On the Application of Measurement Theory in Software Engineering

Lionel Briand
Centre de Recherche Informatique de Montréal (CRIM)
Software Engineering Group
1801 McGill College av.
Montréal, PQ, H3A 2H4
Canada
e-mail: lbriand@crim.ca

Khaled El Emam
Centre de Recherche Informatique de Montréal (CRIM)
Software Engineering Group
1801 McGill College av.
Montréal, PQ, H3A 2H4
Canada
e-mail: kelemam@crim.ca

Sandro Morasca
Dipartimento di Elettronica e Informazione
Politecnico di Milano
Piazza L. Da Vinci 32,
I-20133, Milano
Italy
e-mail: morasca@elet.polimi.it

Abstract

Elements of measurement theory have recently been introduced into the software engineering discipline. It has been suggested that these elements should serve as the basis for developing, reasoning about, and applying measures. For example, it has been suggested that software complexity measures should be additive, that measures fall into a number of distinct types (i.e., levels of measurement: nominal, ordinal, interval, and ratio), that certain statistical techniques are not appropriate for certain types of measures (e.g., parametric statistics for less-than-interval measures), and that certain transformations are not permissible for certain types of measures (e.g., non-linear transformations for interval measures). In this paper we argue that, inspite of the importance of measurement theory, and in the context of software engineering, many of these prescriptions and proscriptions are either premature or, if strictly applied, would represent a substantial hindrance to the progress of empirical research in software engineering. This argument is based partially on studies that have been conducted by behavioral scientists and by statisticians over the last five decades. We also present a pragmatic approach to the application of measurement theory in software engineering. While following our approach may lead to violations of the strict prescriptions and proscriptions of measurement theory, we demonstrate that in practical terms these violations would have diminished consequences, especially when compared to the advantages afforded to the practicing researcher.

1. Introduction

In empirical software engineering — like in other empirical sciences (e.g., experimental psychology) where measurement is noisy, uncertain, and difficult — the definition of sensible measures, their statistical analysis, and the search for patterns amongst variables are difficult activities. In recent years, measurement theory has been proposed and extensively discussed [Z91, F94] as a means to evaluate the software engineering measures that have been proposed in the literature, and to establish criteria for the statistical techniques to be used in data analysis and in the search for patterns. As discussed below, measurement theory is a very convenient theoretical framework to explicitely define the underlying theories upon which software engineering measures are based. This means that measures are not defined out of context and that the theories on which they are based can be discussed, adapted, and refined.

Some software engineering researchers [F94, Z91] have advised that a number of powerful statistical techniques (i.e., parametric statistics) should be proscribed when one cannot prove the measures s/he uses fulfill basic scale requirements (i.e., those of the interval scale). However, this very restrictive and rigid view of the use of measurement theory is not shared by many statisticians and data analysts (e.g., see [Tuk86a, Gar75, VW93]). There has been, for almost fifty years (since the publication of Steven's 1946 paper [Ste46]), another side to a very intense debate. Many of these statisticians and data analysts have stated that a more pragmatic approach for evaluating measures and for selecting data analysis methods is called for; an approach which is likely to result in more results and which can be deemed reasonably reliable and accurate. In many practical applications of measurement in software engineering, a purely mathematical, rigid, and schematic viewpoint on measurement theory turns out to be rather sterile in

terms of results. This can cause a serious hindrance in the development of a discipline such as empirical software engineering, whose state of the art is still in a rather exploratory phase. It is our position that a more flexible and pragmatic approach would allow researchers and practitioners to obtain a number of badly needed results — i.e., patterns that can explain the relevant phenomena.

Our goals in this paper are threefold: (1) review and discuss the current usage of measurement theory in software engineering, (2) provide a balanced view of the issues and solutions related to the use of measurement theory than has existed thus far so that our field can benefit from the existing debate, and (3) present a pragmatic approach to applying measurement theory in empirical software engineering research. In the following section, we will present the basic concepts of measurement theory so that the reader who is not acquainted with it can understand the remainder of the paper. In Section 3, we present a critical review of the current usage and interpretation of measurement theory in our field. Subsequently, in Section 4 we propose a pragmatic approach for applying measurement theory in data analysis. We justify our approach by demonstrating that parametric statistics are robust to violations of interval scale properties (Section 5), that resorting to nonparametric statistics results in loss of statistical power (Section 6), and that performing transformations that are not admissible from a measurement-theoretic viewpoint can still produce useful results (Section 7). Section 8 concludes the paper with a summary of the main points.

2. Basic Concepts of Measurement Theory

For the reader's convenience, we now present some basic definitions and notations of measurement theory, as defined in [Z91, pp. 40 - 51], based on [R79].

A relational system **A** is an ordered tuple $(A,R_1,\ldots,R_n,o_1,\ldots o_m)$ where A is a nonempty set of objects, the R_i, i = 1,…,n are ki-ary relations on A and the o_j, j=1,…,m are closed binary operations. For measurement we consider two relational systems: **the empirical** and **formal relational systems**.

Empirical Relational System:

A *= $(A,R_1,\ldots,R_n,o_1,\ldots o_m)$.*

A *is a non-empty set of empirical objects that are to be measured (in our case program texts, flowgraphs, etc.).*

R_i *are ki-ary empirical relations on A with i = 1,…,n. For example, the empirical relation "equal to or more complex than".*

o_j *are binary operations on the empirical objects A that are to be measured (for example a concatenation of control flowgraphs) with j=1,…,m.*

The empirical relational system describes the part of reality on which measurement is carried out (via the set of objects A) and our empirical knowledge of the objects' attributes that we want to measure (via the collection of empirical relations R_i's). Depending on the attributes we want to measure, different relations are used. For instance, if we are interested in program length, we may want to use the relation "longer than" (e.g., "program P_1 is longer than program P_2"); if we are interested in program complexity, we may want to use the relation "more complex than" (e.g., "program P_3 is more complex than program P_4"). Binary operations may be seen as a special case of ternary relations between objects. For instance, suppose that o_1 is the concatenation operation between two programs. We may see it as a relation Concat(Program1, Program2 ,Program3), where Program3 is obtained as the concatenation of Program1 and Program2, i.e., Program3 = Program1 o_1 Program2. It is important to notice that an empirical relational system does not contain any reference to measures or numbers. Only "qualitative" statements are made, based on our understanding of the attribute. These statements are then translated into relations that belong to a formal relational system, as explained below.

Formal Relational System:

B *= $(B,S_1,\ldots,S_n,\bullet_1,\ldots\bullet_m)$.*
B *is a non-empty set of formal objects, for example numbers or vectors.*
S_i *are ki-ary relations on B such as "greater than" or "equal to or greater than".*
$\bullet_j$ *are closed binary operations B such as addition or multiplication.*

The formal relational system describes (via the set B) the domains of the measures for the studied objects' attributes. For instance, these may be integer numbers, real numbers, vectors of integer and/or real numbers, etc. A formal relational system also describes (via the collection of relations S_i's) the relations of interest between the measures. The link between the empirical relational system and the formal relational system is provided by measures and scales, as follows.

Definition 4.1 *(Measure μ):*
A ***measure*** *μ is a mapping $\mu:A\rightarrow B$ which yields for every empirical object $a\in A$ a formal object (measurement value) $\mu(a) \in B$. Of course, this mapping may not be arbitrary. This leads to the following definition of a scale.*

Definition 4.2 *(Scale):*[1]
Let $\mathbf{A} = (A,R_1,\ldots,R_n,o_1,\ldots,o_m)$ be an empirical relational system and $\mathbf{B} = (B,S_1,\ldots,S_n,\bullet_1,\ldots\bullet_m)$ a formal relational system and μ a measure. The Triple (**A**,**B**,μ) is a scale if and only if for all i,j and for all $a_1,\ldots a_k$, b, c ∈ A the following holds

$R_i(a_1,\ldots a_k) \Leftrightarrow S_i(\mu(a_1),\ldots,\mu(a_k))$ and $\mu(b\ o_j\ c) = \mu(b) \bullet_j \mu(c)$

If B = R is the set of real numbers, the Triple (**A**,**B**,μ) is a **real scale**.

Every object *a* of A is mapped into a value of B, i.e., it is measured according to measure μ(a). Every empirical relation R_i is mapped into a formal relation S_i. For instance, the relation "more complex than" between two programs is mapped into the relation ">" between the complexity measures of two programs. The formal relations must preserve the meaning of the empirical statements. For instance, suppose that R_1 is the empirical relation "more complex than," S_1 is the formal relation ">," and μ is a complexity measure. Then, we must have that program P_1 is more complex than program P_2 if and only if $\mu(P_1) > \mu(P_2)$.

