Is the Selection of Statistical Methods Governed by Level of Measurement?

BRUNO D. ZUMBO
University of Ottawa

DONALD W. ZIMMERMAN
Carleton University

Abstract
The notion that nonparametric methods are required as a replacement of parametric statistical methods when the scale of measurement in a research study does not achieve a certain level was discussed in light of recent developments in representational measurement theory. A new approach to examining the problem via computer simulation was introduced. Some of the beliefs that have been widely held by psychologists for several decades were examined by means of a computer simulation study that mimicked measurement of an underlying empirical structure and performed two-sample Student *t*-tests on the resulting sample data. It was concluded that there is no need to replace parametric statistical tests by nonparametric methods when the scale of measurement is ordinal and not interval.

Stevens' (1946) classic paper on the theory of scales of measurement triggered one of the longest standing debates in behavioural science methodology. The debate – referred to as the levels of measurement controversy, or measurement-statistics debate – is over the use of parametric and nonparametric statistics and its relation to levels of measurement. Stevens (1946; 1951; 1959; 1968), Siegel (1956), and most recently Siegel and Castellan (1988) and Conover (1980) argue that parametric statistics should be restricted to data of interval scale or higher. Furthermore, nonparametric statistics should be used on data of ordinal scale. Of course, since each scale of measurement has all of the properties of the weaker measurement, statistical methods requiring only a weaker scale may be used with the stronger scales. A detailed historical review linking Stevens' work on scales of measurement to the acceptance of psychology as a science, and a pedagogical presentation of fundamental axiomatic (i.e., representational) measurement can be found in Zumbo and Zimmerman (1991).

Many modes of argumentation can be seen in the debate about levels of measurement and statistics. This paper focusses almost exclusively on an empirical form of rhetoric using experimental mathematics (Ripley, 1987). The term experimental mathematics comes from mathematical physics. It is loosely defined as the mimicking of the rules of a model of some kind via random processes. In the methodological literature this is often referred to as monte carlo simulation. However, for the purpose of this paper, the terms experimental mathematics or computer simulation are preferred to monte carlo because the latter is typically referred to when examining the robustness of a test in relation to particular statistical assumptions. Measurement level is not an assumption of the parametric statistical model (see Zumbo & Zimmerman, 1991 for a discussion of this issue) and to call the method used herein "monte carlo" would imply otherwise. The term experimental mathematics emphasizes the modelling aspect of the present approach to the debate.

The purpose of this paper is to present a new paradigm using experimental mathematics to examine the claims made in the levels of measurement controversy. As Michell (1986) demonstrated, the concern over levels of measurement is inextricably tied to the differing notions of measurement and scaling. Michell further argued that fundamental axiomatic measurement or representational theory (see, for example, Narens & Luce, 1986) is the only measurement theory which implies a relation between measurement scales and statistics. Therefore, the approach advocated in this paper is linked closely to representational theory. The novelty of this approach, to the authors knowledge, is in the use of experimental mathematics to mimic representational measurement. Before describing the methodology used in this paper, we will briefly review its motivation.

Admissible Transformations

Representational theory began in the late 1950's with Scott and Suppes (1958) and later with Suppes and Zinnes (1963), Pfanzagl (1968), and Krantz, Luce, Suppes & Tversky (1971). Recent expositions by Roberts (1979) and Narens (1985) exemplify the prevailing wisdom in representational measurement. Narens and Luce (1986) demonstrate some exquisite mathematical ideas, stemming from abstract algebra and the foundations of mathematics. Representational theory views measurement as a mapping or function (i.e., a homomorphism or isomorphism) of some underlying empirical structure to a representational structure. There is usually more than one function possible in the representation and therefore the uniqueness of the function is in question. The uniqueness is demonstrated via a uniqueness theorem which tells the researcher how the functions that constitute the representation (i.e., scale) relate to one another. Stevens' (1946, 1951) notion of admissible transformations is used with the uniqueness theorem to define a scale type. That is, a ratio scale is one in which the admissible transformations are of the form f(x) = ax, where $a > 0$. An interval scale is one in which positive linear transformations are admissible. An ordinal scale is one in which the admissible transformations are monotone increasing functions. Finally, a nominal scale is one in which the admissible transformations are one-to-one functions. For further discussion of the issue of similarity, linear and affine transformations see Stine (1989a; 1989b).

