severity scale with values 1 (least severe), 2, 3, 4 (most severe). It does not make sense to say that severity 2 failures are twice as severe as severity 1 failures, as the truth value of this statement depends on the specific choice of values. The truth value of the statement changes with another scale with values, say, 2, 15, 34, 981.

As the weight example shows, it may be possible to map one scale into another. In the weight example, we can map one scale into another by multiplication by a suitable constant, i.e., any proportional transformation provides a legitimate scale with a different weight unit. This leads to the definition of admissible transformation.

Definition 7. Admissible Transformation. Given a scale $<ERS,NRS,m>$, the transformation of scale f is admissible if $m'=f \circ m$ (i.e., m' is the composition of f and m) and $<ERS,NRS,m'>$ is a scale.

An admissible transformation may not always exists. For instance (see [5]), let $ERS = <E,R>$, with $E = \{r,s,t\}$ and $R = \{<r,s>,<r,t>\}$, and let $FRS = <Real, >_1 >$, where *Real* is the set of real numbers and $>_1$ and $>_1$ is a binary relation on *Real* such that $x >_1 y$ if and only if $x > y+1$. Let m' be such that $m'(r) = 2$, $m'(s) = 0$, $m'(t) = 0$, and m'' such that $m''(r) = 2$, $m''(s) = 0.1$, $m''(t) = 0$. Both m' and m'' are legitimate scales, but there is no admissible transformation that transforms m' into m''. Scales of this kind do exist in MT, where they are called irregular scales.

The scales for which admissible transformations exist may be classified according to the set of admissible transformations they can undergo, since a specific set of admissible transformations entails a specific set of meaningful statements. Next we describe some basic scale types of MT (though other types exist as well).

Nominal Scales. The values of these scales are labels for categories in which the entities are classified, with *no notion of order* among the categories, i.e., the set of entities is partitioned in a set of equivalence classes, each associated with a value of the measure. These scales can be transformed into other scales through one-to-one transformations. The invariant property is that two entities belong to the same equivalence class according to some measure m if and only if they belong to the same equivalence class according to some other measure m'. An example of a nominal scale may be the programming language used to write a program.

Ordinal Scales. The values of these scales are labels of categories in which the entities are classified, with a *total ordering* across categories. These scales are transformed into other scales via strictly monotonic transformations, which preserve the invariant property of the ordering. An example of an ordinal scale is failure severity.

Interval Scales. Each entity is associated with a numerical value, so that we quantify the difference between values. Strictly speaking, given a scale m, a new scale m' can be obtained only through transformations of the kind $m'=am+b$, with $a>0$, i.e., we can change the origin of the values (by changing b) and the unit measure (by changing a). Given four entities e_1, e_2, e_3, e_4, the invariant statements are of the kind $(m(e_1) - m(e_2))/(m(e_3) - m(e_4))$: the ratio of the lengths of two intervals is invariant. An interval scale may be the date of a milestone of a software project.

Ratio Scales. Each entity is associated with a numerical value that allows us to provide a quantification for the ratio between values. Given a scale m, a new scale m' can be obtained only through transformations of the kind $m'=am$, with $a>0$, i.e., we can change only the unit measure by changing a. Given two entities e_1, e_2 the invariant statements are of the kind $m(e_1)/m(e_2)$, i.e., the ratio of the values is invariant. An example of a ratio scale may be LOC as a measure of the size of a program.

The larger the set of admissible transformations, the less "precise" the scale. To measure the control-flow complexity of program segments, a nominal scale would allow us to classify segments into different classes, but not to order them. An ordinal scale would allow us to order the classes of the program segments, but not to measure the "distance" in control-flow complexity. An interval scale would allow us to measure the "distance" in control flow complexity among the program segments, but not the ratios between control-flow complexities; a ratio scale would allow us to measure the ratios between control-flow complexities. In practical use, there is a divide between the fully numerical scales (interval, and ratio ones) and the scales whose values may not necessarily be numerical (nominal and ordinal ones).

2.2 Issues in Applying MT in Software Engineering Measurement

Representational Measurement Theory is a consistent theory that has been developed in the last decades to give a solid mathematical basis to measurement for the social sciences. Here, we show some problems that may arise when applying MT in software engineering measurement. We argue that MT may be too demanding at this stage of technology, so a weaker approach may be more appropriate.