Measurement values cannot be used in a totally unrestricted manner. One may obtain results that do not make sense by applying mathematical operation to numbers obtained through measurement. The transformational approach reported below has been used to classify measures according to their *level of measurement* in order to understand what kind of mathematical operations can be carried out on the measurement values they yield.

Definition 4.3 *(Admissible Transformation):*
*Let (**A**,**B**,μ) be a real scale. A mapping g: $\mu(A) \rightarrow B$ is an admissible transformation if and only if (**A**,**B**,g ∘ μ) is also a scale (where g ∘ μ stands for the composition of the two functions g and μ, i.e., $g(\mu(x))$).*

Definition 4.4 *(Meaningfulness):*
A statement is meaningful if and only if its truth value is invariant against all admissible transformations.

Scales are also defined by admissible tranformations. For real scales there is a classification of scales according to their admissible tranformations [Ste46, Ste51, R79]:

[1] Definition 4.2 was slightly changed as compared to the one in [Z91].

Name of the Scale	*Transformation g*
Nominal Scale	*Any one to one g*
Ordinal Scale	*g: Strictly increasing function*
Interval Scale	$g(x) = a\,x + b$
Ratio Scale	$g(x) = a\,x$
Absolute Scale	$g(x) = x$

In other words:

1. a nominal scale measure is just a classification of the objects; the only possibile transformations are those that preserve the fact that objects are different, i.e., one-to one transformations
2. an ordinal scale measure is a ranking of the objects, according to some ordering criterion
3. an interval scale measure is such that differences between values are meaningful, but not the values of the measure itself (e.g., temperature, measured on a Celsius or Fahrenheit scale)
4. a ratio scale measure is such that there is a meaningful "zero" value, and the ratios between values are meaningful
5. an absolute scale measure is such that no transformation is meaningful (e.g., count of the objects).

The above types of measurement scales[2] (i.e., levels of measurement) are ordered from "less powerful" to "more powerful." In particular, the more powerful scales (interval, ratio, and absolute) provide more information and are more useful for measurement purposes.

Michell [M86] makes a distinction between scale-specific statements and scale-free statements. The former specifies the unit of measurement, while the latter does not specify a unit. For example, “the weight of object *ob1* is twice as much as the weight of object *ob2*” is a scale-free statement, while “the weight of object *ob1* in *kilograms* is twice the weight of object *ob2* in *kilograms*” is a scale specific statement. He then argues that a necessary condition for meaningfulness of statements is being able to infer the scale-free version of a statement from its scale specific version. This basically means that statements are meaningful if they do not depend on the particular unit/scale of measurement. Therefore, under admissible scale transformations (i.e. different units of measurement), the truth value of the statement should remain invariant. For instance, let’s consider the scale-specific statement given above. This statement can be expressed as: f(*ob1*) = 2f(*ob2*), where f(x) is the weight of x in kilograms. This statement is meaningful if and only if its truth value is preserved under all admissible transformations g (i.e., other measures of weight such as pounds). This means that the above statement is equivalent to: (g o f)(*ob1*) = 2[(g o f)(*ob2*)]. Given that the measurement of weight is on a ratio scale, then one can define g(x)=ax, with a>0. With this transformation, the above statement becomes: af(*ob1*)=2af(*ob2*). This statement can be reduced to the original statement, and therefore the transformation g, which is admissible for a ratio scale, preserves the truth value of the original statement (i.e., the statement is meaningful for ratio scales).

As another example, let’s take the statement: f(*ob1*) = b Ln f(*ob2*). Under a transformation g, this statement becomes: (g o f)(*ob1*) = b Ln (g o f)(*ob2*). If we take g(x) = ax as we did above, then the previous statement becomes: af(*ob1*) = b Ln a + b Ln f(*ob2*). It is clear that this transformed statement cannot be made equivalent to the original statement, and therefore the original statement is meaningless for ratio scales.

[2] There are other types of measurement scales that we will not consider here because they are less common. For example, Stevens [Ste62] presents the logarithmic interval scale, and Roberts [R79] presents the difference scale.

3. Usage and Interpretation of Measurement Theory in Software Engineering

Before discussing the current state of practice in experimental software engineering, we will first discuss the important concept of complexity. This is necessary since measurement theory has been applied heavily in order to better define and refine the notion of software complexity (especially control flow complexity). The issues related to this important topic will be discussed in subsection 3.1. In subsection 3.2, we will focus on reviewing the current state of the art that has been mainly captured by Zuse's work and identify what we believe are harmful misinterpretations of the theory of measurement.

3.1 Complexity Measures Do not Have to Be Additive

Several different concepts are used in software measurement, e.g., size, complexity, cohesion, coupling. However, people use them in a very unrestricted way, so they are used inconsistently throughout the literature. Often, people refer to cohesion and coupling (and even size) as internal attributes of product complexity. Some authors define metrics for such concepts, but they do so without providing any precise definition for the concept they intend to measure. This is the reason why software measurement often becomes a matter of belief. People do or do not "believe" that a proposed complexity metric is indeed a complexity metric. We understand that these concepts are somewhat subjective, but, until some kind of consensus is reached on the meaning behind those words, any significant progress will be difficult. Therefore, if software measurement is to become a scientific discipline, the meaning of these concepts must be made clear and unambiguous, and similarities and differences between measurement concepts must be pointed out. Measurement Theory, if properly applied, is one of the most precious tools in this task.

In [BMB94], we proposed a set of properties to characterize size, complexity, cohesion, coupling, and length. Our goal was not to provide "the" sets of properties for each of those concepts, but to point out the need for a clear separation between them. Measurement concepts that are different in nature should be characterized by different sets of properties. For the sake of discussion, we will here examine and compare two fundamental and essentially different concepts related to the internal attributes of software products [F94]: size and complexity. The former is the only measurement concept that can be said to be fairly well understood and for which there is some sort of implicit consensus. The latter, instead, is the source of most controversies, since, as we outlined above, it is used to cover many *different* measurement concepts.

Suppose that a program P is made of the sequence of two blocks, B_1 and B_2. It seems natural, and even trivial, to say that the size of the program is the sum of the size of B_1 and the size of B_2. Therefore, it seems intuitive to believe that size is characterized by an additivity property. When two disjoint pieces are put together, then the size of the resulting piece is the sum of the sizes of the two original pieces. Several physical variables satisfy this additivity property, e.g., weight, volume, and height, provided that a suitable definition is provided for the operation of "putting together" two pieces. (For instance, the length of a wooden bench made of two wooden benches is the sum of the lengths of the two wooden benches only if they are "put together" along their short side.) Size measures are characterized by a set of nice analytical properties, so they are easier to study and use than other measures. That is the reason for their success.

However, let us turn to complexity, and let us consider again program P composed of the two blocks B_1 and B_2. Why should the complexity of P be equal to the mere sum of the complexities of B_1 and B_2? In our opinion, and though this is subjective, we would be very surprised if we found someone with a different viewpoint who did not agree that complexity is related to the relationships between the elements of a system, while size is related to the amount of elements in the system. In normal life, when one says

that a system is complex, he or she is referring to the "intricacy" of the relationships between elements of the system. Often, people are faced with systems that are small, but complex, because of the relationships between their elements. Therefore, in program P, if we look at the relationships between elements, we see that there are relationships between the elements of B_1 considered in isolation, relationships between the elements of B_2, considered in isolation, and relationships from the elements of B_1 to the elements of B_2 and from the elements of B_2 to the elements of B_1. If complexity were additive in the same way as size, we should not consider the relationships across the two blocks, i.e., from elements of B_1 to elements of B_2 and from elements of B_2 to elements of B_1. We also have to add that it is these relationships that are created across the blocks that make it more difficult to understand a program than to understand its component blocks — although complexity is not the only factor that has an impact on program understandability. We have been very general, because we are not describing a specific type of complexity, such as control flow or data flow complexity, but all types of complexity related to the internal attributes of software products. Therefore, our position is that complexity measures *do not have* to be additive, under some kind of composition operation between program segments or, more generally, subsystems. That does not mean that they *must not* be additive, either. We only want to point out that additivity must not be a mandatory property for complexity measures, as opposed to size measures.

3.2 Ratio Scale vs. Additive Ratio Scale and Extensive Structure

Unfortunately, despite the arguments presented in the previous section, the argument that complexity measures should be additive [Z91] is gaining increasing acceptance (see [BO94, CK94, F94]). In this section, we perform an in-depth critical analysis of the reasoning behind such a standpoint.