The uniqueness theorem not only aids in the classification of the scale type but it also puts limitations on the mathematical operations which will preserve truth about the empirical structure (see, Adams, Fagot, & Robinson, 1965; Roberts, 1979; Siegel & Castellan, 1988). This notion of preserving truth is known as invariance or the appropriateness criterion. That is, a numerical statement is appropriate if and only if its truth (or falsity) remains unchanged under all admissible transformations of the scale involved. The appropriateness criterion directs the debate away from the statistical model and onto the statistical hypothesis. That is, it does not impose restrictions concerning the computation of statistics for a scale, but it does impose restrictions concerning certain statements about those statistics.

Adams, Fagot, and Robinson's theory of appropriateness is cited by textbooks such as Hays (1988), Conover (1980) and Siegel & Castellan (1988) as the reason why parametric statistics should not be used with ordinal data. This is certainly consistent with Adams et al.; but Hays and the other authors never state that the appropriateness criterion is not applicable for systems of measurement for which there are not a clearly defined set of permissible transformations (i.e., measurement scale). Unfortunately, most of the scales in behavioural sciences are those for which we do not know the set of permissible transformations. Therefore, clearly the mathematically elegant appropriateness criterion has very little applicability for most of the behavioural sciences and therefore has not resolved the measurement statistics debate for most behavioural measurements.

The above limitation of the appropriateness criterion was also stated by Adams et al.. For many of the other counter arguments to representational theory's claims to appropriate statistics see Zumbo & Zimmerman (1991).

The present study

With this study we examined the measurement-statistics controversy within a framework of experimental mathematics. Previous empirical studies (see, for example, Baker, Hardyck, & Petrinovich, 1966; or Gregoire & Driver, 1987; and a commentary by Rasmussen, 1989) have begun with an ordinal representational system, and then applied illegitimate transformations to this representational system in order to examine the robustness of the test. The limitations of this approach are discussed by Townsend and Ashby (1984), Stine (1989a), and Zumbo and Zimmerman (1991).

The conclusion drawn from the empirical studies is that the parametric tests perform well on "normally" distributed ordinal data; however, the nonparametric tests perform better than the parametric tests for several nonnormal distributions. Furthermore, the performance of the parametric tests is not hindered by ordinal measurement. However, these empirical investigations focussed their attention solely on the notion of invariance of the representations or mappings with respect to each other as expressed in the appropriateness criterion.

What Baker et al. and Gregoire and Driver did was start with a numerical representational system (or the observed scores which researchers would usually obtain), which they assumed was of ordinal scale. Unfortunately, from the observed scores one cannot discern the scale of measurement. Therefore, generating integers on the interval 1 to 10 (as in Baker et al.) or 1 to 100 (as in Gregoire & Driver) does not guarantee an ordinal scale. Hence, a test of whether the measurement scale is important when we apply statistical tests would require mimicking ordinal measurement so that the statistic computed under conditions of perfect measurement can be compared with the statistic based on imperfect measurements.

According to Baker et al., to state the problem of invariance of results under scale transformations raises the following question: 'Can we make correct decisions about the nature of reality if we disregard the nature of the measurement scale when we apply statistical tests?' (p. 293). For the representational theorist, the nature of reality is reflected within the underlying structure.

A test of this question would seem to require a comparison of the statistical decision made on the underlying structure with that of the statistical decision made on the observed numerical representational structure. Therefore, what is required is to mimic representational measurement. If the statistical decision based on the underlying structure is consistent with the decision made on the representational structure then level of measurement is not important when applying statistical tests. If the statistical decision based on the underlying structure is not consistent with the decision made on the representational structure then level of measurement is important when applying statistical tests. Since statistical statements are probabilistic, consistency of the statistical decision is reflected by the power functions for the given test. The power functions can be interpreted, then, as similar to truth functions in formal logic.

