Psychological Methods
1998, Vol. 3, No. 1, 55-67


1082-989X/98/$3.00

# Analysis of Likert Scale Data in Disability and Medical Rehabilitation Research

Michael J. Nanna
Rehabilitation Institute of Michigan

Shlomo S. Sawilowsky
Wayne State University

Many clinical evaluations are subjective, resulting in ordinal level measurements. A widely used example in medical rehabilitation is the Functional Independence Measure (FIM), which provides a measure of disability. The FIM is an 18-item, 7-point Likert scale ranging from complete dependence to complete independence. Parametric statistics are commonly used for the analysis of ordinal data. However, Likert scales often lead to violation of many underlying assumptions. This study examined the comparative power of the $t$ test with the Wilcoxon rank-sum test using real pretest/posttest data sets measured on an ordinal scale. FIM scores were obtained on 714 geriatric patients at admit and discharge from a rehabilitation hospital. A Fortran 77 program was written to sample with replacement from each admit and discharge data distribution. Results indicated the Wilcoxon rank-sum test outperformed the $t$ test for almost every sample size and alpha level examined.



There is increasing attention being given to assessment of functional outcome in the field of rehabilitation medicine. This information is becoming more important to medical institutions and governmental health care agencies that are confronted with decreasing levels of funding and to insurance companies who are adhering to more stringent criteria for reimbursement. In fact, "third party payors are increasingly reimbursing only in cases where functional improvements are documented" (Baldrige, 1993, p. 3). Moreover, "functional status at admission and/or discharge from rehabilitation has even been proposed as a basis for hospital payment" (Stineman et al., 1996, p. 1101). (See also Batavia, 1988; Harada, Sofaer, & Kominski, 1993; Stineman et al., 1994; Wilkerson, Batavia, & DeJong, 1992.)

Unlike objective measurements that are based on sensitive instruments with well-defined calibrations (e.g., sphygmomanometer, dynamometer), evaluations in rehabilitative medicine are often subjective and are based on instruments with limited psychometric information, such as reliability and validity. Obviously, this makes accurate documentation of functional improvement tenuous. Complicating this issue, instruments used in clinical assessment are often ordinally ranked evaluations (i.e., rating scales) indicating higher or lesser degrees of a particular construct (e.g., functional performance). Indeed, Heeren and D'Agostino (1987) noted, "Often biomedical research focuses on the comparison of two independent samples on some dependent measure that is an ordinal variable with only three, four, or five levels" (p. 79).

## Ordinal Measurement

According to Siegel (1956), an ordinal scale is "irreflexive, asymmetrical, and transitive" (p. 24). That is, the following conditions apply (a) if it is not true for any x that $x > x$ (irreflexive); (b) if $x > y$, then $y \not> x$ (asymmetrical); and (c) if $x > y$ and $y > z$, then $x > z$ (transitive; Siegel, 1956, p. 24). Thus, for objects to be on an ordinal scale, there must be a hierarchical relationship of "greater than" or "less than." However, the difference in magnitude from one level to the next may not represent equal amounts. Although 4 may be greater than 3, and 3 greater than 2, the amount of the construct in question may be different between 4 and 3, as compared with that between 3 and 2.

Michael J. Nanna, Research Department, Rehabilitation Institute of Michigan, Detroit; Shlomo S. Sawilowsky, Educational Evaluation and Research Department, College of Education, Wayne State University.

This research is based on Michael J. Nanna's previous work for his master's thesis submitted to the College of Education, Wayne State University.

Correspondence concerning this article should be addressed to Michael J. Nanna, who is now at the Institute for Professional Development, 4615 East Elwood, Phoenix, Arizona 85040. Electronic mail may be sent to mjnanna@apollogrp.edu.