In several papers that appeared in the scientific literature on measurement properties [Z91, Z92, Z94, Z95], Zuse advocates the need for a complexity measure to be additive, and for the underlying empirical system to "assume an extensive structure." For the reader's convenience, we will now present the definition of extensive structure used by Zuse in [Z91, p. 57] (in what follows, we will use [Z91] as the main reference, since it is complete, easily available, and consistent with the other references [Z92, Z94, Z95]). In [Z91, p. 46], Zuse assumes that programs are represented by their flowgraphs, and that P is the set of all flowgraphs.

Theorem 4.2 *(Modified Extensive Structure):*
Let ***P*** *be a non-empty set,* $\bullet\geq$ *a binary relation on P, and o a binary operation on* ***P****. The relational system* (***P***, $\bullet\geq$, *o) is an* ***Extensive Structure*** *if and only if the following axioms hold for all P1, …, P4*$\in P$.

A1: $(P, \bullet\geq)$	*is a weak order*
A2: $P1 \circ (P2 \circ P3) \approx (P1 \circ P2) \circ P3$	*axiom of weak associativity*
A3: $P1 \circ P2 \approx P2 \circ P1$	*axiom of weak commutativity*
A4: $P1 \bullet\geq P2 \Rightarrow P1 \circ P3 \bullet\geq P2 \circ P3$	*axiom of weak monotonicity*
A5: If $P3 \bullet> P4$ *then for any P1, P2 there exists a natural number n, such that* $P1 \circ n\ P3 \bullet> P2 \circ n\ P4$	*Archimedean Axiom*[3]

For our purposes, the empirical relation "$\bullet>$" has the meaning "more complex than," the empirical relation "$\approx$" the meaning "as complex as," and the empirical relation "$\bullet\geq$" the meaning "more complex than or as complex as." We will also say that the binary operation o between two objects is the composition of the two objects. The above theorem (by Bollmann [B84]) provides a set of axioms for extensive structures as a modification of the original and accepted ones, provided in [K71, R79]. In other words,

[3] The notation n P denotes the composition of P with itself n times.

- axiom A1 states that there is an order relation between objects which is complete, reflexive, and transitive [Z91, p. 47]
- axiom A2 states that the result of a series of compositions does not depend on the order in which they are carried out
- axiom A3 states that the complexity of the composition of two objects does not depend on which object comes first and which comes second in the composition
- axiom A4 states that composition preserves the ordering between objects, i.e., if we compose two objects P1 and P2 with the same object P3, then the order relationship between P1 and P2 is the same as that between P1 o P3 and P2 o P3
- axiom A5 states that given any two objects P1 and P2, with P1 less complex than P2, and two other objects P3 and P4, with P3 more complex than P4, we may always compose P1 with P3 a sufficient number of times n and P2 with P4 n times so that P1 composed n times with P3 is more complex than P2 composed n times with P4.

It is our position that some of the axioms of extensive structures are not suitable for software complexity measurement. The argument used in [Z91] in favor of the use of extensive structures is that additive measures (whose underlying empirical system is an extensive structure) are ratio scale measures. Therefore, extensive measurement is a priori appealing because it is a way of achieving the ratio scale. However, the converse is not true, i.e., not all ratio scale measures are additive and assume an extensive structure (for instance, in [Z91], p. 46, one can read: "Hence, a ratio scale is not always additive" and in p. 51: "However, there exist other possibilities to give necessary and sufficient conditions for the ratio scale."). An extensive structure is a *sufficient* condition for obtaining a ratio scale measure, but *by no means a necessary one* (also see [SZ63] who make the same point). Extensive measurement is remarkable in the sense that, assuming the properties of the extensive structure to be true, one can prove that a measure is defined on a ratio scale. Moreover, additivity is in many circumstances a convenient property.

As a parallel with monotonic functions from real numbers to real numbers, one may study linear functions, which are an important subset of monotonic functions provided with nice and elegant mathematical properties (e.g., linear superposition). However, there are monotonic functions that are not linear, but that are nevertheless useful for studying physical phenomena and interesting from a purely mathematical point of view. In much the same way, there are ratio scale measures that are not additive measures, but are both useful in software engineering measurement and an interesting subject for theoretical studies. We believe that measures whose underlying empirical relational system is an extensive structure represent too narrow a set of measures to be useful for all software engineering measurement concepts, and, most of all, for complexity.

We will now point out the problems deriving from the use of extensive structures as presented by Zuse by examining his review of Weyuker's set of properties for complexity measures [W88]. Based on the axioms of extensive structures, Zuse [Z91] states that Weyuker's properties for complexity measures are inconsistent from a measurement-theoretic point of view. In particular, according to Zuse, Weyuker's property W9 requires the ratio scale (i.e., property W9 is meaningful for transformations of the form $g(x)=ax$), while Weyuker properties W6 and W7 reject the ratio scale[4] (i.e., properties W6 and W7 are not meaningful for transformations of the form $g(x)=ax$). However, that is not true, as we now show. In what follows, P, Q, and R will represent program bodies, as defined in [W88].

Property W6

$\exists P, \exists Q, \exists R (\mu(P) = \mu(Q) \text{ and } \mu(P;R) \neq \mu(Q;R))$
$\exists P, \exists Q, \exists R (\mu(P) = \mu(Q) \text{ and } \mu(R;P) \neq \mu(R;Q))$

[4] By W6, W7, and W9, we will denote Weyuker's properties 6, 7, and 9 of [W88]. These properties correspond to properties 5, 6, and 8 in the review of Weyuker's properties reported in [Z91, pp. 92-96].

[Z91, p. 94] says that "if this property of software complexity measures would become a general accepted property, the way to the ratio scale by the Extensive Structure would be blocked. The axiom of monotonicity is an axiom of the Extensive Structure. The Extensive Structure… is the way to come to the ratio scale, which is required in the literature for software complexity measures."

Property W7

“There are two program bodies P and Q such that Q is formed by permuting the order of the statements of P, and $\mu(P) \neq \mu(Q)$.”

[Z91, p.95] says that "Weyuker does not require the axiom of commutativity which is required by Bache /BACH87/. The axiom of commutativity is also a prerequisite of the Extensive Structure. If the order of statements should be reflected by a software complexity measure as proposed above, then there is no way to come to the ratio scale via the Extensive Structure and the proposed concatenation of program by Weyuker."

Property W9

$\exists P, \exists Q (\mu(P) + \mu(Q) \leq \mu(P \circ Q))$

(It is worth noting that in Zuse's description of this property [Z91, p. 95] the existential quantifiers are substituted with universal quantifiers. The above formula — with existential quantifiers — is the one that appears in the original published paper [W88]. Actually, Weyuker explicitly rejects the version with universal quantifiers [W88, unnumbered property after property W9].) In [Z91, p. 96] one can read that W9 "is not meaningful for an interval scale but is meaningful for a ratio scale."

All of the above comments quoted for properties W6, W7, and W9 from [Z91] are appropriate. Properties W6 and W7 are inconsistent with two of the requirements of extensive structures, i.e., weak monotonicity and weak commutativity, and property W9 is valid on a ratio scale. Therefore, Zuse correctly argues that "the way to the ratio scale by the Extensive Structure would be blocked." However, it is incorrectly concluded that "The property W6 denies the ratio scale via the Extensive Structures while property W9 requires the ratio scale. That together is a contradiction." ([Z91, p. 96]). Two paragraphs later, Zuse writes: " ... with property W9 Weyuker requires the ratio scale and rejects the ratio scale with property W7 and W6. That is a contradiction." These arguments are incorrect because properties W6 and W7 do not exclude the ratio scale, but only the ratio scale VIA extensive structures (as mentioned by Zuse himself on pp. 94 and 96!). Moreover, by applying the ratio scale admissible transformations, one can easily prove that properties W6, W7, and W9 are meaningful for the ratio scale.

The above discussion of Weyuker's properties also shows that some of the extensive structure axioms, namely weak commutativity and weak monotonicity as provided by [Z91], are NOT obviously intuitive for complexity measurement. Therefore, they should not be imposed just to achieve a ratio scale via extensive measurement. For example, most well accepted data flow complexity metrics [Ovi80] do not assume commutativity. For example, P1 o P2 might not contain as many definition-use pairs as P2 o P1. From a more general perspective, the set of relationships between elements contained in P1 o P2 might not be identical to P2 o P1, regardless of the particular relationship studied, nor must it be as complex. As Zuse's analysis of property W7 shows, a consequence of the weak commutativity axiom of Extensive Structures is that any two programs composed of the same statements are equally complex. That implies that the complexity of a program does not at all depend on the order of statements, i.e., the way they are related to each other, but only on the statements it contains. Therefore, requiring commutativity keeps researchers from studying relationships and therefore, we believe, the very nature of complexity.