The point of contention in the measurement-statistics debate is whether it is appropriate to use parametric statistics on ordinal data. When alternative statistical tests are available, a criterion for choosing among them is statistical power (given that they maintain type I error rate). Statistical power reflects a correct rejection of the hypothesis. Therefore, a comparison of the power functions of a statistical test based on the underlying empirical structure and the ordinal structure sheds light on the measurement-statistics debate and avoids the notion of invariance of arithmetic operations and Adams et al.'s appropriateness criterion.

In summary, previous researchers have focussed on the invariance of arithmetic operations and representations. We believe our novel paradigm approaches the problem with a different form of invariance. That is, we use the underlying structure as the frame of reference for the invariance rather than the representations. Developments in modern representational theory (i.e., post-Stevenson) are used to mimic measurement. Experimental mathematics is used to generate an underlying structure with specified distributional forms, variances, and mean differences and then a measurement is obtained from this.

It should be noted that statistical power is dependent not only on the sample size, the predetermined type I error rate, and population noncentrality parameter and variances (see Hays, 1981) but also the test under consideration. Therefore, the results of this study are limited to the test and conditions examined.

The test statistic selected for study was Student's *t*-test for two independent samples. The nonparametric alternative to the *t*-test is the Wilcoxon-Mann-Whitney test. The *t*-test is one of the most commonly used statistics in the behavioural sciences and has the advantage of having been studied thoroughly (for a recent discussion, see Sawilowsky & Blair, 1992). For the purpose of this paper we examined cases with equal sample sizes and equal variances.

Method

An ordinal scale is one of a family of scales for which the basic rules for assigning numbers are the determination of order. The basic form of ordinal measurement is the rank ordering of information. A rank ordering is an order-preserving mapping of a set of numbers onto the set of the first N integers. In the case of rank ordering in two-sample tests, N is the sum of the two sample sizes.

Ranks are not the only ordinal scale that is possible. Ordinal measurement which involves further loss of the original information in the empirical structure would involve a distortion of the ranks by creating ties. This is also referred to as a partially-ordered structure. To simulate further loss of information than ranks, we developed *pseudoranks*. Pseudoranks, involve even more loss of information than ranking. To calculate the pseudoranks we divided each rank value by 1.10 and then invoked a function which would simply truncate the decimal portion of the dividend and return the integer value. One problem with this procedure is that the rank of one always becomes zero. So, by default, in the pseudorank procedure, the rank of one was not changed.

In measurement, loss of information is not the only distortion of the empirical structure. Measurement error also plays an important role. Measurement error has been discussed extensively in classical test theory (Lord & Novick, 1968; Rozeboom, 1966a; Zimmerman, 1969, 1975a, 1975b, 1976). Interpreted within the framework of this study, measurement error should not be related to the rank, as well the error distribution should have a mean of zero. To simulate measurement error, a value, randomly selected from a Gaussian (Normal) distribution with mean zero and unit variance, was added to the rank of each score.

In summary, we have mimicked measurement via the use of experimental mathematics (computer simulation). An underlying empirical structure was created with two independent samples of equal means and variances. The underlying empirical structure was generated with varying sampling distributions. The distribution used in this study was the Gaussian or Normal distribution. This was chosen because it was the distribution used in the derivation of the *t*-test. Furthermore, the purpose of this paper was to introduce the new paradigm and give some preliminary results. In this study we wished to examine the relative Type I error rate and power for the *t*-test on the empirical structure and ordinal representations. Of particular interest was the relative power under conditions of similar population distributions (i.e., with identical shape and variance). Future research will be discussed in the final section of this paper.

Computer Simulation Method

A computer program generated random samples based on pseudorandom numbers from the distribution described below and performed five significance tests of location on each pair of samples. Sample sizes were $N_1 = N_2 = 4$, 8, 12, 16, and 25, and all tests were nondirectional. These sample sizes were chosen because they are representative of research in psychological and social sciences, and are similar to sample sizes used in previous simulation studies (see for example, Boneau, 1960; Baker, Hardyck, & Petrinovich, 1966).