## Ordinal Measurement in Medical Rehabilitation

Some constructs frequently measured on ordinal scales in rehabilitation medicine include specific aspects of disability such as performance of activities of daily living (ADL) skills in persons with disability (Dinnerstein, Lowenthal, & Dexter, 1965; Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963; Mahoney & Barthel, 1965) and functional outcome (Harvey & Jellinek, 1981; Keith, Granger, Hamilton, & Sherwin, 1987). Some examples of specific tests with an ordinal level of measurement include the Katz Index of ADL (Katz et al., 1963), the Kenny Self-Care Evaluation Scale (Schoening & Inersen, 1968), the Functional Life Scale (FLS; Sarno, Sarno, & Levita, 1973), and the Ashworth Scale (Ashworth, 1964). The Katz Index of ADL was developed to study the results of treatment and prognosis in the elderly and the chronically ill. This Index is scored by ranking individuals on their performance in a variety of areas, such as bathing, dressing, and toileting (Katz et al., 1963). The Kenny Self-Care Evaluation Scale was developed to test physical activities necessary for self-care in a home environment. This instrument consists of six major categories and is scaled on a 5-point Likert scale (Baldrige, 1993). The FLS was designed to assess patients in the home or community, rather than in a hospital, and is composed of 44 items across five categories and scored on a 5-point scale. The Ashworth Scale was developed to assess muscle spasticity by manually moving a limb through range of motion to passively stretch specific muscle groups. It is scored on a 5-point Likert scale for grading encountered muscle resistance (Bohannon & Smith, 1987).

## Functional Independence Measure

One of the most widely used assessment instruments in rehabilitation is the Functional Independence Measure (FIM; Keith, Granger, Hamilton, & Sherwin, 1987). In fact, "about 60% of rehabilitation facilities nationwide use the FIM" (Stineman et al., 1996, p. 1101; Granger, Hamilton, Keith, Zielezny, & Sherwin, 1986). It was developed in 1983 and is used as part of a national database system. The FIM was developed to provide uniform assessment of severity of disability and medical rehabilitation outcome, although some subscores purport to assess cognitive and social variables associated with disability. Ordinal scores are obtained from this 18-item, 7-point Likert scale, consisting of scores that range from 1 (*complete dependence*) to 7 (*complete independence*). The scale was originally designed so that ratings on all 18 items were summed into a single score estimating overall burden of care (Stineman et al., 1996). Total FIM scores range from 18 (*complete dependence*) to 126 (*complete independence*).

FIM scores are based on the observation of a patient meeting specific objective behavioral criteria and are usually rated by clinical observation at the time of admission, and again prior to discharge from rehabilitation services. It is intended to measure levels of disability regardless of any underlying pathological condition and is considered independent of the rater's clinical background (Brynes & Powers, 1989; Keith et al., 1987; Granger, Cotter, Hamilton, Fiedler, & Hens, 1990). Previous studies have indicated high levels of instrument reliability (.95; Brynes & Powers, 1989) and interrater agreement (.93 and .97; Hamilton, Laughlin, Granger, & Kayton, 1991). It was concluded in a recent study by Ottenbacher, Hsu, Granger, and Fiedler (1996), that the FIM "provided good interrater reliability across a wide variety of raters with different professional backgrounds and levels of training. The median interrater reliability value was .95 and was based on a large cumulative sample of patients representing a wide variety of disability levels and medical conditions" (p. 1230).

## Statistical Analysis of Ordinal Scores

Classical parametric statistics are the most commonly accepted and widely used techniques for the analysis of data. Parametric tests have underlying assumptions, such as population normality and homogeneity of variance. The discrete nature of Likert scales, however, is conducive to nonnormality, and the limited possible outcomes on Likert scales are likely to produce ceiling or floor effects. Bevan, Denton, and Myers (1974) noted, "Such data are often analyzed by using analysis of variance techniques, though often in face of anxiety on the part of the investigator, who is aware that he has violated the normality assumption" (p. 199). For these reasons, nonparametric or distribution-free statistics have been suggested for the analysis of ordinal scores, because they make no assumption regarding the shape of the population from which samples were drawn.