A similar discussion could be held about the axiom of weak monotonicity. Given two programs P1 and P2, with P1 •≥ P2, and a third program P3, there is no clear reason why this would necessarily imply that P1 o P3 •≥ P2 o P3. For the sake of discussion, suppose that new relationships are introduced when the composition P2 o P3 is carried out, with respect to the relationships in P2 and P3, while no new relationships are introduced when the composition P1 o P3 is carried out, with respect to the relationships in P1 and P3 separately. There is no intuitive reason why P1 o P3 should necessarily be "more complex than or as complex as" P2 o P3. Once again, one should not overlook the relationships created between the two parts being composed.

Unfortunately, this misconception about extensive structures has already had an important impact on the scientific work and the literature. Several authors have quoted Zuse's results in major journals or used his conclusions to validate their work. Examples can be found in journals such as IEEE Transactions on Software Engineering [F94, CK94, BO94].

4. Applying Measurement Theory

4.1 Scale Types and Proscribing Statistics

We will now take a more general perspective, i.e., we will focus on the role that measurement theory should have in empirical software engineering. Several books and papers on the topic of measurement theory are conveying the idea that scale types should be used to proscribe the use of "inappropriate" statistical techniques. For example, a table similar to the one shown in Figure 1 is given in [F91]. This table, for instance, proscribes the use of the Pearson product moment correlation for scale types that are either nominal or ordinal. Such proscriptions, of course, are not unique to software engineering. For instance, they serve as the basis of the classic text of Siegel on nonparametric statistics [SC88], and serve as an integral part of the decision tree developed by Andrews et al. [AKD+81] to guide researchers in the selection of the most appropriate statistics. Accordingly, if a researcher's measures do not reach the interval level, it is advised that s/he use non-parametric statistics (i.e., tests which make less stringent assumptions, such as the Mann-Whitney U test[5]).

Scale Type	Examples of Appropriate Statistics	Type of Appropriate Statistics
Nominal	Mode Frequency Contingency Coefficient	Nonparametric Statistics
Ordinal	Median Kendall's tau Spearman's rho	
Interval	Mean Pearson's correlation	Nonparametric and Parametric Statistics
Ratio	Geometric Mean Coefficient of Variation	

Figure 1: Appropriate statistics for various scale types.

The logic behind the above proscriptions is that statistical measures should remain invariant under the admissible transformations for a particular scale. Stevens [Ste68, Ste62] defines two types of invariance. First, invariance in value, whereby the numerical value of the statistic remains unchanged under the admissible transformations. For example, the Pearson product moment correlation keeps its value under linear transformations (admissible for interval level scales). Second, invariance in reference, whereby the

[5] This test is not truly "distribution free" as all nonparametric statistics are believed to be. For instance, Boneau [Bon62] shows through simulation that the Mann-Whitney U test is more sensitive to distribution differences than is the parametric t test.

value of the statistic may change but it would still refer to the same item or location. For example, the value of the median may change but would still refer to the item at the middle of the distribution under monotonic increasing transformations, and the item at the mean would remain the same under linear transformations even though the value of the mean changes[6].

The above proscriptions are deemed controversial for at least two reasons. First, it is not always obvious that all statistics classified as nonparametric make use of only rank order information. Second, it is quite difficult to determine precisely the scale type of a measure.

An example of a nonparametric statistic that makes use of more than rank order information is the Wilcoxon signed rank test for paired differences (see [SC88]). One of the steps in calculating this statistic involves taking the difference between the scores of the paired observations. As Stevens himself notes [Ste62], this difference would have no meaning if the scores are rankings on an ordinal scale, and therefore an increasing monotonic transformation would change the values of such differences. This means that the nonparametric Wilcoxon signed rank test would not be appropriate for scales deemed to be at the ordinal level. Furthermore, in one of his early papers [Ste46], Stevens states that a rank order correlation statistic (such as Spearman's rho) *"assumes equal intervals between successive ranks and therefore calls for an interval scale"*. However, in other papers he considers that rank correlations are appropriate for ordinal level scales [Ste62], and he considers that rank correlations are appropriate for *both* ordinal and interval level scales in another publication [Ste51]. Therefore, from these articles, it is not clear whether rank order correlation is appropriate or not for ordinal level scales. Caution should then be exercised when following the broad prescriptions or proscriptions that the choice of certain classes of statistics should be based on scale types.

In order to select the most "appropriate" statistics, a researcher has to know the type of scale(s) that s/he is using. The problem is that, in sofware engineering, like in other scientific disciplines, often it is very difficult to *determine* the scale type of a measure. For example, what is the scale type of cyclomatic complexity? Can we assume that the distances on the cyclomatic complexity scale are preserved across all of the scale? This is difficult to say and the answer can only be based on intuition. Despite a few available techniques to help the researchers in particular situations (e.g., extensive structures in [LK88], as discussed in Section 3), the answer to those questions is hardly ever straightforward.

A good example of the confusion about scale types are User Information Satisfaction (UIS) instruments (e.g., see [IOB83]). This kind of subjective measure is often used by researchers in the related discipline of Management Information Systems (MIS) as a surrogate for Information Systems effectiveness [Kim89]. In one article, the authors state that UIS is measured on an *ordinal scale* [GL89]. However, for the kind of scaling model used by this instrument (which is the summative or "Likert" scaling model [MC81]), some authors argue that it produces measures on an *interval scale* [MC81]. What kind of statistics should be used by researchers employing the UIS instrument?

Therefore, there are many cases where researchers cannot demonstrate that their scales are interval, but they are confident that they are more than only ordinal. By treating them as ordinal, researchers would be discarding a good deal of information. Therefore, as Tukey [Tuk86a] notes *"The question must be 'If a scale is not an interval scale, must it be merely ordinal?'"*

Is it realistic to answer questions about scale type with absolute certainty, since their answers always rely on intuition and are therefore subjective? Can we know for sure the scale types of the measures we use? Knowing the scale type of a measure with absolute certainty is out of the question in the vast majority of cases. And in those cases, should we just discard our practical questions — whose answers may have a real impact on the software process — because we are not 100% positive about the scale types of the measures we are using? To paraphrase Tukey [Tuk86b], "Science is not mathematics" and we are not

[6] Here, of course, we assume that there is a specific item at the median and mean values.

looking for perfection and absolute proofs but for *evidence* that our theories match reality as closely as possible. The other alternative, i.e., reject approximate theories, would have catastrophic consequences on most sciences, and in particular, on software engineering. What is not acceptable from a theoretical perspective may be acceptable evidence and even a necessary one from an engineering or an experimental perspective.

It is informative to note that much of the recent progress in the social sciences would not have been possible if the use of "approximate" measurement scales had been strictly proscribed. For example, Tukey [Tuk86a] states after summarizing Stevens' proscriptions *"This view thus summarized is a dangerous one. If generally adopted it would not only lead to inefficient analysis of data, but it would also lead to failure to give any answer at all to questions whose answers are perfectly good, though slightly approximate. All this loss for essentially no gain."* Similarly, in the context of multiple regression, Cohen and Cohen [CC83] state: *"The issue of the level of scaling and measurement precision required of quantitative variables in [Multiple Regression/Correlation] is complex and controversial. We take the position that, in practice, almost anything goes. Formally, fixed model regression analysis demands that the quantitative independent variables be scaled at truly equal intervals ... Meeting this demand would rule out the use of all psychological tests, sociological indices, rating scales, and interview responses ... this eliminates virtually all kinds of quantitative variables on which the behavioral sciences depend."* Even Stevens himself, with respect to ordinal scales, concedes that [Ste46]: *"In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than relative rank-order of data. On the other hand, for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results."*

The above, evidently more pragmatic, view is under-represented in software engineering, however. This stems from the fact that some of the most influential books in our field (the ones considered as standard references on measurement theory in software engineering such as [F91][Z91]) are only presenting one side of the debate, (i.e., the side claiming that scale types should be used to proscribe data analysis techniques).

4.2 A Pragmatic Approach to Applying Measurement Theory

In most cases, the questions to answer (i.e., our measurement goals) determine the scale under which data must be used, and not *vice versa*. One should use the appropriate statistical technique assuming the level of measurement required by the question. If a pattern is detected, then the scientist should start thinking about the validity of the assumption s/he made about the scale types. In addition, it is sometimes possible to use different statistics assuming different scale types and compare the results.