The problem of finding a measure that satisfies the Representation Condition (even when the Representation Condition is not fully explicitly stated) has been largely debated for a number of software attributes (e.g., size, complexity, cohesion, coupling, ...) in several different ways in the literature. The main point of most of these discussions is that, for virtually any measure, a "counterexample" can be found that questions the measure's consistency with intuition, i.e., more rigorously, its satisfaction of the Representation Condition.

The first part of the Representation Condition (marked with (1)) requires that $\forall i \in 1..n \; \forall <e_1, .., e_{ni}> \in E^{ni} \; (<e_1, .., e_{ni}> \in R_i \Leftrightarrow <m(e_1), .., m(e_{ni})> \in S_i)$. Let us explain this condition through our on-going example about control-flow complexity. Suppose that we have chosen m as a measure of control-flow complexity. The first part of the Rep-

resentation Condition says that, for any two program segments e_1, e_2, we have $<e_1,e_2> \in$ *more_complex_than* if and only if $m(e_1) > m(e_2)$.

Suppose we use the cyclomatic number as a measure of the control flow complexity of software code. Since the cyclomatic number of a program unit can be computed as the number of decision points in the software unit plus one (where an n-way decision point contributes as (n-1) two-way decision points), using the cyclomatic number as a software complexity scale implies that we rate as more complex a program unit *C10* with a single `case` statement with 10 exits than a program unit *W8* with 8 nested `while` loops. It is unclear if a wide consensus can be found on this ordering. If we believe that program unit *W8* is more complex than program unit *C10* we cannot use the cyclomatic number to measure control flow complexity.

This problem may not be due to a poor choice of a measure for a software attribute. Fenton [2] argues that no ordinal control-flow complexity scales exist because this would first imply finding a strict weak order (Definition 8) among the control-flow graphs according to their complexity and then finding a measure. The same situation is likely to occur for a number of software engineering attributes. Since we cannot find a strict weak order among the entities, we cannot define measures. We now provide the notion of strict weak order [5].

Definition 8. Strict Weak Order. A strict weak order is a pair $<E,R>$, where E is a set of entities and $R \subseteq E \times E$ is a binary relation that is

- Asymmetric: $\forall\ e_1,e_2 \in E\ R(e_1,e_2) \Rightarrow \neg R(e_2,e_1)$
- Negatively transitive: $\forall\ e_1,e_2,e_3 \in E\ \neg R(e_1,e_2) \wedge \neg R(e_2,e_3) \Rightarrow \neg R(e_1,e_3)$ or, equivalently, $\forall\ e_1,e_2,e_3 \in E\ R(e_1,e_2) \Rightarrow R(e_1,e_3) \vee R(e_2,e_3)$.

Strict weak orders can also be characterized through Theorem 1 [5], which requires the following notion of indifference relation.

Definition 9. Indifference Relation. Given a pair $<E,R>$, where E is a set of entities and $R \subseteq E \times E$, the relation $I \subseteq E \times E$ such that $I_R(e_1,e_2) \Leftrightarrow (\neg R(e_1,e_2) \wedge \neg R(e_2,e_1))$ is said to be the indifference relation derived from R.

Theorem 1. Strict Weak Order. A pair $<E,R>$, where E is a set of entities and $R \subseteq E \times E$, is a strict weak order if and only if R is

1. Asymmetric: i.e., $\forall\ e_1,e_2 \in E\ R(e_1,e_2) \Rightarrow \neg R(e_2,e_1)$
2. Transitive: i.e., $\forall e_1,e_2,e_3 \in E\ R(e_1,e_2) \wedge R(e_2,e_3) \Rightarrow R(e_1,e_3)$
3. and the indifference relation I_R is an equivalence relation.

Theorem 1 (from [5]) shows that the entities in a Strict Weak Order can be organized in a totally ordered set of equivalence classes, each of which contains elements that cannot be ordered. Not every partial order is a strict weak order. For instance, given $E = \{e_1,e_2,e_3,e_4\}$ and $R = \{<e_1,e_2>,<e_1,e_4>,<e_2,e_4>,<e_3,e_4>\}$, relation R is asymmetric and transitive, but it does not generate a relation I_R that is an equivalence relation as the pairs $<e_1,e_3>$ and $<e_2,e_3>$ belong to I_R, but $<e_1,e_2>$ does not.