The choice of a parametric or nonparametric statistic is often said to depend on the level of measure-

ment. According to this position (Siegel, 1956; Stevens, 1946), a variable measured on an ordinal scale should be analyzed with a nonparametric statistic, whereas a variable measured on an interval or ratio scale should be analyzed with a parametric statistic. Rules such as this are repeated frequently (e.g., Findley, 1991, p. S89). These rules have been debated in the statistics and measurement literature for decades in the context of the "weak measurement versus strong statistics" controversy. On the basis of considerable simulation evidence (see, e.g., Hunter & May, 1993; Sawilowsky, 1990, 1993; Zumbo & Zimmerman, 1993), we dismiss level of measurement from consideration in choosing between parametric and nonparametric tests.

The use of parametric procedures when underlying assumptions are violated may affect the test's robustness and power properties. *Robustness* refers to the ability of a statistic to preserve valid probability statements applied to it even though underlying assumptions are violated. That is, it retains its characteristics that were based on normal theory even under nonnormal conditions. Type I error refers to the incorrect rejection of a true null hypothesis, and Type II error refers to the failure to reject a false null hypothesis. The statistical power of a test is its ability to detect a false null hypothesis (i.e., to detect a treatment effect).

Hsu and Feldt (1969) investigated the effect of score scale limitations on the probability of Type I error rates in analysis of variance layouts. They noted that distributions of scores on an ordinal level are frequently skewed, platykurtic, and that the variability of scores may differ from treatment population to treatment population, severely violating the normality and homogeneous variance assumptions. They randomly selected scores from discrete score distributions patterned after actual social and behavioral science data sets. The $F$ test displayed excellent control over Type I error rates with 3-, 4-, and 5-point Likert scales. Hsu and Feldt concluded, "Experimenters need not hesitate to use $F$-tests with data based on scales of three or more points" and even "data from a two-point scale may be validly analyzed" for sample sizes of 50 or more (p. 526). They did not, however, examine the comparative power of the parametric test with nonparametric competitors under these situations. Bevan et al. (1974) found similar levels of robustness with the $F$ test applied to 7-point Likert scale data.

Heeren and D'Agostino (1987) investigated the robustness properties of the two independent samples $t$ test when applied to data scaled at the ordinal level. They "generated the full sampling distribution of the $t$ statistics over a range of sample size combinations" (p. 81) and demonstrated that the $t$ test was robust by comparing nominal alpha to actual Type I error rates for 3-, 4-, and 5-point Likert scales. However, they too examined only the robustness of the $t$ test and did not compare it with respect to the power of a nonparametric counterpart, such as the Wilcoxon rank-sum test.

## Limitations of Previous Simulation Studies

Micceri (1989) canvassed social and behavioral science literature and obtained 440 distributions from applied research studies and standardized test databases. He found that nonnormality in the form of extreme asymmetry or lumpiness was quite typical. In fact, only 3% of the distributions were symmetric with light tails, and none of them passed traditional tests of normality. Thus, statistical procedures based on strict assumptions, such as normality, are being applied to data that are nonnormal. Micceri raised the question regarding the potential deleterious effects these violations might have in terms of the robustness and power of classical parametric statistics. In general, studies investigating these statistical properties were based on simulations with mathematical distributions but were not based on the characteristics of real data sets.

Micceri's (1989) study suggested the need to investigate the robustness and power properties of parametric statistics using real data. Sawilowsky and Blair (1992) conducted a Monte Carlo investigation of the robustness with respect to Type I error of the $t$ test to departures from population normality. Distributions were selected as being representative of those commonly found in social and behavioral science research. This was done to provide a more realistic test of the $t$ test's performance under various nonnormal real data situations. In agreement with earlier studies, they found that the $t$ test was reasonably robust under previously reported conditions—large and equal sample sizes and one-tailed rather than two-tailed tests.

