Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Chapter 2

A Reanalysis of Lord's Statistical Treatment of Football Numbers

'I checked it very thoroughly,' said the computer, 'and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.'

Douglas Adams - The Hitchhiker's Guide to the Galaxy

Abstract

Stevens' theory of admissible statistics (1946) states that measurement levels should guide the choice of statistical test, such that the truth value of statements based on a statistical analysis remains invariant under admissible transformations of the data. Lord (1953) challenged this theory. In a thought experiment, a t-test is performed on football numbers (identifying players: a nominal representation) to decide whether a sample from the machine issuing these numbers should be considered non-random. This is an inadmissible test, since its outcomes are not invariant under admissible transformations for the nominal measurement level. Nevertheless, it results in a sensible conclusion: the number-issuing machine was tampered with. The present aim is to show that the thought experiment contains a serious flaw. First, the assumption that the numbers are nominal is shown to be false. Second, it is argued that the football numbers represent the amount of bias in the machine. The level of bias in the machine, indicated by the population mean, conforms to a bisymmetric structure, which means that it lies on an interval scale. In this light, Lord's thought experiment interpreted by many as a problematic counterexample to Stevens' theory of admissible statistics conforms perfectly to Stevens' dictum.

2.1 Admissible Statistics

In a typical introductory statistics class, psychology students are taught that the level of measurement should be taken into account when choosing a statistical test. For example, a t-test should not be performed on data that are of a nominal or ordinal level. Exactly why this rule should be followed is rarely explained and not widely known among psychologists; therefore the rationale for it is reiterated. Suppose mathematical proficiency of children was measured on an ordinal level. In such a case, one is justified in transforming the data by taking the square for example, because the ordinal property in the data, the original ordering of the children, is preserved. For an ordinally measured property, all monotonically increasing, or order-preserving transformations of the data are completely equivalent in their representational capacities, i.e. they all represent the measured property equally well. In fact, in measurement theory this is the defining feature of scale levels (Krantz, Luce, Suppes, & Tversky, 1971).

With respect to the results of parametric statistical analyses that may be executed on the differently transformed data, however, no such equivalence exists. For instance, it is possible that when scores on the aforementioned mathematical proficiency test are analyzed for sex differences with a t-test, different results are obtained for the original and transformed scores. Boys may significantly outperform girls when analyzing the original scores, while boys and girls may not differ significantly in their performance when analyzing the transformed scores (or vice versa; see Hand, 2004, for some interesting examples). Since there is no sense in which the original scores are preferable or superior to the transformed, squared scores, this means that research findings and conclusions depend on arbitrary, and usually implicit, scaling decisions on part of the researcher. This hinders scientific progress because it obscures a factor, namely the choice of scaling, that is influential in determining conclusions based on empirical research. It is important, therefore, to have a clear understanding of how level of measurement can affect our conclusions.

Stevens (1946) introduced the concept of measurement levels and rules for choosing a statistical test according to the dependent variable's measurement level in order to prevent arbitrary scaling decisions from affecting research outcomes. His theory has become known as the theory of admissible statistics (e.g., Robinson, 1965). The basic idea is that one should only perform statistical tests that yield conclusions that are invariant under all so-called ‘admissible transformations’ (admissible in the purely technical and non-pejorative sense of having the same representational capacities) of the data. These admissible transformations are one-to-one transformations for nominal data, order-preserving transformations for ordinal data, positive linear transformations for interval data and transformations that multiply with a positive constant for ratio data (see Krantz et al., 1971; Suppes & Zinnes, 1963). Using admissible statistics ensures that conclusions about the measured property are not dependent on numerical values arbitrarily chosen to encode the data. Thus, in respecting the measurement level, one removes a threat to the validity of the conclusions, which is a good thing.

2.1.1 The Measurement-Statistics Debate

On first sight, one may expect this simple fact to be universally appreciated as a powerful insight into the relation between measurement and statistics. Therefore, it may be considered surprising that the concept was not uniformly welcomed by statisticians, some of whom vehemently rejected the suggestion that measurement levels could have any bearing on data analysis. Several arguments have been adduced in support of this view. Some argue that the level of measurement is often very hard to determine. Velleman and Wilkinson (1993) maintain that in real situations, data do not always fall neatly into the scale levels describes by Stevens (1946), which is problematic when applying his rules. Others have argued that theoretically inadmissible statistics can be arbitrary, but that in practice they rarely are (Baker, Hardyck, & Petrinovich, 1966). That is, inadmissible parametric statistics tend to agree with their more cumbersome admissible counterparts, so that real harm is rarely done by executing statistical analyses that are strictly inappropriate.