For the simulations the Type I error rate and power were calculated for a = .05[1], .01 and .10 on data from independent populations of homogeneous variance. For Type I error rate and power 5000 replications were carried out for each degree of separation between means, that is, for each point of the power function. For each replication, N_1 values were generated according to the specified probability distribution and then N_2 values were generated independently from the same distribution and the test statistic was calculated. The Type I error rate was the proportion of replications in which the obtained test statistics exceeded the critical values (i.e., declared statistically significant). For the power of the statistical tests, an effect size of $ES = 0.50$ to 4.0 standard units of difference (in increments of .50) was added to N_2. For each effect size the statistics were calculated and compared with their critical values. The power of the statistical test, at a given effect size, was represented by the proportion of replications for which the test was declared significant.

The order of tests are as follows: first, a Student t-test for independent groups was performed on the original values. Next, these values were ranked, and a Wilcoxon-Mann-Whitney test statistic was calculated in the usual way from rank sums. At this point, the Student t-test was performed on the ranks. Then, the pseudoranks and the ranks with measurement error were introduced.

In the case of pseudoranks, the partially-ordered structure, the ranks were replaced by integral-valued functions of the ranks. The sixteen ranks assigned to the original combined samples of $N_1 = N_2 = 8$, were replaced by integers from the set {1,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 11, 12, 13, 14}. The thirty-two ranks assigned to the original combined samples of $N_1 = N_2 = 16$, were replaced by integers from the set of the sixteen integers of the prior set, as well, {15, 16, 17, 18, 19, 20, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29}, and so on. It should be noted that ties were evident in the pseudoranks.

Next, the transformation mapping ranks to ranks with measurement error (rme) was investigated. For each case, a number was selected from the Gaussian distribution, with mean zero and unit variance, and added to the initial rank. In this way, the sample values associated with the rme were no longer all integral valued and vary from sample to sample.

Student t-tests were applied to the pseudoranks and to the ranks with measurement error. Differences between population means were introduced, so that the probabilities of both Type I and Type II errors and points on the power function could be obtained. The differences were expressed in units of the standard error of the difference and ranged from 0 to 4.0 standard error units, in increments of 0.5.

When using power functions, an important property of the functional relation is the effect size, ES. The effect size is a ratio of the mean differences and standard deviations in the population. For present purposes, an issue that should be addressed is the effect for the power functions of the representations (i.e., ranks, pseudoranks, and ranks with measurement error). The fundamental point of contention in the measurement-statistics debate is whether the ordinal mapping preserves the effect (i.e., mean difference). Therefore, the present simulation was designed so that the effect size was introduced only for the underlying empirical structure. The empirical structure was then mapped to the ordinal representations and statistical tests were performed to examine

[1] For sample size of 16, the significance level for W and t was set at .0475 because for this sample size there is no tabled integral value of W associated with an exact probability of .05. The critical values were $t = 2.0765$, and $W = 75.495$. See Zimmerman & Zumbo (1989) for a note on this matter.

whether the mean difference was preserved.

METHODS OF SIMULATING VARIATES

The methods of simulating the variates were as follows (see Lehman, 1977; Morgan, 1984; Devroye, 1986, and Ripley, 1987, as standard texts). In all cases X_1 and X_2 are independent pseudorandom numbers on the interval [0,1]:

Gaussian (Normal) Distribution. Normal deviates were generated using the method of Box and Muller (1958). The relation used was $X = \sqrt{-2 \log X_1} \cos 2 \Pi X_2$. The pseudorandom numbers were generated using a well-known and thoroughly tested prime-modulus multiplicative congruential generator described by Lewis and Orav (1989) and Lewis, Goodman, and Miller (1969).

Results

Tables 1 and 2 contain the empirical power functions for sample sizes of 4 through 25 and Table 3 presents the results for varying apriori Type I error rates, a. The tables were constructed such that, for each sample size, the column labelled *ES* represents the effect size..The effect size was measured in terms of standard units of difference between population means (This is also referred to as the noncentrality parameter, see Hays, 1981, pp. 287-289 for details) ranging from 0 to 4 in increments of .50. Due to space limitations, Tables 2 and 3 report three points of the power functions rather than nine[2]. The numbers in the columns represent the probability measure (i.e. the proportion of test statistics declared significant). The functional relation of the effect size and probability values is the power function given that the Type I error rate was close to that set apriori. The probability values at 0 units of difference is the Type I error rate.