It can be shown theoretically [5] that an ordinal scale exists iff the Empirical Relational System is a strict weak order. Fenton [2] contends that it would be impossible to find a sufficiently wide consensus on any negatively transitive order among control-flow graphs, so no ordinal control-flow complexity scales can possibly be defined. With reference to the control-flow graphs in Figure 1, it is argued that program segment *y* may be plausibly ranked as more complex than program segment *x*, but other pairs, such as *x* and *z* or *y* and *z* look incomparable. So, it would be impossible to find a wide consensus on a strict weak among these control-flow graphs.

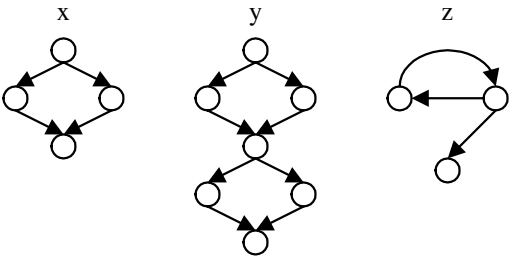

Fig. 1. Complexity ordering among control-flow graphs? (from [2])

Summarizing, we can provide an intuitive order for some of the pairs of entities, but not for all of them. In this situation, MT allows only for the existence of nominal measures. However, this implies a loss of information, since we have lost the piece of information about the ordering of the entities that we are actually able to order.

This is due to the nature of condition (1) of the Representation Condition, which states a property of the kind "if and only if," i.e., a double link is required between the Empirical Relational System and the Numerical Relational System. This double link may be too demanding for the current state of the art of software engineering measurement. It should be studied how this double link could be relaxed and what can be gained and lost by doing so, as we explain next.

3 Weak Measurement Theory

We propose a modified form of Measurement Theory, which we call Weak Measurement Theory (WMT for short). The difference between MT and WMT is that we do not require a double link between the Empirical Relational System and the Numerical Relational System in the Representation Condition, as follows.

Definition 10. Weak Representation Condition. A measure must satisfy the following two conditions

$$\forall i \in 1..n \; \forall <e_1, .., e_{n_i}> \in E^{n_i} \; (<e_1, .., e_{n_i}> \in R_i \Rightarrow <m(e_1), .., m(e_{n_i})> \in S_i) \quad (2)$$

$$\forall j \in 1..m \; \forall <e_1, e_2> \in E \times E \; (m(e_1 \; o_j \; e_2) = m(e_1) \bullet_j m(e_2))$$

This definition leads to the definition of weak scale.

Definition 11. Weak Scale. A weak scale is a triple $<ERS,NRS,m>$, where *ERS* is an Empirical Relational System, *NRS* is a Numerical Relational System, and *m* is a measure that satisfies the Weak Representation Condition.

In our complexity-related example, the Weak Representation Condition states that, for any two program segments e_1, e_2, whenever $<e_1,e_2> \in$ *more_complex_than* we have $m(e_1) > m(e_2)$ and $m(e_1 \oplus e_2) = m(e_1) + m(e_2)$. Since we have lost the implication from the Numerical Relational System to the Empirical Relational System, we also allow cases in which $<m(e_1), .., m(e_{n_i})> \in S_i$, but not $<e_1, .., e_{n_i}> \in R_i$, i.e., "false positives" are possible. In our complexity-related example, this means that we may have $m(e_1) > m(e_2)$, but not $<e_1,e_2> \in$ *more_complex_than*. We need to point out, however, that $<e_1,e_2> \notin$ *more_complex_than* does not imply $<e_2,e_1> \in$ *more_complex_than*. Instead, we do not have sufficient knowledge or we have not reached a sufficiently broad consensus on how to order e_1 and e_2 based on their complexity. As a proof of this statement, if we know that $<e_2,e_1> \in$ *more_complex_than*, then we must have $m(e_2) > m(e_1)$, due to condition (2) of the Weak Representation Condition and not $m(e_1) > m(e_2)$. Thus, having just the implication from the Empirical Relational System to the Numerical Relational System means that a measure must order correctly all those entities whose complexity is ordered in the Empirical Relational System. Whenever $<e_1,e_2> \notin$ *more_complex_than* and $<e_2,e_1> \notin$ *more_complex_than*, we are indifferent as for the relative values that the measures may associate with e_1 and e_2, i.e., we may indifferently have either have $m(e_1) > m(e_2)$, $m(e_1) = m(e_2)$, or $m(e_2) > m(e_1)$.