We usually come up with an important question first (e.g., "Is there a linear relationship between two variables?"), relevant to our measurement goals. Subsequently, we usually look around for available data or develop measurement scales that we think can help us answer our questions at a reasonable cost. Also, most of the time, our learning process is exploratory and we have a very limited understanding of the phenomena we are studying, e.g., the impact of component coupling on software defect density. Some important questions require interval or ratio scales but we are not sure if the scales we are using are actually interval or ratio. Should we not analyze the data? For instance, if someone is asking the question: "If I can reduce coupling by 10% through a better design technique, how much would I gain in terms of reduction of defect density?" The answer to this — quite common — kind of question requires a ratio scale for coupling, since the reduction is given in terms of proportions (i.e., ratios), and there is a natural zero level of coupling (when modules are not related to each other). Intuitively, we can be quite sure defect density is defined on a ratio scale too. However, with respect to coupling, the level of measurement of the scale is quite difficult to determine, regardless of the definition used.

For example, if a statistically significant linear relationship is found between coupling and defect density through linear regression then, theoretically, the researcher must start wondering if the computed level of significance is real (or close to reality) since there is some uncertainty with respect to the type of the coupling scale. External information may be examined in order to confirm or otherwise the scale assumption. For example, assuming we want to model defect density, programmers may be surveyed by asking them to score the relative "difficulty" of programs with different coupling levels. If the scores confirm that the distance is, on average, preserved along the studied part of the scale (hopefully, the relevant one for the environment under study), then the equal interval properties may be assumed with greater confidence. In addition, thorough experience and a good intuitive understanding of the phenomenon under study can help a great deal. For example, in a given environment, very often programmers know the common causes of errors and their relative impact. Scales may thus be validated with the help of experts.

However, for the approach presented above, three important issues that need to be addressed are:

1. The researcher may end up promoting his/her measure to the next higher level of measurement. For instance, if a researcher's measure straddles the ordinal/interval boundary, s/he could treat it as an interval scale. What are the consequences of this choice? Suppose that we use a parametric test because we assume a measure of coupling is approximately interval. What is the impact of such approximations on the validity of parametric tests with respect to errors of Type I (i.e., the null hypothesis is falsely rejected). This question is addressed in detail in Section 5, where we demonstrate that commonly used parametric tests are robust to non-linear transformations (but not those that are as extreme as, for example, an exponential transformation) of the interval scale.

2. The researcher may end up not demoting his/her measure to the next lower level of measurement. For example, a researcher has developed a measure of design complexity and wants to validate it by investigating its relationship with some other variable. Unfortunately for the researcher, s/he is unable to demonstrate that his/her measure is on an interval scale, so s/he prudently demotes the measure to the ordinal level. The proscriptions would then dictate that the researcher use an appropriate nonparametric (or distribution free) statistic to validate the measure. What are the consequences of this choice on the researcher's ability to validate the measure? This question is addressed in detail in Section 6, where we demonstrate that nonparametric statistics are, in general, less sensitive than anologous parametric statistics, and therefore the researcher's chances of validating the measure are reduced.

3. During the search for patterns in the data by the researcher, it is frequently necessary to perform transformations. According to measurement theory, certain transformations would be proscribed for certain scale types. What are the consequences of ignoring such proscriptions? This question is addressed in Section 7, where we demonstrate that a proscribed transformation for interval level measures, the logarithmic transformation, has been most useful in the area of developing effort estimation models, and therefore, despite being proscribed, it has demonstrated substantial utility.

5. Robustness of Parametric Tests

Various studies have shown that, most of the time, using parametric tests for scales that are not strictly interval does not lead, except in extreme departures from the interval scale, to wrong statistical decisions, i.e., one rejects the null hypothesis when one should not. Below, we present instances of well known studies in the behavioral sciences where this issue has been investigated. Unfortunately, there are no equivalent studies in software engineering on which we can base our arguments.

5.1 Simulations By Baker et al.

In [BHP66], Baker et al. attempted to test empirically the robustness of the Student t test when used to compare the mean of samples coming from an identical distribution. The question they wanted to answer is stated as follows: "Can we make correct decisions about the nature of reality if we disregard the nature of the measurement scale when we apply statistical tests?" As is the case in software engineering, they note that, most of the time, scales in psychology are somewhere between the ordinal and interval levels of measurement.

The issue of invariance of statistical test results when measurement scales are transformed is a problem similar to the one stated above but can be investigated empirically. The authors decided to consider a scale comprising scores from cardinal numbers 1 to 30. This scale was assumed to be the interval scale of reference for the simulation. Then, populations of 1000 scores each were generated to approximate as closely as possible the expected frequencies for (1) a normal distribution, (2) a rectangular distribution, and (3) an exponential distribution.

A set of 35 non-linear transformations were applied to the reference interval scale. These transformations were however of different nature and the authors claimed they represented cases frequently encountered in psychology. The various categories of transformations may be intuitively described as follows:

- Transformations 1-15: 30 random scores were generated for each transformation within pre-set maximum limits, the first score being 1. For transformations 1-5, 6-10, 11-15, the maximum limits were 2, 10, and 25, respectively.

- Transformations 16-20: larger distortions of the scale are introduced at the extremes and no distortion at the center. More precisely, a possible maximum multiplicative factor of 15 was applied at the extremes down to 1 at the center of the scale.

- Transformations 21-25: Similar to the category above but the maximum multiplicative factor is 45 down to 3 to the center, decreasing by three steps units.

- Transformations 26-35: unit intervals are retained for scores 1 to 15, the other scores being varied randomly.

For each type of distribution and transformation, 4000 pairs of samples (of various sizes such as 5,5 ; 15,15 ; 5,15) were drawn out of a pool of 1000 scores. These pairs of samples were then used to test statistically the difference of their means using a Student t-test, the null hypothesis being that there is no difference between the means. It was assumed that, if the transformations did not have much effect on the decision outcome of the test, then t values should not vary significantly.

Results varied somewhat depending on the type of distribution of the population and the type of transformation. However, they showed that the scale transformations did not affect dramatically the t values resulting from the comparison of sample pairs. The authors concluded that, under the conditions of equal sample sizes and two-tailed t tests, transformations had a neglectible effect on the probability of rejection of the null hypothesis. In addition, t values across transformation types were strongly correlated to t values of the reference interval scale.

Townsend and Ashby

In [TA84], Townsend and Ashby’s main criticism of Baker et al [BHP66] is that "in general we have no idea as to the degree of transformation that may occur in nature." In other words, the authors think that some ordinal scale may in fact be an extreme distortion of an interval scale without the scientist realizing it. They point out that Baker et al's study does not explore "monotonic transformations that stretch or

shrink one part of the scale more than another **..."** (The result of such a scale transformation will be denoted as a “dichotomous” scale.[7]) They conclude by saying that such simulations cannot be representative of real-life robustness of parametric statistics if one does not have knowledge about the kind of distortions that may actually occur when measuring the phenomenon under study. However, as discussed below, we think that empirical software engineers usually understand the problem at hand well enough to assess whether the scale of a measure is closer to being “dichotomous” or more towards interval. It would be, on the other hand, extremely difficult to determine exactly if a scale is strictly interval.

5.2 Simulations by Labovitz

In one study by Labovitz [Lab70], the author states that "Although some small error may accompany the treatment of ordinal variables as interval, this is offset by the use of more powerful, more sensitive, better developed, and more clearly interpretable statistics with known sampling errors." In this sociological study, the relationship between "occupational prestige" and suicide rates was studied. The former variable was based solely on the principle of ordinal ranking. Based on the original prestige ratings resulting from the study, various scoring systems were generated from a computer by randomly assigning numbers between 1 and 10,000 such that the ordinal ranking of the original ratings is preserved. In addition, any ties in the ordinal ranking were assigned identical numbers.

First, the Pearson intercorrelations between scoring systems, assuming in turn that each scoring system is the "true one," were computed. The results indicated consistently high correlations showing a high degree of interchangeability among the 20 scoring systems. All were above 0.9 and 157 out of 190 were above 0.97. Also, results showed the correlation between the various scoring systems and suicide rates did not vary significantly from a scoring system to another.

In addition the study showed that the greater the number of ranks N, the greater the confidence in assigning an interval scoring to ordinal data. In other words, it was observed that the standard deviations among the correlation coefficients decrease as N increase.

The author however ponders his position by acknowledging that these results may not hold for "extreme" nonlinear monotonic transformations of ordinal measures. There exists a point beyond which measures are not interchangeable.

One additional argument for using interval level statistics is that they offer well-developed and interpretable multivariate analysis techniques, e.g., multivariate regression analysis, and principal components analysis [DG84]. The author, in order to support this point, showed an example where the additive combination of occupational prestige, income and education results had a much stronger impact on suicide rates than each of them considered independently. Treating occupational prestige as ordinal would not have allowed this analysis and these highly suggestive conclusions could not have been drawn.

In summary, the following conclusion is drawn by the author: "certain interval statistics can be given their interval interpretation with only negligible error if the variable is nearly interval."