These arguments against the idea of admissible statistics are pragmatic in character (i.e., determining the level of measurement is too hard, parametric analyses are easier to use, etc.). Hence it may seem as if substantial agreement exists among scholars regarding the general validity of Stevens’ ideas, even if adhering to this idea is not generally advisable. This is not the case, however. Some are of the opinion that there is a principled problem with Stevens’ view and think that there exists no connection between the levels of measurement and the validity of results attained through statistical analyses at all. Statistical tests simply help the researcher to decide whether their data – in the form of a set of numbers – is likely to be a random sample drawn from a larger population of numbers having a specific distribution. What the numbers measure is irrelevant to this particular decision, and hence issues concerning the level of measurement are irrelevant as well (Burke, 1953; Gaito, 1980).

2.1.2 Lord's Controversial Argument

One of the most important arguments supporting this 'statistical' perspective was provided by Lord (1953). He is uniformly cited among opponents of the theory of admissible statistics (Gaito, 1960; Anderson, 1961; Baker et al., 1966; Gaito, 1980; Velleman & Wilkinson, 1993; Kampen & Swyngedouw, 2000; Harwell & Gatti, 2002; Pell, 2005). Lord introduces a thought experiment in which the use of an inadmissible parametric test on nominal numbers leads to a legitimate, useful and seemingly non-arbitrary conclusion. Lord thus appears to present a clear counterexample to Stevens' theory of admissible statistics. In doing so, he lends support to the view that levels of measurement should not influence one's choice of statistical analysis. Lord's argument sparked what is now called the measurement-statistics debate, and must be considered the most influential and certainly the most entertaining critique of the theory of admissible statistics to date. As such, the two-page letter about a nutty statistics professor has become something of a locus classicus in the literature on psychological measurement and statistics.

Notwithstanding its rhetoric force, Lord's contribution was severely criticized. Some responded by clarifying the basic principles of Stevens theory, giving examples where computations on nominal data lead to absurd conclusions (Behan & Behan, 1954; Bennet, 1954; Stine, 1989). Others pointed out that although computations can be performed on a nominal variable, the results have no reference to the empirical world and so they are irrelevant (Townsend & Ashby, 1984). In our view, however, most of the published criticisms have not gotten to the heart of the matter, in that they fail to explain why the conclusion in Lord's thought experiment is useful, while at the same time the statistical test is inadmissible.

The goal here is to unravel this problem, and to show that Lord's thought experiment does not provide a valid counterexample to Stevens' theory of admissible statistics. We start by revisiting Lord's thought experiment in detail. After analyzing an important, but implicit assumption, it is argued that the validity of the thought experiment hinges on the question of what property the numbers represent in relation to the statistical question that is asked. It is then shown that in relation to this statistical question, the numbers clearly do not represent a nominal property.

This conclusion is enough to disqualify Lord's thought experiment as a valid counterexample to Stevens' theory. Additionally however, it is possible to identify a property that actually is relevant to the statistical question. The structure of this property and its measurement level is explored. It is argued that the data can represent this newly identified property on an interval level, which provides a genuinely new outlook on Lord's thought experiment. For in this new light, Lord's thought experiment is not a counterexample, but instead a perfect illustration of Stevens' theory of admissible statistics. Related arguments made by critics of the theory of admissible statistics are addressed, and it is argued that statistics and measurement cannot be viewed separately when one wants to make meaningful inferences about the properties that one intends to measure. Finally, ways are discussed in which researchers can deal with the implications of our discussion of Lord and incorporate decisions about measurement levels in their research.

2.2 The Treatment of Football Numbers

2.2.1 The Nutty Professor

Lord (1953) describes a university professor who loves to compute means and standard deviations of his students' grades, which measure proficiency on an ordinal level only. He knows this is against Stevens' rules and he feels so guilty that he goes into early retirement. Instead of a gold watch, the university gives him an enormous amount (a hundred quadrillion) of two-digit cloth numbers and a vending machine. He can sell these numbers to the football teams, so they can use the numbers to distinguish players on the field. In somewhat oblique terms, one could say that the numbers 'measure' the uniqueness of the players, obviously on a nominal level. After making an inventory of the numbers, the professor shuffles them, puts them in the vending machine, and sells a large pile of numbers (1600 to be exact), first to the sophomore team and then to the freshman team. After a few days the freshmen come back with a complaint. The sophomores have been making fun of them for having received lower numbers. The professor now faces a problem: The freshmen receiving lower numbers could be either a coincidence or the result of foul play. The professor decides to ask a statistician for help.