The presence of "false positives," also entails that, in the control-flow complexity example, when we have $m(e_1) > m(e_2)$, we cannot draw the conclusion that $<e_1,e_2> \in$ *more_complex_than* with certainty. Thus, there may be a loss of information when one uses the Weak Representation Condition. This potential loss should be compared against the benefits of using a less strict approach, as we now explain.

First, it should be borne in mind that, for a number of currently widely used measures, the Representation Condition simply does not hold. Two alternatives exist:

- discard all the measures that do not satisfy the original Representation Condition;
- study them and their properties with the Weak Representation Condition.

Second, MT is a relatively recent development and its usefulness should be assessed based on its applicability to real cases. It is a fact that, so far, MT has provided very little guidance in the definition of new software engineering measures and has had a mainly "destructive" function in finding out problems with the existing software engineering measures. We would like to add that the Weak form of MT has been suggested in the past as an alternative to MT [1], but, to the best of this author's knowledge, this avenue of research has not been pursued any further.

Third, in several cases, it is not possible to provide a strict weak order of the entities according to some attribute of interest, at least at this stage of the software measurement field. The examples of [2] are a clear proof of that. Multidimensional measurement has been suggested as an alternative in these cases. Instead of considering an attribute as a single one, it should be considered as composed of several different (sub)attributes, each of which can be measured independently. However, it may not be easy to identify the (sub)attributes that compose the attribute

easy to identify the (sub)attributes that compose the attribute of interest. Also, even if all the (sub)attributes were identified, a consensus on a strict weak order on the entities according to the overall attribute of interest cannot usually be obtained.

We propose instead that WMT be investigated so as to provide a less demanding mathematical theory as a foundation for measures. This implies providing new definitions in WMT for a number of concepts that exist in MT, starting from meaningful statement. The most natural way to provide a new definition of meaningful statement is to use the same notion as in Definition 6, except that the new definition involves replacing a weak scale with another weak scale.

Definition 12. Weak Meaningful Statement. A statement is called a weak meaningful statement if its truth value does not change if a weak scale is replaced by another weak scale. Formally, if $S(m)$ is a statement based on measure m and $S(m')$ is the same statement obtained by replacing m with m', we have $S(m) \Leftrightarrow S(m')$.

A weak meaningful statement is invariant across all weak scales. Based on the notion of weak meaningful statement, a theory of levels of measurement can be defined in WMT. Being more liberal than MT, Weak Measurement Theory imposes less strict constraints on the scales belonging to each category.

Weak Nominal Scales. The meaningful statements of this class of scales are of the form $m(e_1) = m(e_2)$, for at least one pair of entities e_1, e_2. If $m(e_1) = m(e_2)$ for a pair of entities e_1, e_2 and one scale m, we must have $m'(e_1) = m'(e_2)$ for all other scales m'. Whenever the set of values V of the measures has at least the same cardinality as the set of values of the entities E, we cannot impose that the entities e_1, e_2 be necessarily different. Otherwise, scales that are classified as nominal in MT would be excluded by the definition. In MT, when $|V| \geq |E|$, there are at least as many values as entities, so we can define a nominal measure just by assigning a different value to each entity. Thus, there may not exist two distinct entities e_1, e_2 such that $m(e_1) = m(e_2)$. However, when $|V| < |E|$, we are forced to provide the same value to at least two entities, in both MT and WMT. In both MT and WMT, values are labels for subsets of entities, with the following difference. In MT, the set of entities is partitioned in a set of equivalence classes, each of which receives a different label by every scale. In WMT, two different classes may receive the same label by some scale.

Weak Ordinal Scales. $<ERS, NRS, m>$ is a weak ordinal scale if $m(e_1) > m(e_2)$ is a weak meaningful statement for at least one pair of entities e_1, e_2. It is not required that $m(e_1) > m(e_2)$ or $m(e_1) = m(e_2)$ be weak meaningful statements for all pairs of entities e_1, e_2. The values of these scales are labels for categories in which the entities are classified, with a *not necessarily total ordering* across the categories.

Weak Interval Scales. $<ERS, NRS, m>$ is a weak interval scale if $(m(e_1)-m(e_2))/(m(e_3)-m(e_4))=k$ is a weak meaningful statement for at least one 4-tuple of entities e_1, e_2, e_3, e_4, i.e., k is a constant value across all scales. It is not required that this statement be meaningful for all 4-tuples of entities. Thus, we can quantify the difference between measurement values for some entities.