When analyzing ordinal (or below interval level) variables, the following research strategies may be adopted:

(1) Assign a linear scoring system according to existing evidence or knowledge about distances between ranks,

---

[7] In such extreme cases, Labovitz (see Section 5.2) denotes that phenomenon as a “dichotomy” of scale. It comes from the fact that, in that situation, scores will be clustered together on a part of the scale far away from the other scores. This will virtually “split” the scale into different clusters of scores.

(2) Use all available rank categories without collapsing them into a smaller number,

(3) Report the actual scale of your data and interpret interval statistics with care by performing further exploration and tests to confirm or otherwise distance-related assumptions.

In an earlier independent study, Labovitz [Lab67] investigated the difference between two (hypothetical) therapies and the impact of interval scale transformations on the study of these differences. Patients having undertaken the therapies provides a rank from one to 4 according to the level of effect they think therapy had on them. The author then assigned a random number between 1 and 10 to each of the possible ranks seven times, thereby generating 7 scoring systems for the subjective response categories. Consistently with the study above, Labovitz showed that the differences in scoring systems had little effect on Point-Biserial correlation coefficients (i.e., a correlation coefficient between a dichotomous and a continuous variable [Nun78]) between the therapies and the subjective scoring of patients. Similarly, he shows that test of differences between means for the two therapies are not significantly affected by the variation in scoring systems. His conclusions are similar to the ones presented above. Even though he suggests caution when interpreting results obtained when using such parametric statistics with ordinal scales, he claims that parametric statistics should be used in a larger number of situations, especially when there is evidence or belief that the scale is "near-interval." Thus, no information about the data (i.e., about distances) is wasted and more sensitive/powerful techniques (see next Section) can be used. In that case, it is advised that the researcher should analyze the problem at hand carefully by looking at the robusteness of the statistics he/she wishes to use, how wrong is the interval scale assumption, and the power/sensitivity of alternative techniques considering the size of the data set and other factors described in the next section.

Mayer

However, in a response to Labovitz, Mayer [May71] claims that under certain circumstances, treating ordinal data as interval can be disastrous. Assuming one wished to compute a linear correlation between two variables Y1 and Y2, Mayer shows that if the real interval scale is Log Y2, then such an exponential distortion will lead the scientist to strongly underestimate the relationship between the two variables. He shows that the degree of underestimation depends on the variance on the Log Y2 scale. The larger the variance, the larger the underestimation. However, Mayer's argument was disclaimed by Labovitz [Lab71] as being extreme because such an exponential distortion nearly “dichotomizes” the scale (i.e., clusters the scores in one part of the scale to the extent where they are almost not distinguishable and streches the other part of the scale, thereby “splitting” it) and therefore, as discussed in [Lab70], the author recommends great care in treating the data as interval if scale dichotomization is suspected for any reason.

5.3 Summary

The above simulations seem to indicate that, to the extent that they are not “extreme,” nonlinear transformations do not strongly affect usual statistics such as the t-test or correlation coefficients. In addition, the counter-arguments of their opponents do not appear to apply in most situations. Such results lead us to believe that parametric statistics are applicable in a larger number of circumstances than what was originally thought. It is important to remember, however, that the above simulation studies only present a partial answer to our questions and that further studies (if possible representative of classical software engineering contexts) are needed. On the other hand, one could question the usefulness of parametric statistics in such situations since non-parametric statistics are available and safer to use. This issue is partially discussed in the next section where we show that using non-parametric statistics may lead to a loss of information and power (i.e., "sensitivity") with respect to statistical inference. Other

arguments related to model interpretation and variable interactions could be developed to support the use of parametric tests. However, we have chosen to focus on only one of the main points and to provide substantial evidence to support it.

6. Power of Nonparametric Tests

The choice of a nonparametric statistic over an analogous parametric statistic will, in general, result in some loss of statistical power. For example, if a researcher is validating a design complexity metric, then less power means that the probability of successfully validating the metric is reduced even if the metric is valid. To demonstrate this, we first present a brief overview of the concept of statistical power.

6.1 Statistical Power

When analyzing their data, researchers often state (before, during or after data collection) a null hypothesis which they hope to reject. For instance, the null hypothesis may be that a design complexity metric is not related to maintenance effort. When testing this null hypothesis using inferential statistical procedures, researchers run the risk of two types of errors. A Type I error is the incorrect rejection of the null hypothesis. The probability of committing a Type I error is represented by the level of significance, α. For instance, if α=0.05, then the researcher is running a 5% risk of incorrectly rejecting the null hypothesis if s/he were to repeat his/her study a large number of times. A Type II error is the acceptance of the null hypothesis when in fact it is false. The probability of committing a Type II error is expressed by β. The value $1-\beta$ is the probability of correctly rejecting the null hypothesis, which is statistical power. For example, if statistical power of a given inference test is 0.8, then there is a 0.8 probability that the statistical test will reject the null hypothesis if there is a relationship between the measure of design complexity and maintenance effort.

Statistical power is closely related to three other parameters: (a) sample size, (b) α, and (c) magnitude of the effect (e.g., the magnitude of the correlation coefficient). Intuitively, the following statements characterize the relationships between these parameters and power (in each case assuming all other parameters are held constant):

- the greater the power required, the greater the necessary sample size;
- the smaller the sample size, the smaller the power;
- the more stringent the α - level, the less the power;
- the greater the magnitude of the effect size, the less power is necessary; and
- the greater the power, the less the necessary magnitude of the effect size.

In practical terms, researchers should use the most powerful tests available to them[8]. For a given sample size, the greater the power then the more sensitive the test and therefore the greater the probability of finding significant results. Also, the greater the statistical power, the smaller the sample size necessary for a study, and hence the greater the feasibility of conducting the study (for example, because of reduced costs or because of the greater availability of data). This is very important for software engineering, where we cannot always expect to have huge amounts of data, and where the cost of data collection can be quite substantial.

6.2 Nonparametric vs. Parametric Tests

In general, parametric tests are more powerful than their nonparametric counterparts. A number of simulation and analytical studies have demonstrated this point. For instance, a simulation study by

[8] Of course, one should also check the distributional and other assumptions of the test to determine its appropriateness for the task at hand, as well as the test's robustness to violations of these assumptions.

Boneau [Bon62] compared the empirical power of the parametric t test with the nonparametric Mann-Whitney U test. In the case of a normal distribution for both samples, equal variances, and equal group sizes, the t test clearly showed superior power for the two-tailed test at a nominal α-level of 0.01. This remained unchanged as the sample sizes were increased. When the two groups are of different sizes, the t test has greater power for one and two tailed tests at various α levels.

Now, if we return to our researcher, s/he may choose to use correlation coefficients to investigate the relationship between the design complexity metric and maintenance effort. Figure 2 shows the sample sizes necessary for different magnitudes of two commonly used correlation coefficients, one parametric (Pearson's product moment correlation) and one nonparametric (Spearman's rank correlation)[9,10,11]. As can be seen, for a specified level of power, Spearman's correlation always requires a larger sample size than Pearson's coefficient (approximately 20% larger for strong relationships). In general, for *large samples*, to achieve the same power as the Spearman correlation, a test using Pearson's coefficient would require only approximately 91% of the former's sample size [SC88]. This is called the asymptotic relative efficiency (ARE) [Gib71]. The ARE for various nonparametric tests[12] and their parametric counterparts are shown in Figure 3.

	Power = 90%			Power = 80%		
Corr.	**Pearson**	**Spearman**	**% Difference**	**Pearson**	**Spearman**	**% Difference**
0.1	854	1013	84%	618	733	84%
0.2	212	250	85%	154	183	84%
0.3	93	107	87%	68	79	86%
0.4	51	62	82%	38	46	83%
0.5	32	39	82%	24	30	80%
0.6	21	26	81%	16	20	80%
0.7	15	19	79%	12	15	80%

Figure 2: Minimal sample sizes required for Pearson's and Spearman's correlations for two levels of power (90% and 80%) at one tailed alpha = 0.05.

Non-parametric Test	Analogous Parametric Test	Asymptotic Relative Efficiency
Wilcoxon Signed Ranks Test	t-test	95.5%
Wilcoxon-Mann-Whitney Test	t-test	95.5%
Friedman 2-way ANOVA	F-test	64% (for k=2)[13], 72% (for k=3)
Kendall's Tau	Pearson's r	91%

Figure 3: Asymptotic Relative Efficiency for various nonparametric tests.

Under the violation of the assumptions of parametric tests, their power levels do not change markedly, and they retain, *in general*, their greater power when compared to the nonparametric tests [Coh65]. For example, one simulation study concluded that the power of the t-test remains essentially unchanged when the homogeneity of variances assumption is violated [Bon62]. Furthermore, the nonparametric Mann-Whitney U test remains less powerful than the parametric t test when the above assumption is violated.