TABLE 1
Power function for a Gaussian distribution for various sample sizes (Tolerance band in parentheses).

			N = 4	
ES	Underlying	W/rank	Pseudo-ranks	Ranks/error
0	054 (048, 059)	058	058	061*
0.5	079 (071, 087)	082	082	087
1.0	150 (135, 165)	148	148	149
1.5	259 (233, 285)	250	250	242
2.0	400 (360, 440)	394	394	371
2.5	576 (518, 634)	553	553	509*
3.0	726 (653, 799)	704	704	624*
3.5	842 (758, 926)	821	821	724*
4.0	921 (829, 100)	905	905	812*

			N = 8	
ES	Underlying	W/rank	Pseudo-ranks	Ranks/error
0	046 (041, 051)	043	051	046
0.5	071 (064, 078)	069	080*	071
1.0	155 (140, 171)	149	170	150
1.5	298 (268, 328)	288	311	283
2.0	467 (420, 514)	448	474	445
2.5	648 (583, 713)	620	645	609
3.0	809 (728, 890)	782	802	771
3.5	911 (820, 100)	886	902	875
4.0	965 (869, 100)	955	962	945

Note: Values reported in table without decimal point. * indicates not within the tolerance band.

It was previously noted by Conover and Iman (1981) that the Wilcoxon-Mann-Whitney test is statistically equivalent to a *t*-test on ranked values. This statistical equivalence was discussed by Zimmerman & Zumbo (1989, 1993). The definition of equivalence was further elaborated such that two tests were said to be alpha-equivalent if, when performed on the same sample data using the same significance level, they always lead to the same statistical decision (Zimmerman & Zumbo, 1990). Therefore, for clarity of presentation, one power function labelled *W/rank* represented the statistical decisions for the Wilcoxon-Mann-Whitney and *t*-test on ranks. It should be noted however, that the rank transformation has recently been demonstrated to be detrimental for some designs other than the two independent samples case discussed in the present paper (see, for example, Blair, Sawilowsky, & Higgins, 1987; Sawilowsky, Blair, & Higgins, 1989; Sawilowsky, 1989, 1990).

[2] A more complete list can be obtained from the authors.

The power functions for *t*-tests on the underlying empirical structure are labelled *Underlying*. The *t*-tests on pseudoranks are labelled *Pseudoranks*. And finally, the *t*-tests on ranks with measurement error are labelled *ranks/error*.

A criterion for judging whether the representational (i.e., rank, pseudo ranks, and ranks with error) power functions were consistent with the empirical structure was set to plus or minus 10% of the value for the underlying distribution. This is a tolerance band and is listed in parentheses next to power estimate for the underlying empirical structure throughout all of the tables. The criterion of 10% was chosen somewhat arbitrarily but was a compromise between the rather conservative measurement which would treat each probability entry as a series of 5000 Bernoulli trials and calculate a standard error of the point estimate, and the rather liberal "eye-balling or interocular test" of the results[3].

Examining Tables 1, 2 and 3 it seems evident that conducting a Wilcoxon-Mann-Whitney test or a *t*-test on the pure ordinal representation (i.e., ranks) will consistently yield the same statistical decision. This is evident from the fact that the two power functions could be summarized as one. That is, the power functions are alpha-equivalent. Generally, the *t*-tests on pseudoranks and ranks with measurement error are within the tolerance regions of the *t*-tests on the underlying structure. The exceptions being for small sample sizes of 4 per group with large effect sizes and the three cases where the values were barely outside the tolerance band.

TABLE 2
Three points of the power function for a Gaussian distribution for various sample sizes (Tolerance band in parentheses).

		$N = 12$		
ES	Underlying	W/rank	Pseudo-ranks	Ranks/error
0	050 (045, 055)	052	050	052
1.5	305 (275, 336)	299	289	290
3.0	824 (742, 906)	806	798	794

		$N = 16$		
ES	Underlying	W/rank	Pseudo-ranks	Ranks/error
0	051 (046, 056)	049	051	047
1.5	297 (267, 327)	280	286	282
3.0	822 (740, 904)	804	810	803

		$N = 25$		
ES	Underlying	W/rank	Pseudo-ranks	Ranks/error
0	044 (040, 048)	050	044	044
1.5	286 (257, 315)	290	275	273
3.0	823 (741, 905)	821	807	806

Note: Values reported in table without decimal point.