The statistician computes means and standard deviations for the population and the freshmen sample, computes a critical ratio test statistic, which is essentially a one group t-test comparing the sample mean to the population mean using the population standard deviation. He then applies Chebyshev's inequality (since the population is not normally distributed) and finds a very small p value. The professor of course protests heavily that such a test is inadmissible for a variable measured on a nominal level. The statistician responds by challenging him to draw new samples from the vending machine and to see how many times he finds a mean equal to, or lower than the freshman mean. The professor does so many times and finds only two such values. He is now satisfied that the machine was tampered with and provides the freshmen with new numbers. He is so heartened by this meaningful use of a parametric test on a variable measured on a nominal level, that he decides to come out of retirement.

2.2.2 The Implicit Argument

The covert moral in this parable seems to be that Stevens' theory of admissible statistics is incorrect, because the argument provides a counterexample where an inadmissible test leads to a meaningful result. But does this conclusion necessarily follow from Lord's thought experiment? A more structured approach may clarify this issue. The implicit argument Lord makes can be represented by a logical statement about two propositions:

- P1: performing interval manipulations on data that represent measurement on the nominal level results in a meaningless conclusion (i.e. Stevens' theory of admissible statistics is valid);
- P2: performing a parametric test on the football numbers results in a meaningless conclusion.

The logical statement is: if P1, then P2; If Stevens' theory of admissible statistics is valid, then what Lord's statistician does, results in a meaningless conclusion. But the results are not meaningless, for they lead to a useful conclusion; based upon the results of the test, it is concluded that the machine was tampered with and decided that the freshmen should receive new numbers. So, P2 is false, which, by modus tollens, entails that P1 is false; since performing a parametric test in this situation is sensible, performing inadmissible statistical tests on data must be justified, at least in some instances. Lord's parable thus leads the reader to the conclusion that levels of measurement are not always relevant to the choice of statistics.

The above representation, however, does not paint a full picture of Lord's thought experiment. There is an implicit assumption concerning the measurement level of the football numbers that is not included as a proposition. Lord's parable is better represented by making this assumption explicit:

- P1a: performing interval manipulations on data that represent measurement on the nominal level results in a meaningless conclusion (i.e. Stevens' theory of admissible statistics is valid);

- P1b: the football numbers measure a property on the nominal level;

- P2: performing a parametric test on the football numbers results in a meaningless conclusion.

The logical statement now becomes: if P1a and P1b, then P2; if Stevens' theory of admissible statistics is valid and the football numbers measure a property on the nominal level, then what the statistician does is nonsensical. It now becomes clear that the reason that P2 does not hold could lie elsewhere. If P2 is false, either P1a, or P1b, or both, must be false. P1b is not questioned in Lord's parable; the football numbers obviously measure a property on the nominal level. But is it really so obvious that the relevant property in Lord's thought experiment is a property on a nominal scale? If it can be shown that P1b is false, then P1a, and with it the theory of admissible statistics, does not have to be rejected.

2.2.3 What Do the Football Numbers Measure?

The professor in Lord's thought experiment repeatedly emphasizes that the numbers are nominal representations of the uniqueness of the players. Now, the numbers can certainly be used to distinguish players on the field; but this is not the property for which the statistician uses the numbers. Instead, the professor asks a question and draws a conclusion about the machine, namely that it was unlikely to be in its original state (randomly shuffled by the professor) when the freshman numbers were issued. Thus, while an informative inference to the state of the world has been made, this inference does not concern the uniqueness of the football players at all. The level of measurement that the numbers have with respect to these players is completely irrelevant to the thought experiment. Since the players are where the numbers get their status as nominal measurements from, a further conclusion must be drawn: whether a nominal level of measurement plays any role at all in the argument is as yet unsubstantiated. For this reason, premise P1a cannot figure in the argument as described by Lord.