[9] The values in this table are based on the tables provided in [KT87] and [Coh88].

[10] The calculations of sample sizes assume that the assumptions of the tests are met.

[11] Where there are analogous tables in [Coh88], the sample size values are only slightly different from [KT87] (approximately ± 2 difference).

[12] These values were obtained from [SC88][Gib71][Gib93a][Gib93b].

[13] Where k is the number of treatments.

When sampling from nonnormal distributions (e.g., rectangular and exponential distributions), it was also found that the t test has more power than the Mann-Whitney U test[14].

Moreover, a review of published empirical research in the related discipline of Management Information Systems (MIS) [BO89] found that the power of parametric tests (such as regression and correlation) was always higher than for conventional nonparametric tests (such as the Mann-Whitney U test and the Wilcoxon test)[15,16]. This means that, for commonly used sample sizes in MIS research, the parametric tests are more powerful.

Returning to our researcher, if the available sample size is not large enough to attain a reasonable level of power, the nonparametric test chosen will likely not lead to the rejection of the null hypothesis. The researcher then concludes that his metric is not valid[17]. Given that the researcher could not produce evidence demonstrating the validity of the new metric, the chances of an acceptable publication are reduced remarkebly, and the researcher would be better off moving to greener pastures and developing another metric that can be validated. Of course, for a given sample size and α-level, the researcher had a greater probability of rejecting the null hypothesis (and finding his metric to be valid) had s/he used a parametric counterpart[18].

To summarize, our point was to demonstrate the costs of resorting to nonparametric statistics. If the choice of nonparametric statistics is driven by the proscriptions of measurement theorists, then one should be very careful in weighing the loss of power consequences. However, if the choice is driven by *extreme* violations of the parametric statistics' assumptions, then one would be justifiable in using nonparametric statistics. Despite the clear advantages of non-parametric statistics, it appears that, from a practical perspective, parametric statistics have a larger realm of application than originally thought.

7. Scale Transformations

During data analysis and empirical model building, it is very common for researchers to transform their data. There are many reasons why researchers would choose to transform their data. For example, to make nonlinear relationships more linear or to make the data more congruent with the assumptions of a data analysis technique (for instance, to address heteroscedasticity in regression analysis) [CC83]. Thus, equation parameters can be more easily estimated, models can be interpreted in a more straightforward manner, and interpolation is made easier [MT77].

According to the principles of measurement theory, there are admissible transformations applicable to each scale type (see Section 2). An examination of these transformations clearly indicates that nonlinear

---

[14] The power of the t-test is slightly less than that of the Mann-Whitney U test when sampling from populations whose distributions are different (e.g., normal and exponential); however, this is not the case as the sample sizes of the two groups is increased (from 5 to 15) [Bon62].

[15] For this comparison, the Effect Size was kept constant by classifying it into one of three groups defined by Cohen [Coh88]: "small", "medium", and "large".

[16] The list of nonparametric statistics excluded the common Chi-Square statistic.

[17] When null hypotheses are not rejected, few software engineering researchers consider the lack of statistical power as a possible contributing factor.

[18] This assumes that the effect sizes would be the same for the parametric and nonparametric tests (i.e., that Pearson's r and Spearman's rho will give the same coefficient). If this is not the case, for example, if with the nonparametric test the effect size is larger, then the nonparametric test may lead to the rejection of the null hypothesis.

transformations would not be admissible for interval and ratio level scales. In the context of software engineering, this has some serious implications.

A very common transformation used in software engineering is the logarithmic transformation. This is applied frequently in the construction of effort estimation models using linear regression. As has been noted by Bailey and Basili [BB81] and Basili [Bas80], a general form of such models is:

$$E = a\, L^b$$

where:

E = effort
L = some measure of size (usually LOC)
a, b = constants

Examples of this kind of model include the one developed by Walston and Felix [WF77]:

$$E = 5.2\, L^{0.91}$$

and the one developed at the NASA SEL [Bas80]:

$$E = 1.4\, L^{0.93}$$

As is common, when using ordinary least squares regression to develop models of this general form, an analyst would use the following estimating equation:

$$\ln E = \ln a + b \ln L$$

Thus, using the above equation, ordinary least square regression estimating formulas could a priori be used to estimate, in a straightforward manner, the parameters *a* and *b*. However, it is clear from the equation above that both the effort variable and the size variable are transformed using the nonlinear logarithmic transformation. Since we know that these transformations (i.e., E*=ln E and L* = ln L) are not admissible if these two variables are measured on an interval or ratio scale, then *ln E* and *ln L* cannot be used as if they were defined on an interval or ratio scale. Therefore, following strictly the precepts of measurement theory, we are not allowed to use such scale transformations and still use linear regression analysis.

This naturally leads to the conclusion that all effort estimation models of this general form that use either or both of the above variables and that use linear regression for model construction are meaningless[19]. They are meaningless because the transformations that are used are not admissible according to measurement theory. However, many software engineering researchers and practitioners would argue that these models are being used to good effect and are useful, irrespective of the measurement theoretic perspective. Furthermore, the purpose of a transformation is to effect the resultant model. If the model parameters are invariant, which would be the case if only admissible transformations were performed, then why bother with the transformation? For example, on an interval scale, the only permissible transformations are linear ones. Linear transformations will result in the same R^2 value. If one wanted to improve the goodness of fit of the model to the data, it does not make sense to use a permissible transformation because it will have no effect on goodness of fit.

[19] An alternative conclusion would be that effort and size are all measured on an ordinal scale. However, it would hardly be acceptable for, say, a project manager to consider effort to be on an ordinal scale.

In general, not being able to use nonlinear transformations, such as the logarithmic transformation, is a substantial hindrance to the data analyst. Given that there is even a push towards higher scale levels (e.g., interval and ratio) in software engineering, this means that we would be less able (or permitted) to use these nonlinear transformations because they would then not be applicable for our scale types. This would limit the tools available to contemporary researchers and practitioners to empirically model software engineering phenomena. Clearly, this is not an acceptable state of affairs, and therefore, we suggest that a more pragmatic perspective should be adopted.

8. Conclusions

In this paper, we have shown that, in most cases, the interpretation and application of measurement theory in our field is too rigid and even questionable. Furthermore, numerous studies in other fields show that, for the last fifty years, there has been a very interesting and important debate on the issues related to measurement theory and its applications. Arguments on various sides show that there are divergent points of view which are either not known or not considered by the software engineering community.

The points that we have discussed in this paper in support of our position are:

1. There is little basis for mandating that software complexity measures should be additive and that they should assume an extensive structure.

2. It is not obvious that all nonparametric statistics make use of only rank-order information in their calculation. Therefore, caution should be exercised even when strictly following the commonly found proscriptions on the use of statistical procedures.

3. In software engineering, as in other disciplines, it is frequently difficult to know for certain the scale type of a measure. This is true even for well known measures such as cyclomatic complexity.

4. The use of parametric statistics is more risky than the use of nonparametric statistics in a field where measurement scales are not well understood. However, with care and after thorough reflection, such a risk appear to be worth taking when one considers the following facts:

    - Many common parametric techniques (e.g., product moment correlation, t-test) are robust to non-linear distortions of interval scales as long as they are not "too extreme", e.g., exponential. If this is the case, these statistics will have a tendency to be conservative and underestimate existing relationships or differences. However, in many cases, even though it is difficult to judge whether we have an interval scale or not, we can reasonably determine if a scale is closer to being exponential or to being a near-linear distortion of an interval scale. For example, even though we cannot be sure that cyclomatic complexity is interval, few of us would claim that it is an exponential distortion of an interval scale. Based on our intuitive idea of control flow complexity, it is fair to say that most of us feel that the distance between scores (number of independent paths in a control flow graph) on the cyclomatic complexity scale is approximately equivalent (i.e., near-linear scale) and does not increase/decrease exponentially.

    - Parametric techniques are, in general, more powerful (i.e., more sensitive) than nonparametric statistics. An increased use of the less powerful nonparametric statistics may lead to: (a) a reduction in empirical research because of the infeasibility of the larger sample sizes required to attain reasonable power with nonparametric statistics, and/or (b) an increase in “non-significant’ findings which may cause many interesting relationships being overlooked. Both of the these consequences are, of course, undesirable.

Considering the state of the art and even though measurement scales are extremely useful concepts, we believe they should not be used to broadly proscribe statistical techniques but should serve as useful

concepts to assess (the lack of) observed patterns in a data set. More precisely, measurement scales are useful to better evaluate how valid and credible the statistical level of significance of an observed pattern is, e.g., difference of means between samples or relationships between variables. Or, if there is no pattern, one may wonder if the scale was inadequate, or the statistical test was not powerful enough (see Sections 6.1and 6.2). However, the search for a pattern or trend in one's data should not be sacrificed or ignored because of doubts about the scale type of a measure.