Discussion

The measurement-statistics debate was examined by the vehicle of experimental mathematics. Ordinal measurement was simulated by utilizing axiomatic representational theory's formalisms of a representation. Representational theory, like its predecessor, Stevens' scale type theory (Stevens, 1946, 1951, 1968), places limitations on the use of parametric statistics on ordinal measurements (Townsend & Ashby, 1984).

The objection to the use of parametric statistics on ordinal measurement is that the statistical decision on the ordinal structure will not be consistent with the nature of the decision on the unobservable structure[4]. The present paper uses experimental mathematics so that the latent structure becomes observable. Simulating an artificial latent structure with preset specifications allows an examination of the statistical decision on the

[3] The value of 10% was decided upon because it is the subjective value the authors use when examining published monte carlo studies. This issue is directly related to the definition of robustness in monte carlo studies as discussed by Bradley (1978).

[4] We use Neyman-Pearson two-point hypothesis testing as a mechanism to examine the measurement-statistics debate. This use should not be interpreted as indicative of our support of the blind use of hypothesis testing in practical research settings.

TABLE 3
Three points of the power function for a Gaussian distribution for various alpha levels and $N = 8$ (Tolerance band in parentheses).

α = .01				
ES	Underlying	W/rank	Pseudo-ranks	Ranks/ error
0	010 (009, 011)	010	010	012*
1.5	105 (095, 116)	110	108	114
3.0	529 (476, 582)	511	505	512

α = .10				
ES	Underlying	W/rank	Pseudo-ranks	Ranks/ error
0	097 (087, 107)	100	097	097
1.5	422 (380, 464)	412	402	394
3.0	895 (806, 985)	880	874	857

Note: Values reported in table without decimal point. * indicates outside of the tolerance band.

latent structure relative to the statistical decision on the ordinal structures. This allows us to examine whether the statistical decision is hindered by the use of ordinal measurement; as predicted by Stevens and modern axiomatic representational theorists.

In general, the results indicate that for statistical hypothesis testing of two-sample location problems (i.e., tests of mean differences) it is not detrimental to use parametric tests on ordinal data. That is, if a mean difference is evident in the latent structure, then the *t*-test or Wilcoxon-Mann-Whitney test performed on data from an ordinal representation will indicate a mean difference, at least as often as a *t*-test on the empirical structure. Also, if no mean difference is evident in the latent structure, then the *t*-test or Wilcoxon-Mann-Whitney test performed on data from an ordinal representation will indicate no mean difference as often as a *t*-test on the empirical structure. These findings also are maintained for various sample sizes and significance levels.

Of particular importance is the finding that the power functions for the *t*-tests on the ordinal representations are very similar to the power functions for the nonparametric Wilcoxon-Mann-Whitney test. This indicates that there is no benefit to be gained from excluding the use of parametric statistics on ordinal data. These findings are also true for various sample sizes and significance levels.

An alternative interpretation of our results is offered by Stine (1989c). He argues that given we generate an underlying structure which has a probability distribution we are implicitly imposing an interval metric structure upon our underlying empirical structure. Therefore, he interprets our present results with the caveat that they are only true given that we have an ordinal measurement of a interval latent measure. Stine's (1989a) statement that we cannot resolve the levels of measurement debate via computer simulation is related to the above reinterpretation. That is, any computer simulation must generate interval scaled latent variables because we cannot generate a latent variable without metric structure. That is, when we generate pseudorandom numbers we are imposing a uniformly distributed generating process and therefore has metric structure. Of course, any transformations of the pseudorandom numbers to create other distributional shapes also suffers of Stine's limitations.