It is prudent to note that we are not the first to point out that the professor's conclusion does not pertain to the football players. Several of Lord's critics have argued this point (Behan & Behan, 1954; Bennet, 1954). They emphasize that Stevens' rules of admissible statistics presuppose that the statistics are performed on measurements of some empirical property, and maintain that the numbers as they are used in the test do not represent any such property; hence Lord's parable does not provide a counterexample to refute Stevens' dictum. Adams, Fagot, and Robinson (1965, p. 125) actually state: "In other words, the hypothesis is concerned with the method of assigning numbers, and has nothing to do with any hypotheses about measurable properties of objects". Unfortunately, these critics fail to note that while the numbers do not refer to the football players in any relevant way, they may refer to another property that is relevant, i.e., the degree to which the machine is biased.

This observation puts us back at square one, for the basic premise that should support Lord's logical construction and the many papers that have used it to substantiate arguments against the theory of admissible statistics is left in doubt. The relevance of Lord's argument to Stevens' theory is no longer obvious. The property that fosters measurement level claims (uniqueness of players) and the property that the statistical conclusion refers to (state of the machine) should be one and the same. This is plainly not the case in Lord's parable. Treatment of Lord's parable could cease here, since this conclusion alone is enough to disqualify the thought experiment as a counterexample to Stevens' dictum. However, the intriguing question why the conclusion based on the parametric test seems so sensible remains unanswered. Perhaps Lord's story about the nutty professor does have some relevance to Stevens' theory, but in a way different from how it is generally viewed. To evaluate such relevance if there is any we need to reconsider the basic question what the football numbers measure and, especially, what the associated level of measurement may be.

It could be argued that, since the statistician performs a one-sample t-test with a fixed population mean, the results will not be invariant under any transformation of the data, so that the statistician is actually assuming an absolute scale. However, in such a line of reasoning one attempts to ascertain the measurement level of the data by looking at the invariance of statistical results, which is not in line with the representational measurement literature or Stevens' dictum. Establishing a measurement level implies the invariance of statistical tests with respect to the class of admissible transformations, but the reverse is not necessarily true: the invariance of statistical tests with respect to a class of admissible transformations does not imply a measurement level. Otherwise one could infer that one has interval level measurements from the fact that the results of a t-test are invariant under linear transformations of the data; this is clearly not the case, because the results of a t-test are always invariant under linear transformations of the data, whether these are at the interval level or not.

2.3 Vending Machine Bias

Lord's statistician uses the statistical results to make an inference about the state of the vending machine, and decides that the freshman mean did not come from the machine in its original state. Thus, Lord's inference concerns the state of the machine relative to another (possible) state of the machine. His reference class is not a set of football players, but a set of possible states of the machine (e.g., fair and biased states). Insofar as measurement is taking place in the thought experiment, therefore, it relates to the assignment of a label to the machine. And in this regard, Lord's example is in perfect accordance with Stevens' theory of admissible statistics.

To see this, it is necessary to first consider how the machine came to be in its altered condition. If we have a better idea what we mean by the state of the machine, better insight may be developed into the nature of the inference being made. Lord gives us a clue how the state of the machine could have been altered by hinting that the professor suspects foul play by the sophomores. Now, the sophomores could have tampered with the machine in several ways. For example, the vending machine could be imagined as an enormous stacked pile of numbers. The numbers are issued one by one from the top of the stack. The sophomores could have tampered with the vending machine by replacing or removing specific numbers at the top of the stack. This way they could ensure that the freshmen received an inordinate amount of even numbers, or numbers ranging from 20 to 30, or prime numbers; the possibilities are endless.

2.3.1 Nominal Bias

If we do not make any assumptions about the way in which the machine was tampered with, the question posed to the statistician may be reformulated into the following research hypothesis: is the machine tampered with or not? This state of tampering can be conceived of as a property of the machine that can be represented nominally with two distinct categories: tampered with and not tampered with. However, it is clear that this way of thinking about Lord's method is in poor accordance with the statistical procedure utilized. That is, if the question were merely "was the machine tampered with or not?", then the test that the statistician performs is not a very good one. It is possible, for instance, that the sophomores removed all the numbers from 01 to 30 and from 70 to 99. The machine would, in this case, clearly be tampered with, but the expected sample mean would not be different from the population mean in the long run. The tampering would not show up in a test were the mean is used to detect a deviation in the sample from the population; the sensitivity of the assignment procedure would be very low.