Weak Ratio Scales. <*ERS*, *NRS*, *m*> is a weak ratio scale if $m(e_1)/m(e_2) = k$ is a weak meaningful statement for at least one pair of entities e_1, e_2, i.e., k is a constant value across all scales. Each entity is associated with a numerical value that provides a quantification for the ratio between some measurement values.

As admissible transformations across scales may not exist even in MT, we have defined the above levels of measurement based on the notion of weak meaningful statement. The notion of admissible transformation is less likely to be useful in WMT for this purpose, because the set of weak scales of WMT includes the set of scales of MT. Thus, the set of transformations that lead from one weak scale to another weak scale may not be easy to identify. For instance, the set of admissible transformations that lead from one weak ordinal scale to another includes all those transformations that preserve only the truth value of the meaningful statements, i.e., the ordering among those pairs of entities e_1, e_2 for which we must have $m(e_1)>m(e_2)$ for all measures m. Take two entities e_3, e_4 for which we need not have $m(e_3)>m(e_4)$ for all measures m, i.e., $m(e_3)>m(e_4)$ is not a weak meaningful statement. Given a measure m' for which $m'(e_3)>m'(e_4)$, there exists an admissible transformation of m' that leads to a measure m'' for which $m''(e_3)>m''(e_4)$ and another admissible transformation that leads to a measure m''' for which $m'''(e_4)>m'''(e_3)$.

At any rate, it is true that the admissible transformations used for the different levels of measurement of MT can be safely used to transform one weak scale into another weak scale in WMT. For instance, by applying any non-decreasing strictly monotonic transformation to a weak ordinal scale we still obtain a weak ordinal scale, since it certainly preserves the truth value of all meaningful statements.

4 Existence of Weak Scales for the Ordinal and Ratio Levels

Here, we first show how the theory of ordinal scales of MT can be extended to weak ordinal scales in WMT (Section 4.1). Then, we show how a theory of weak extensive measurement can be built in WMT (Section 4.2). In what follows, we need to use in the Empirical Relational Systems the concept of hierarchy, which we define next.

Definition 13. Hierarchy. A hierarchy is a pair <*E*, *R*>, where $R \subseteq E \times E$ is a binary relation on E such that it does not contain any cycle, i.e., any sequence of pairs $\{<e_1,e_2>, <e_2,e_3>, \ldots, <e_i,e_{i+1}>, \ldots, <e_n,e_{n+1}>\}$ of any length n with $\forall i \in 1..n\ R(e_i,e_{i+1})$ such that $e_1=e_{n+1}$.

For instance, hierarchy <*E*,*R*> where $E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ and $R = \{<e_6,e_2>, <e_5,e_3>, <e_4,e_3>, <e_3,e_2>, <e_3,e_1>\}$ can be graphically represented as in Figure 2.

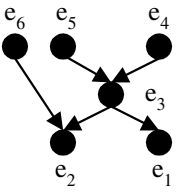

Fig. 2. A hierarchy

4.1 Weak Ordinal Scales

Based on operations research, Adams [1] proved Theorem 2, analogous to Theorem 1 for MT. (A different proof is in [4], with the proofs for all other theorems).

Theorem 2. **Necessary and Sufficient Condition for the Existence of a Weak Ordinal Scale.** Given the Empirical Relational System $ERS=<E,R>$, with $R \subseteq E \times E$ and the Numerical Relational System $NRS=<Real_+, >>$ ($Real_+$ is the set of positive real numbers), a weak ordinal scale $<ERS,NRS,m>$ exists iff R is a hierarchy.

When the implication from the Numerical Relational System to the Empirical Relational System is removed, the theorem shows that relation R of the Empirical Relational System may not be a strict weak order for a weak scale to exist. So, we have a scale type in WMT that is between the Nominal and the Ordinal ones of MT.

4.2 Weak Extensive Structure and Weak Ratio Scales

We prove a theorem on how we can even extend the results of Extensive Structures of MT to define Weak Extensive Structures in WMT and weak ratio scales.

Definition 14. Weak Extensive Structure. Let E be a set, $R \subseteq E \times E$ is a binary relation on E, and o a total function o: $E \times E \rightarrow E$. The triple $<E, R, o>$ is said to be a Weak Extensive Structure if and only if the following axioms hold.