In that context, the software engineering community should remain informed of the debate that is going on in the statistics and the social science communities. Above all, we should not accept measurement theory prescriptions and proscriptions as though they were absolute (almost religious!) and unquestionable. Pragmatism and common sense, combined with a reasonable dose of rigour, should always prevail.

9. Acknowledgements

We would like to thank all the people who helped us improve both the content and form of this paper: Victor Basili, Warren Harrison, Guy Lafond, Filippo Lanubile, Denis St-Pierre, and Norman Schneidewind.

10. References

[AKD+81] F. Andrews, L. Klem, T. Davidson, P. O'Malley, and W. Rodgers: *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*, Institute for Social Research, University of Michigan, 1981.

[BB81] J. Bailey and V. Basili: "A Meta-Model for Software Development Resource Expenditures." In *Proceedings of the International Conference on Software Engineering*, pages 107-116, 1981.

[BHP66] B. Baker, C. Hardyck, and L. Petrinovich: "Weak Measurements vs. Strong Statistics: An Empirical Critique of S. S. Stevens' Proscriptions on Statistics." In *Educational and Psychological Measurement*, 26:291-309, 1966.

[Bas80] V. Basili: "Resource Models. " In *Tutorial on Models and Metrics for Software Management and Engineering*, IEEE Computer Society Press, V. Basili (ed.), 1980.

[BO89] J. Baroudi and W. Orlikowski: "The Problem of Statistical Power in MIS Research." In *MIS Quarterly*, pages 87-106, 1989.

[BO94] J. Bieman and L. M. Ott: "Measuring Functional Cohesion." In *IEEE Trans. Software Eng.*, 20(8): 644-657, August 1994.

[B84] P. Bollman: "Two Axioms for Evaluation Measures in Information Retrieval." In *Research and Development in Information Retrieval*, ACM, British Computer Society Workshop, Series, pp. 233-246, 1984.

[Bon62] C. Boneau: "A Comparison of the Power of the U and t Tests." In *Psychological Review*, 69(3):246-256, 1962.

[BMB94] L. Briand, S. Morasca, and V. Basili: "Property Based Software Engineering Measurement." *Technical Report*, CS-TR-119, University of Maryland, November 1994.

[CK94] S. R. Chidamber and C. Kemerer: "A Metrics Suite for Object Oriented Design." In *IEEE Trans. Software Eng.*, 20(6): 476-493, June 1994.

[Coh65] J. Cohen: “Some Statistical Issues in Psychological Research.” In *Handbook of Clinical Psychology*, B. Woleman (ed.), McGraw-Hill, 1965.

[Coh88] J. Cohen: *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 1988.

[CC83] J. Cohen and P. Cohen: *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 1983.

[DG84] W. Dillon and M. Goldstein: *Multivariate Analysis: Methods and Applications*, Wiley & Sons, 1984.

[F91] N. Fenton: *Software Metrics: A Rigorous Approach*, Chapman & Hall, 1991.

[F94] N. Fenton: “Software Measurement: A Necessary Scientific Basis.” In *IEEE Transactions on Software Engineering*, 20(3):199-206, March 1994.

[GL89] D. Galletta and A. Lederer: “Some Cautions on the Measurement of User Information Satisfaction.” In *Decision Sciences*, 20:419-438, 1989.

[Gar75] P. Gardner: “Scales and Statistics.” In *Review of Educational Research*, 45(1):43-57, Winter 1975.

[Gib71] J. Gibbons: *Nonparametric Statistical Inference*, McGraw-Hill, 1971.

[Gib93a] J. Gibbons: *Nonparametric Statistics*, Sage Publications, 1993.

[Gib93b] J. Gibbons: *Nonparametric Measures of Association*, Sage Publications, 1993.

[IOB83] B. Ives, M. Olson, and J. Baroudi: “The Measurement of User Information Satisfaction.” In *Communications of the ACM*, 26(10):785-793, October 1983.

[Kim89] K. Kim: “User Information Satisfaction: Toward Conceptual Clarity.” In *Proceedings of the 11th International Conference on Information Systems*, pages 183-191, 1989.

[KT87] H. Kraemer and S. Thiemann: *How Many Subjects? Statistical Power Analysis in Research*, Sage Publications, 1987.

[K71] D. Krantz, R. Luce, P. Suppes, and A. Tversky: *Foundations of Measurement*, Vol. 1, Academic Press, 1971.

[Lab67] S. Labovitz: “Some Observations on Measurement and Statistics.” In *Social Forces*, 46(2):151-160, December 1967.

[Lab70] S. Labovitz: “The Assignment of Numbers to Rank Order Categories.” *In American Sociological Review*, 35:515-524, 1970.

[Lab71] S. Labovitz: “In Defense of Assigning Numbers to Ranks.” In *American Sociological Review*, 36:521-522, 1971.

[LK88] R. Luce and C. Krumhansl: “Measurement, Scaling, and Psychophysics.” In *Stevens’ Handbook of Experimental Psychology*, Wiley, 1988.

[May71] L. Mayer: “A Note on Treating Ordinal Data as Interval Data.” In *American Sociological Review*, 36:519-520, 1971.

[MC81] J. McIver and E. Carmines: *Unidimensional Scaling*, Sage Publications, 1981.

[M86] J. Michell: “Measurement Scales and Statistics: A Clash of Paradigms”. *In Psychological Bulletin*, 100(3):398-407, 1986.

[MT77] F. Mosteller and J. Tukey: *Data Analysis and Regression*, Addison-Wesley, 1977.

[Nun78] J. Nunnally: *Psychometric Theory*, McGraw-Hill, 1978.

[Ovi80] E. Oviedo: “Control Flow, Data Flow, and Program Complexity”. In *Proceedings of COMPSAC*, pp. 146-152, November 1980.

[R79] F. Roberts*: Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley, 1979.

[SC88] S. Siegel and J. Castellan: *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, 1988.

[Ste46] S. Stevens: “On the Theory of Scales of Measurement.” In *Science*, 103(2684):677-680, June 1946.

[Ste51] S. Stevens: “Mathematics, Measurement, and Psychophysics.” *In Handbook of Experimental Psychology*, S. Stevens (ed.), John Wiley, 1951.

[Ste62] S. Stevens: “Measurement, Psychophysics, and Utility.” In *Measurement: Definitions and Theories*, C. Churchman and P. Ratoosh (eds.), John Wiley, 1962.

[Ste68] S. Stevens: “Measurement, Statistics and the Schemapiric View.” In *Science*, 161:849-856, 1968.

[SZ63] P. Suppes and J. Zinnes: “Basic Measurement Theory”. *In Handbook of Mathematical Psychology*, Vol. 1, R. Luce, R. Bush, and E. Galanter (eds.), John Wiley, 1963.

[TA84] J. Townsend and F. Ashby: “Measurement Scales and Statistics: The Misconception Misconceived.” In *Psychological Bulletin*, 96(2): 394-401, 1984.

[Tuk86a] J. Tukey: “Data Analysis and Behavioral Science or Learning to Bear the Quantitative Man’s Burden by Shunning Badmandments.” In *The Collected Works of John W. Tukey*, Vol. III, Wadsworth, 1986.

[Tuk86b] J. Tukey: “The Future of Data Analysis.” In *The Collected Works of John W. Tukey*, Vol. III, Wadsworth, 1986.

[VW93] P. Velleman and L. Wilkinson: “Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading.” In *The American Statistician*, 47(1):65-72, 1993.

[WF77] C. Walston and C. Felix: “A Method of Programming Measurement and Estimation.” In *IBM Systems Journal*, 1:54-73, 1977.

[W88] E. Weyuker: “Evaluating Software Complexity Measures.” In *IEEE Transactions on Software Engineering*, 14(9):1357-1365, 1988.

[Z91] H. Zuse: *Software Complexity: Measures and Methods*, de Gruyter, 1991.

[Z92] H. Zuse: “Measuring Factors Contributing to Software Maintenance Complexity.” In *Proceedings of the 2nd International Conference on Software Quality*, Triangle Research Park, NC, October 1992.

[Z94] H. Zuse: "Software Complexity Metrics/Analysis." In *Encylopedia of Software Engineering*, J. Marciniak, (ed.), Volume I, pp. 31-166, John Wiley & Sons, 1994.

[Z95] H. Zuse an T. Fetcke: “Properties of Object-Oriented Software Measures.” *In Proceedings of the Annual Oregon Worshop on Software Metrics*, 1995.