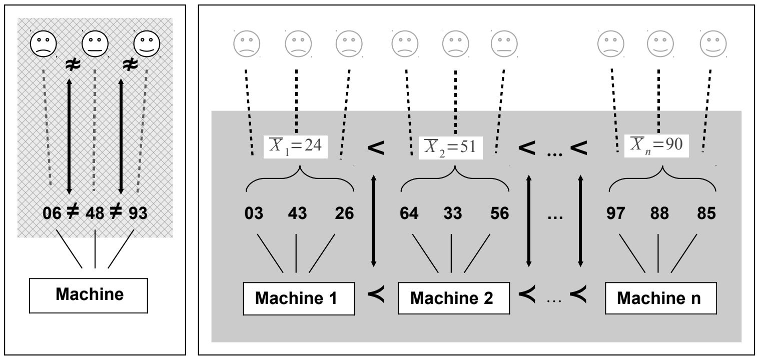
2.3.2 Ordinal Bias

Now consider the fact that Lord provided the statistician with a good argument to use the mean to discover any tampering. The freshmen do not just complain that their sample is different from the expected population, but that the numbers are lower. The freshmen's distaste for low numbers does not magically imbue the numbers with a higher level of measurement, of course, but it does give us more information on how the machine was tampered with. This enables us to refine our understanding of the property of the machine that we are interested in. Knowing that low numbers would upset the freshmen could have prompted the sophomores to remove high numbers from the top of the stack of numbers in the machine. If we focus on this specific type of tampering we can say not only whether the machine was tampered with or not, we can now also say whether the machine was tampered with to a greater or lesser extent. Bias can be introduced into the machine (still consisting of a stack of shuffled numbers) by replacing high numbers with low numbers (or vice versa), resulting in a lower (or higher) population mean. The tampering method suspected by Lord's professor and tested in the statistical analysis is clearly one that introduces a bias toward lower numbers. The sophomores could have been very subtle and removed only a few high numbers, or they could have been extremely overt in their mischief and removed all the numbers larger than 10. The more consistently this replacement is performed, the more bias will be present.

More formally, the amount of bias can be thought of as a variable by imagining a population of vending machines with varying amounts of bias. This idea is illustrated in the right panel in Figure 2.1 by machine 1 to n, which is a subset of the total conceivable population of machines, in which every amount of bias possible is present. The amount of bias in the machine can be represented pragmatically by taking the population mean of the numbers as an indicator for the ordering of the locations of the distributions in question, or by using a nonparametric concept such as stochastic ordering for this purpose. The mean of the extracted numbers may then be compared to the population mean of a fair machine. If there is no bias, the expected value of the sample mean is expected to be equal to this population mean.

To compare this point of view to Lord's sketch of the situation, consider the left panel in Figure 2.1. The property of uniqueness or non-identity of the players is denoted by the relational symbol $\not\approx$. This property is repre-

Figure 2.1: Graphical representation of measurement on the nominal level of the uniqueness of the football players (left panel) and measurement on the interval level of the bias in n vending machines (right panel)



: Measurement on the nominal level of uniqueness of the players, the vending machine is irrelevant.

: Measurement on the interval level of bias in the conceptual machines, the players are irrelevant.

: Indicates correspondence of empirical (≠,≺) relations between objects on the attribute of interest on one hand, and numerical (≠ , <) relations between numbers on the other hand.

$\overline{X}_n$: The mean of machine n.

sented by using different numbers to represent unique players, denoted in the figure by the symbol $\neq$. In the right panel of Figure 2.1, bias in a machine is represented by a mean, in the same way a football player's uniqueness is represented by a single football number. Clearly, the players can only be judged to be different from one another. Equally clearly however, the machines are not merely different from one another; they can also be ordered according to the amount of bias they possess. Normally, relations between objects on the empirical level are described in qualitative terms and then the numbers come in to represent these relations on the numerical level. It might seem a bit strange that in our case the objects on the empirical level already consist of numbers, but this is the nature of the vending machine as constructed by Lord. That is, the numbers in the machine are numbers, not representations, and therefore relations between these numbers can be used (in this case by the sophomores) to make up empirical relational systems that are subsequently represented by a separate numerical relational system.