In another study, Sawilowsky and Hillman (1992) investigated the Type II error properties of the independent samples $t$ test with a real data set. Treatment effects were modeled using Monte Carlo methods to

sample with replacement from a normal distribution and a real distribution that Micceri (1989) found to be prevalent with "onset variables." Onset variables typically have a mass at zero with a gap before nonzero values appear. Sawilowsky and Hillman found that the independent samples *t* test maintained power levels for the nonnormal distribution consistent with that expected from normal curve theory. However, they repeated the comments pointed out by Scheffé (1959), that in the presence of nonnormality, the preservation of "power calculated under normal theory should not be confused with that of their efficiency against such alternatives relative to other kinds of tests" (p. 351). That is, achieving the power predicted by normal curve theory does not rule out the possibility that a nonparametric test might be considerably more powerful in this nonnormal data context.

A useful feature common to the Monte Carlo studies by Sawilowsky and Blair (1992), and Sawilowsky and Hillman (1992) was that the distributions selected to study were documented by Micceri (1989) to be prevalent in applied psychology and education research. Micceri noted that most of the early small samples research was conducted on theoretically expedient and mathematically well-known distributions, which unfortunately have little relevance to applied researchers.

The power portion of the above-mentioned studies, however, were restricted to modeling treatment effects in terms of shift in location parameter (i.e., simulated treatment effects) or a combination of shift in location parameter with a change in scale (which are also simulated treatment effects). A second restriction common to these studies (seven of the eight distributions examined in Sawilowsky & Blair, 1992, and the distribution studied in Sawilowsky & Hillman, 1992) was that data were sampled from data sets describing relatively continuous or smooth curves, as opposed to the discrete nature of ordinal data obtained from Likert scales.

To summarize the research to date, we briefly recount below the history of asymptotic procedures or Monte Carlo methods used to compare the parametric *t* test with competitors, primarily with the nonparametric Wilcoxon test:

- asymptotic studies based on theoretical (e.g., Gaussian, Cauchy, chi-square, exponential, *t*, uniform) continuous distributions (e.g., Chernoff & Savage, 1958; Dixon, 1954; Hodges & Lehmann, 1956);
- Monte Carlo studies based on theoretical continuous distributions and synthetic treatment effects modeled as shift in location (e.g., Blair & Higgins, 1980a, 1980b, 1985; Blair, Higgins, & Smitely, 1980; Neave & Granger, 1968; Posten, 1982; Randles & Wolfe, 1979; Van der Brink & Van der Brink, 1989), change in scale,[1] or shift in location with change in scale (e.g., Gibbons & Chakraborti, 1991; O'Brien, 1988; Zimmerman, 1987);
- Monte Carlo studies based on real continuous data sets and synthetic treatment effects modeled as shift in location (e.g., Bridge & Sawilowsky, 1997; Sawilowsky & Blair, 1992; Sawilowsky & Brown, 1991; Sawilowsky & Hillman, 1992) or shift in location with change in scale (i.e., Sawilowsky & Blair, 1992);
- a study based on theoretical discrete (Likert scale) data sets and synthetic treatment effects modeled as shift in location (i.e., Zumbo & Zimmerman, 1993, who called their procedure "experimental mathematics");
- a Monte Carlo study based on a real discrete (Likert scale) data set and synthetic treatment effects modeled as shift in location (i.e., Sawilowsky & Blair, 1992), or shift in location with change in scale (i.e., Sawilowsky & Blair, 1992).

## The Current Study

The purpose of the current study was to address the comparative power properties of the *t* test with the Wilcoxon rank-sum test using scores obtained from ordinal measurements on a 7-point Likert scale. The importance of this study should be evident, given the inordinate amount of assessment instruments used in the biomedical, behavioral, and social sciences, which are scaled at the ordinal level of measurement. Previous studies on Likert scaled data were restricted to an examination of robustness, were conducted on theoretical populations with simulated treatment effects,

[1] Although there are many citations in the literature with reference to change in scale, we do not cite examples because this is the so-called "Behrens–Fisher" problem. As noted by Sawilowsky and Blair (1992), we do not know of a treatment effect or naturally occurring condition that brings about a change in scale, while leaving the means exactly the same.

and, in the case of the two independent samples layout, were limited to 2-point to 5-point scales. The current study goes beyond these previously conducted studies, as the properties of the *t* test and the Wilcoxon test are compared with real Likert scale data sets and real treatment effects.