A1 (Weak Associativity): $\forall\ e_1,e_2,e_3 \in E\ (Eq(e_1 o(e_2 o e_3), (e_1 o e_2) o e_3))$, where Eq is an equivalence relation defined as $Eq(e_1,e_2) \Leftrightarrow \neg R(e_1,e_2) \wedge \neg R(e_2,e_1)$.

Axiom A1 states that the order in which the entities in E are combined does not affect the ordering of the resulting elements. Thus, if $e_4=e_1 o(e_2 o e_3)$ and $e_5=(e_1 o e_2) o e_3$, it is neither $R(e_4,e_5)$ nor $R(e_5,e_4)$. The associativity of the combination operation is weak (like in MT) because it is not required that $e_1 o(e_2 o e_3)=(e_1 o e_2) o e_3$.

A2 (Hierarchy): $<E, R>$ is a hierarchy.

A3 (Monotonicity): $\forall\ e_1,e_2,e_3 \in E\ (R(e_1,e_2) \Rightarrow \neg R(e_2 o e_3, e_1 o e_3))$. The order between two entities e_1 and e_2 cannot be changed by combining e_1 and e_2 with any other entity e_3.

A4 (Archimedean): $\forall\ e_1,e_2,e_3,e_4 \in E\ (R(e_1,e_2) \Rightarrow \exists n \in \mathrm{N}\ \neg R(ne_2 o e_4, ne_1 o e_3))$ where ne is

recursively defined for any $e \in E$ as $1e=e$ and $\forall\, n > 1$ $ne=(n\text{-}1)e\ o\ e$. Axiom A4 states that, for any pair of entities e_3,e_4 and for any ordered pair of entities e_1,e_2, it is possible to find a number n such that the combination of e_1 n times with e_3 is not ordered with respect to the combination of e_2 n times with e_4.

Theorem 3 extends to WMT the theorem of the Extensive Structures for MT.

Theorem 3. Existence of an Additive Scale for a Weak Extensive Structure. Let E be a set, $R \subseteq E \times E$ a binary relation on E, and o a total function $o: E \times E \rightarrow E$. There is a function $m: E \rightarrow Real$, where *Real* is the set of real values, such that $R(e_1,e_2) \Rightarrow m(e_1) > m(e_2) \wedge m(e_1 o e_2) = m(e_1) + m(e_2)$ if and only if $<E,R,o>$ is a Weak Extensive Structure.

Theorem 3 shows that there are several families of measures, one for each Strict Weak Order that can be defined on the transitive closure of R. This is different from MT, where there is only one such family of measures, and one measure m' can be obtained from any another measure m by multiplying m by a suitable coefficient. Each of the measures in WMT is an additive one. It is immediate to show that the scale built on top of an additive measure m is a weak ratio scale.

Theorem 4. Weak Additive Scales and Weak Ratio Scales. Let $<ERS, NRS, m>$ be a weak scale, with $ERS = <E, R, o>$ (where E is a set, $R \subseteq E \times E$, and o a total function $o: E \times E \rightarrow E$), $NRS = <Real, >, +>$, and m a function $m: E \rightarrow Real$, i.e., $R(e_1,e_2) \Rightarrow m(e_1) > m(e_2) \wedge m(e_1 o e_2) = m(e_1) + m(e_2)$. Then, $<ERS, NRS, m>$ is a weak ratio scale.

5 The Mean Value to Evaluate Central Tendency for Weak Scales

One of the most important characteristics of a set of measurements is its central tendency. Several indicators may be defined for measuring the central tendency of a set of measurements, e.g., the arithmetic mean and the median. In this section, we show when the arithmetic mean and the median can be used in WMT.

We would like to point out that, in MT, it is often said that the mean value of a set of ordinal measurements cannot be taken as an indicator of central tendency. The following example is typically used to show that the mean of an ordinal scale should not be used because it leads to meaningless statements. Suppose that m' is an ordinal measure, and that we have obtained two sets of failures $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_p\}$ on two programs. Take statement $\frac{1}{n}\sum_{i \in 1..n} m'(x_i) > \frac{1}{p}\sum_{j \in 1..p} m'(y_j)$, which compares the mean values of the measures we have obtained on the two sets of failures. Conventional wisdom says that this statement is not meaningful since its truth value may depend on the specific ordinal scale chosen. In other words, it would not be true that

$$\frac{1}{n}\sum_{i \in 1..n} m'(x_i) > \frac{1}{p}\sum_{j \in 1..p} m'(y_j) \Leftrightarrow \frac{1}{n}\sum_{i \in 1..n} m''(x_i) > \frac{1}{p}\sum_{j \in 1..p} m''(y_j) \quad (3)$$

for every other ordinal measure m'' that we chose to use instead of m'. Examples can be easily built to show this lack of meaningfulness.