2.3.3 Interval Bias

It is clear that such an empirical relational system can be constructed, and that Lord's measurement problem involves at least an ordinal structure at the level of the machine. However, it appears that more structure than mere order can be established in the property of bias; one suspects, in fact, that this structure is quantitative in the sense that it should be meaningful to say that the difference in bias between two machines is equal to the difference in bias between two other machines. It turns out that it is possible to show that the bias in the machines has quantitative structure and that the number that the statistician uses to represent it (i.e., the mean) is actually an interval measure of this structure. To this purpose, we need to show that additive structure is present in our bias property and that this structure is represented uniquely up to linear transformations by the population mean.

To show that the machines' level of bias towards low numbers possesses quantitative structure, it suffices to show that an operation exists that allows us to concatenate machines, and that the resulting concatenation has the right properties (Krantz et al., 1971). The operation proposed here very loosely follows the analogy of concatenating temperatures in volumes of liquid. Two equal volumes of liquid, each of a particular temperature, can be added to each other. The resulting temperature is the mean of the separate temperatures. A similar operation on the machines can be conceptualized; the bias in two machines could be 'added' by concatenating the numbers drawn from each machine into a new randomly shuffled pile, which functions as the concatenation of the original machines. This operation allows for the establishment of a relation that satisfies the requirements for measurement on an interval level. The operation is based on the representational measurement theorems describing bisymmetric structures by Krantz et al. (1971, p. 294), which were developed for mean structures. A formal treatment of the bisymmetric structure and how it applies to Lord's thought experiment is provided in Appendix A. That the operation results in measurement on an interval level is intuitively clear. Any non-linear transformation would stretch or shrink the scale somewhere and make comparison of differences in bias of machines impossible. Any linear transformation however, would represent the bias towards low numbers in these machines equally well.

It was already shown that Lord's parable does not provide a counterexample to Stevens' dictum, because the assumption that the numbers are on a nominal level is invalid. No further analysis of Lord's thought experiment is needed to make this point. However, when we do take a closer look at its structure, it is clear that there exists at least one conceptualization of Lord's thought experiment in which the statistician is operating in accordance with Stevens' principles. Therefore, in addition we have now shown that the procedure followed can actually be viewed as an illustration of the theory of admissible statistics. Viewing Lord's thought experiment in this way also answers the question why the statistician's conclusion seems so sensible. It is sensible because it is about a relevant property of the machine.

2.4 Conclusion

We have examined extensively why the test in Lord's thought experiment appears to be inadmissible, while at the same time it leads to a scientifically useful and informative conclusion. In doing so Lord's argument was found to depend on the assumption that the football numbers represent a property on the nominal level. Not only was it shown that it is immaterial to the argument that the numbers represent nominal uniqueness of the players, it was also possible to identify another relevant property, namely the level of bias towards low numbers that a machine exhibits. The numbers in fact represent both the property of uniqueness (in relation to the players) on a nominal level and the property of bias (in relation to machines) on an interval level. What is important here is that the property that corresponds to the statistical question must be considered in determining the admissibility of a test. Because the freshmen complain about low numbers and because the statistician uses a test of the mean that is sensitive to order and differences we conclude that Lord's professor was actually interested in inferring something about bias in the machine towards low numbers. This property of bias was argued to have a structure that can be measured on an interval scale by the population mean, thereby transforming Lord's counterexample into a perfect illustration of Stevens' theory of admissible statistics.

2.4.1 Representing Several Properties at Once

Our analysis relies on the assumption that a single set of numbers may have multiple representational purposes. It is interesting, in this respect, that some of Stevens' critics have used the fact that the same data can represent different properties as an argument against Stevens' theory of admissible statistics. Velleman and Wilkinson (1993), for instance, argue that the level of measurement is not a characteristic of the data. They state that the same numbers can relate to different properties at different measurement levels. Why this is an argument against Stevens might (and should) seem oblique to the reader. It was probably incited by Stevens' procedure for determining the level measurement by assessing the rule used to assign numbers.