## *Method*

Monte Carlo techniques were used to sample eight distributions obtained from admit–discharge data sets of FIM scores. As suggested by Micceri (1989), these real data sets are taken to be representative of the admit and discharge populations (as opposed to some mathematically convenient distribution) associated with each FIM score. The FIM consists of a 7-point ordinal scale that designates major gradations in behavior ranging from 1 (*total dependence*) to 7 (*total independence*) for 18 areas of patient performance. The FIM consists of two main domains: motor and cognitive. The motor domain consists of 13 items assessing areas of self-care (e.g., dressing), transfers, and locomotion. The cognitive domain consists of five items comprising the Communication and Social Cognition subscales (Ottenbacher et al., 1996). As originally developed, ratings on all 18 items were summed into a single index to estimate overall burden of care (or functional ability; Stineman et al., 1996). However, FIM scores are often summed to give discipline-specific composite scores. In this case, a composite score was calculated from all FIM items assessing ADL-related tasks.

FIM scores were obtained by evaluating 714 geriatric patients from 1991 to 1994 who were admitted to a large midwestern rehabilitation hospital. Patients were evaluated using the FIM at the time of admission, and again at the time of discharge. Seven (1, 3, 4, 5, 6, 7, 13) of the 18 individual FIM score distributions were selected for further study and are depicted in Figures 1–7. The histograms of the remaining FIM score distributions were similar to these seven, and to conserve space, are not presented here. The composite scale mentioned in the preceding paragraph was used to represent a distribution of scores at a more continuous "interval-like" level of measurement (see Figure 8).

The difference obtained from independently sampling with replacement from the pretest (admit) and posttest (discharge) scores were used to represent real treatment effects, as opposed to artificially modeling treatments, as has been done in previous studies. Thus, scores from the admit data set were representative of pretest scores obtained by the patients, and scores from the discharge data set were representative of posttest scores. Patients received treatment regimens (e.g., physical therapy, occupational therapy, psychological counseling) during the intervening period. This intervention is presumed to be captured by the differences between the two distributions. Therefore, sampling independently from the two distribu-

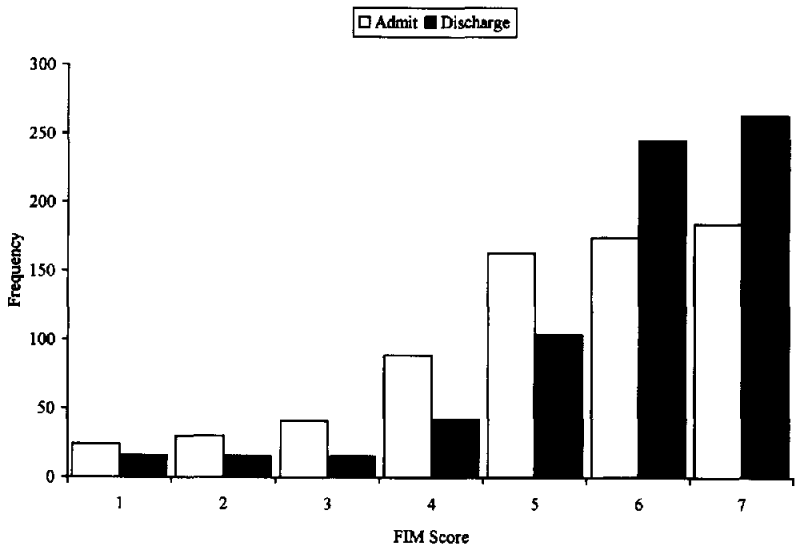


*Figure 1.* Distribution of Functional Independence Measure (FIM) Item 1 (i.e., eating skills) scores.