Surprisingly, this *general* statement is not true. In some cases, statement (3) is true independent of the ordinal measure chosen. To prove this property, which is against conventional wisdom, we proceed from more extreme less obvious examples and then we will build a general theory. Before doing so, we observe that the computation of the mean of an ordinal measure m with t values (e.g., a failure criticality measure that can assume $t = 5$ different values) can also be carried out as $\sum_{k\in 1..t} f_k m_k$, where f_k is the relative frequency with which value m_k is obtained. For instance, suppose that failures can be classified into five classes ($t = 5$) with the class whose index is $k = 1$ being the class of the least critical failures and the class corresponding to $k = 5$ being the class of the most critical failures. Note that k is just the index of the different classes, so it is not the measure chosen. Frequency f_k is the proportion of failures obtained by executing the program that have been assigned to class k, e.g., $f_1 = 0.2, f_2 = 0.25, f_3 = 0.35, f_4 = 0.1, f_5 = 0.1$ (the sum of the frequencies must be 1).

Thus, we can rewrite (3) as

$$\sum_{k\in 1..t} f_k m'_k > \sum_{k\in 1..t} g_k m'_k \Leftrightarrow \sum_{k\in 1..t} f_k m''_k > \sum_{k\in 1..t} g_k m''_k \qquad (4)$$

where g_k is the frequency of class k failures obtained by executing a different program (or even by carrying out a different set of executions on the same program, with a different testing techniques, for instance).

Let us start with the most extreme example. Suppose we have obtained the following two frequency distributions: $f_1=0, f_2=0, f_3=0, f_4=0, f_5=1$ and $g_1=1, g_2= 0, g_3= 0, g_4=0, g_5=0$. Statement (4) is obviously true, since it becomes $m'_5>m'_1 \Leftrightarrow m''_5>m''_1$. Thus, *any* admissible (i.e., strictly monotonic) transformation of measure m' into a measure m'' will not change the truth value of the statement $m'_5 > m'_1$. Let us now move on to an even less extreme example, with the following two frequency distributions: $f_1 = 0.05, f_2 = 0.15, f_3 = 0.2, f_4 = 0.2, f_5 = 0.4$ and $g_1 = 0.1, g_2 = 0.15, g_3 = 0.25, g_4 = 0.3, g_5 = 0.2$. Statement (4) becomes

$$0.05m'_1+0.15m'_2+0.2m'_3+0.2m'_4+0.4m'_5 > 0.1m'_1 + 0.15m'_2 + 0.25m'_3+0.3m'_4+0.2m'_5$$

$$\Leftrightarrow$$

$$0.05m''_1+0.15m''_2+0.2m''_3+0.2m''_4+0.4m''_5>0.1m''_1+0.15m''_2+0.25m''_3+0.3m''_4+0.2m''_5$$

This statement is true, and the reason can be found in the following theorem.

Theorem 5. Let $\{w_1, ..., w_t\}$ and $\{m_1, ..., m_t\}$ be two sets of real numbers. Let the total ordering among the m_i's be known, and, let us suppose that $0 < m_1 < m_2 < ... < m_t$. We have $\sum_{k\in 1..t} w_k m_k > 0$ for every choice of $m_1, m_2, ..., m_t$ such that $0 < m_1 < m_2 < ... < m_t$ if and only if, for all $1\le q\le t$, the sum of the weights $\sum_{k\in q..t} w_k > 0$.

Theorem 5 can be used to compare two mean values by rewriting (4) as

$$\sum_{k\in 1..t} (f_k-g_k)\, m'_k > 0 \Leftrightarrow \sum_{k\in 1..t} (f_k-g_k)\, m''_k > 0$$

to apply Theorem 5 with $w_k = (f_k-g_k)$. Thus, it is not true in general that statement (3) is false in the general case even in MT. In WMT, we can prove a similar theorem for a special case of hierarchy (a forest), whose definition is based on the definition of tree.