In Stevens' thinking, a property can be represented on an interval level if participants can judge intervals on this property to be equal. The rule used to assign numbers thereby determines the measurement level for these particular numbers; once a rule is chosen, the measurement level is set. However, even in Stevens' original papers, one can identify appeals to the requirement that the rules used to assign numbers must yield a numerical structure that is isomorphic to that of the property measured, or to its behavior under empirical operations (Stevens, 1946, p. 677). In a more sophisticated form, this requirement became a cornerstone of representational measurement theory. When one accepts that representational measurement theory has replaced Stevens' original, rather crude theory of levels of measurement, Velleman and Wilkinson's point becomes moot. In representational measurement theory, the level of measurement is determined jointly by the structure of the property of interest and the relation that the numerical assignments bear to that structure. According to this view, nothing prevents the same numbers from representing different aspects of a property, or different properties altogether.

2.4.2 Lord's Forgotten Rejoinder

[h] Reflection on Lord's thought experiment and the critique by Velleman and Wilkinson (1993) nicely illustrates that what our numbers and our conclusions refer to is not always as obvious as we may think. The fact that a simple football number example still puzzles us after more than fifty years shows that measurement issues in relation to legitimate inference (a term

coined by Michell, 1986) concerning a measured property deserve our attention. Researchers should be aware of the inferential power of their statistical conclusions or the lack thereof. Perhaps surprisingly, Lord would probably have agreed. In a response to his critics, published a year after the original article (Lord, 1954), he stated that when one wants to draw an unambiguous conclusion about a property that at best can be represented ordinally, independent of the scale that was used, then nonparametric statistics should be employed.

The point Lord ostensibly intended to make is that Stevens' rules should not be applied mindlessly when choosing a statistical test, but that each situation should be considered anew. In the football numbers argument, Lord attempted to show that there are situations conceivable where these rules do not have to be applied. In all likelihood, however, Lord did not recognize that a relevant property allowing interval level representation could be identified in his example. Had Lord recognized that the football numbers represent the property of bias in the machine on an interval level, he probably would have agreed that his thought experiment does not provide a compelling argument against Stevens' ideas. It remains to be seen if anyone is able to come up with an example where a question about a nominally measured property, answered with a parametric test, results in a truly sensible conclusion about the same nominal property that the numbers refer to. This challenge, of course, stands for all those who argue that statistics and measurement are completely disconnected scientific domains.

2.4.3 Determining the Level of Measurement

Unfortunately, Lord's first publication (1953) has had an enormous influence on the measurement statistics debate; nearly every contributing author refers to this publication. All of Stevens' opponents use Lord's thought experiment and the infamous quote[1] "the numbers don't know where they came from" (Lord, 1953, p. 751) to illustrate their arguments, sometimes even using the quote itself as an argument (Gaito, 1980, p. 565). In contrast, his follow-up publication (1954) has been cited only three times (web of science citation search, at the time of publication). This is regrettable, because Lord's in-

[1]The text actually reads: "The numbers don't know that. Since the numbers don't remember where they came from, they behave just the same way, regardless."

tended point was right on the money: Stevens' rules should not be applied mindlessly.

Careful deliberation is necessary, because one can easily lose sight of the correspondence between the property that is actually being measured and the property about which one wants to make an inference. When choosing a statistical test, considerable thought should be given to the property about which one wants to draw a conclusion, the way this property is measured, and the level it is measured on. Of course, drawing firm conclusions about the achieved measurement level is almost always beyond our reach (see Roberts, 1986, on how the theory of meaningfulness can be applied in psychology). Demanding that the level of measurement for psychological properties is unequivocally determined before continuing with substantive research would bring psychological inquiry to a grinding halt. We certainly would not want to contribute to such a disastrous development. However, researchers need to at least consider the property they want to infer something about and commit to a level of measurement associated with this property, preferably using plausible arguments. Having done so, researchers should consider whether their statistics allow them to draw conclusions that are independent of the specific scale that was used. Of course, future research could always show that the assumptions about the level of measurement were wrong, but this way the research was at least performed in a manner that is internally consistent. To paraphrase Lord: The numbers don't have to know where they came from; researchers have to know where they came from, since they assigned them in the first place.

In conclusion, the current analysis shows that Lord's parable is ill-suited to serve as an argument against the relevance of measurement level to the choice of statistical analysis. However, it may be fruitfully reinterpreted as a warning to researchers that measurement can be much more complicated than it seems, and that measurement levels in fact are important, even in cases where they initially seem irrelevant. Perhaps this point would come across better if we abandon Stevens' interpretation of admissibility of tests, along with the pejorative connotation of this terminology, and encourage psychological researchers to consider the validity of their inferences, not the admissibility of their statistical tests.