Latent Variables

We concur with Stine (1989c) that the present results must be framed within a context of at least an interval scaled latent variable. However, we do not consider this a limitation. For practical purposes, when researchers consider latent variables they implicitly impose at least an interval structure upon them by their operationalization (See Rozeboom, 1966b, on the issue of imposing content on a latent variable). Researchers usually think about latent variables with a metric structure (i.e., distributional shape). Classic examples come from the domain of intelligence where I.Q. scores are normally distributed or cognition and cognitive science where reaction time/response time data are argued to be either Exponential or Gamma distributed. As well, Item Response theory (also referred to as Modern tests theory, see Lord & Novick, 1968) where the latent trait or ability is Normally distributed. It is not unusual to impose an interval structure on unobserved measures. We know of no measures which are conceived of without distributional form. This may in fact be a residue of the interweaving of statistical theory through modern systemic behavioural and social sciences.

In fact, Stine's interpretation of the results reflects a clear line on which scholars involved in the measurement statistics debate are divided (this was also suggested by Michell, 1986 but becomes clearer with the current results). Many of the individuals who state that levels of measurement are not important (for example, Davison & Sharma, 1988, 1990; Lord, 1953) impose an interval structure to the latent variable. Stine (1989c) is suggesting a mathematically interesting scenario wherein the underlying variable has no metric structure; however, almost all behavioural and social science research deals with variables with a metrically defined latent structure.

Therefore, for the practicing researcher to test hypotheses of mean differences, all that is needed to maintain statistical power and type I error rate, is ordinal information from the underlying empirical structure. The results indicate that, when deciding whether to use parametric or nonparametric statistical methods for a two-sample location problem, the level of measurement is not the criterion on which to make this decision. In fact, the form or shape of the probability distribution is a better criterion than levels of measurement.

These findings are, of course, restricted to the parameters (for e.g., two independent samples, equal variances, a location shift model, etc.) examined within this study. Forthcoming work will examine the case where the underlying distributions are not Normally distributed. There is no need to assume that the underlying structure is Normally distributed; it could be one of a multitude of distributional shapes. Also, this paradigm will be used to examine other univariate and multivariate statistical procedures. On a methodological note, forthcoming work will deal with a more precise criterion for comparing power functions, in particular, a possible empirical tolerance band or exploring further the notion of a truth function.

This research was made possible by a Social Sciences and Humanities Research Council of Canada Fellowship and a Carleton University Research Grant to BDZ. The first author is indebted to his previous professors W.W. Rozeboom and K. V. Wilson of the Dept. of Psychology, University of Alberta who unknowingly gave invaluable guidance and suggestions which led to the formulation of this paper. They, of course, are not to be held responsible for his blunders. Send correspondence to: Bruno D. Zumbo, Measurement & Evaluation, Faculty of Education, University of Ottawa, 145 Jean-Jacques Lussier, Ottawa, Ontario, Canada KIN 6N5.

References

Adams, E.W., Fagot, R.F., & Robinson, R.E. (1965). A theory of appropriate statistics. *Psychometrika, 30,* 99-127.

Baker, B.O., Hardyck, C.D., & Petrinovich, L.F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscription on statistics. *Educational and Psychological Measurement, 26,* 291-309.

Blair, R.C., Sawilowsky, S.S., & Higgins, J.J. (1987). Limitations of the rank transform in tests for interaction. *Communications in Statistics: Computation and Simulation,* B16, 1133-1145.

Boneau, C.A. (1960). The effects of violations of assumptions underlying the *t*-test. *Psychological Bulletin, 57,* 49-64.

Box, G.E.P., Muller, M. (1958). A note on the generation of random normal deviates. *Annals of mathematical statistics, 29,* 610-611.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Conover, W.J. (1980). *Practical nonparametric* *statistics* (2nd Edition). New York: Wiley.

Conover, W.J., & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35,* 124-128.

Davison, M.L., & Sharma, A.R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin, 104,* 137-144.

Davison, M.L., & Sharma, A.R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin, 107,* 394-400.

Devroye, L. (1986). *Non-Uniform random variate generation.* New York: Springer-Verlag.

Gregoire, T.G., & Driver, B.L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101,* 159-165.

Hays, W.L. (1981). *Statistics* (3rd Edition). New York: Holt, Rinehart, Winston, Inc.

Hays, W.L. (1988). *Statistics* (4th Edition). New York: Holt, Rinehart, Winston, Inc.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. 1, Additive and polynomial representations.* New York: Academic Press.