Definition 15. Trees and Forests. A tree is a connected hierarchy in which (1) there exists only one node r (the root) such that there is no other entity a for which $<r,a> \in R$ and (2) for any node a_1 there is at most one other node a_2 such that $<a_1,a_2> \in R$. Two trees are disjoint if they do not share any nodes. A forest is the union of a set of disjoint trees. Two nodes c, d in the forest are said to be ordered if the pair $<c,d>$ belongs to the irreflexive transitive closure of R. Given a node d, the set of nodes c for which the pair $<c,d>$ belongs to the irreflexive transitive closure of R or is the pair $<d,d>$ is denoted by $Prec(d)$.

Theorem 6. Let $\{w_1, ..., w_t\}$ and $\{m_1, ..., m_t\}$ be two sets of real numbers. Let a forest ordering among the m_i's be known. We have $\sum_{k \in 1..t} w_k m_k > 0$ for every choice of non-negative $m_1, m_2, ... , m_t$ satisfying the given forest ordering if and only if, for all nodes d, we have $\sum_{k \in Prec(d)} w_k > 0$, i.e., where the sum of the w_k's is carried out over all k for which it is known from the forest ordering that $m_k > m_d$ plus the value w_d.

The result of Theorem 6 can be used to solve the problem in the general case of hierarchies. We first need to transform the hierarchy in a forest. The roots of the trees of the forest are the nodes of the hierarchy that have no successors. One tree will be built for each of these nodes. Then, a depth-first search is carried out from these roots. When a node a with more than one successor is found, replicate that node in each of the trees that are being built. The value of measure $m(a)$ is left unchanged, while the total weight $w(a)$ is distributed across all the new nodes that have been created. Thus, we introduce a set of variables representing the weights of the new nodes with the constraint that their sum must be equal to $w(a)$. When the visit of the hierarchy is completed, we have a forest and a set of linear equations that link the weights of the new nodes created. For instance, starting from the hierarchy in Figure 2, we obtain the forest represented in Figure 3.

So, the general problem is reduced to the following one: Given a finite set of linear equations (derived from the creation of the new nodes) and linear inequalities (derived by the requirements of Theorem 6) involving real valued variables, decide if the polytope determined by these constraints is empty. This problem has a Linear Programming algorithmic solution of relatively low computational cost, which is also the first part of the Symplex algorithm. In a time which is polynomial in the number of the variables we can decide if this set is empty. If not, the algorithm provides a solution point as well. Also, the entire process can be easily automated.

From the viewpoint of WMT, finding a solution to the above problem implies that the comparison between mean values is meaningful, i.e., its truth value does not change for any choice of the measure m' used.

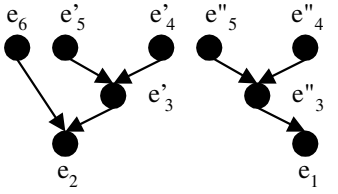

Fig. 3. Transformation of the hierarchy of Fig. 2 into a forest

6 Conclusions and Future Work

In this paper, we have presented a proposal for an approach to defining a theory of measurement with less strict requirements than MT. As such, WMT may be more easily applicable to the current state of the art of software measurement.

A more thorough investigation of the issues outlined in this paper is clearly called for, even though this paper may represent a consistent nucleus around which a full-fledged theory may be developed. Among the issues, we need to investigate further whether the notions of weak measurement levels is adequate or should be refined; how the properties of many important software engineering attributes, e.g., complexity, cohesion, coupling, may be modeled in the framework of WMT; what other gains and losses may result from the use of WMT instead of MT.

Acknowledgements. Special thanks to Stefano Serra-Capizzano for the useful discussions. This work was carried out with the partial support of MIUR and CNR.

References

1. E.W. Adams, "Elements of a Theory of Inexact Measurement" Philosophy of Science, Vol. 32, No. 3, pp. 205–228, July 1965.
2. N. Fenton, "Software Measurement: A Necessary Scientific Basis," IEEE Transactions on Software Engineering, Vol. 20, No. 3, pp. 199–206, March 1994.
3. D.H. Krantz, R. D. Luce, P. Suppes, A. Tversky, "Foundations of Measurement," Academic Press, New York, 1971.
4. S. Morasca, "A Weak Measurement-Theoretic Approach to Software Measurement," Università degli Studi dell'Insubria, Dip. di Scienze CC. FF. MM., Tech. Report. 2002.
5. F.S. Roberts, "Measurement Theory," Addison-Wesley, Reading, 1979.