Lehman, R.S. (1977). *Computer simulation and modeling: an introduction.* Hillsdale, N.J.: Erlbaum.

Lewis, P.A.W., Goodman, A.S., & Miller, J.M. (1969). A pseudorandom number generator for the System 360. *IBM Systems Journal, 8,* 136-146.

Lewis, P.A.W., & Orav, E.J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove, CA: Wadsworth.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750-751.

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

Michell, J. (1986). Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin, 100,* 398-407.

Morgan, B.J.T. (1984). *Elements of simulation.* London: Chapman & Hill.

Narens, L. (1985). *Abstract measurement theory.* Cambridge: MIT Press.

Narens, L., & Luce, (1986). Measurement: the theory of numerical assignments. *Psychological Bulletin, 99,* 166-180.

Pfanzagl, J. (1968). *Theory of measurement.* New York: Wiley.

Rasmussen, J.L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin, 105,* 167-170.

Ripley, B.D. (1987). *Stochastic Simulation.* New York: John Wiley & Sons.

Roberts, F.S. (1979). *Measurement theory, with applications to decisionmaking, utility, and the social sciences.* Reading: Addison-Wesley.

Rozeboom, W.W. (1966a). *Foundations of the theory of prediction.* Homewood, Ill.: Dorsey Press.

Rozeboom, W.W. (1966b). Scaling theory and the nature of measurement. *Synthese, 16,* 170-233.

Sawilowsky, S.S. (1989). *Rank transform: The bridge is falling down.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research,* 60, 91-126.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 112,* 352-360.

Sawilowsky, S.S., Blair, R.C., & Higgins, J.J. (1989). An investigation of the type I error and power properties of the rank transformation procedure in factorial ANOVA. *Journal of Educational Statistics, 14,* 255-267.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic,* 23, 113-128.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd Edition). New York: McGraw-Hill.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens

(Ed.), *Handbook of experimental psychology.* New York: Wiley.

Stevens, S.S. (1959). Measurement, psychophysics and utility. In Churchman, G.W., and Ratoosh, P. (Eds.), *Measurement: Definitions and theories.* New York: Wiley.

Stevens, S.S. (1968). Measurement, Statistics, and the Schemapiric View. *Science, 161,* 849-856.

Stine, W.W. (1989a). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, 105,* 147-155.

Stine, W.W. (1989b). Interoberver relational agreement. *Psychological Bulletin,* 105, 341-347

Stine, W.W. (1989c). Personal communication. During the University of New Hampshire, Durham, *Conference on the History and Theory of Methods,* Durham, New Hampshire, November.

Suppes, P., & Zinnes, J.L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology Vol. 1.* New York: Wiley.

Townsend, J.T., Ashby, F.G. (1984). Measurement scales and statistics: the misconception misconceived. *Psychological Bulletin, 96,* 394-401.

Zimmerman, D.W. (1969). A simplified probability model of error of measurement. *Psychological Reports, 25,* 175-186.

Zimmerman, D.W. (1975a). Two concepts of "true score" in test theory. *Psychological Reports, 36,* 795-805.

Zimmerman, D.W. (1975b). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395-412.

Zimmerman, D.W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement, 36,* 85-96.

Zimmerman, D.W., & Zumbo, B.D. (1989). A note on rank transformations and comparative power of the Student *t*-test and Wilcoxon-Mann-Whitney test. *Perceptual and Motor Skills, 68,* 1139-1146.

Zimmerman, D.W., & Zumbo, B.D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and Student *t* test under simple bounded transformations. *Journal of General Psychology, 117,* 425-436.

Zimmerman, D.W., & Zumbo, B.D. (1993). Relative power of parametric and nonparametric statistical methods. In Gideon Keren & Charlie Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Vol. 1: Methodological Issues* (pp. 481-517). Hillsdale, NJ.: Lawrence Erlbaum.

Zumbo, B.D., & Zimmerman, D.W. (1991). Levels of measurement and the relation between parametric and nonparametric statistical tests: a review of recent findings. *A handbook for data analysis in the behavioural sciences, Vol. 1: Methodological issues.* (pp. 481-517).