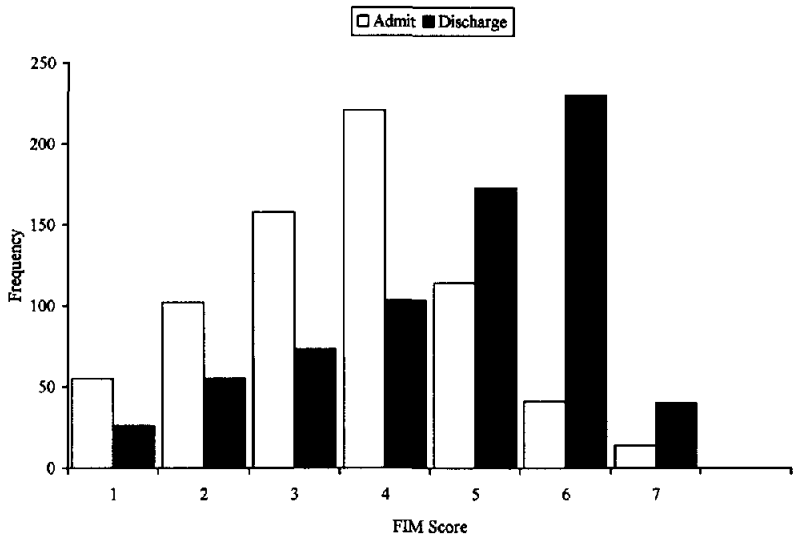


*Figure 2.* Distribution of Functional Independence Measure (FIM) Item 3 (i.e., bathing skills) scores.

tions obviates the need to model synthetic treatment effects in the Monte Carlo study.

FIM 1–6 scores represent self-care skills. According to Keith et al. (1987), FIM 1 is a measure of eating skills, such as "eating and drinking, including opening containers, pouring liquids, cutting meat, buttering bread" (p. 13). FIM 3 measures bathing skills, such as use of "tub, shower, or bed bath" (p. 13). FIM 4 relates to dressing the upper body, including "donning and removing prosthesis or orthosis, when applicable" (p. 13). FIM 5 is similar, except it refers to the lower body. FIM 6 measures perineal care. FIM 7 represents sphincter control and bladder management or "management of equipment necessary for emptying" (p. 13) one's bladder. FIM 13 concerns locomotion skills, specifically with regard to "going up and down 12 to 14 stairs (one flight)" (p. 13).

An Occupational Therapy (OT) subscale score was constructed as a composite score of seven FIM items that relate to domains commonly assessed by occupa-

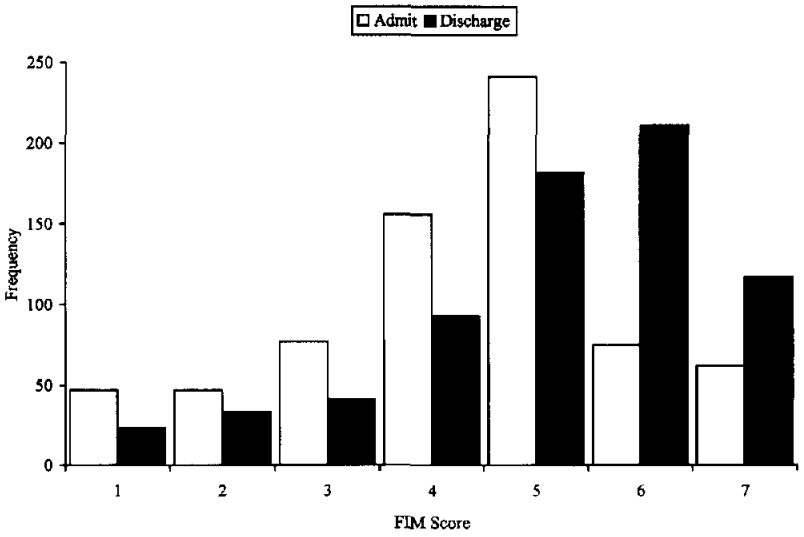


*Figure 3.* Distribution of Functional Independence Measure (FIM) Item 4 (i.e., dressing the upper body) scores.

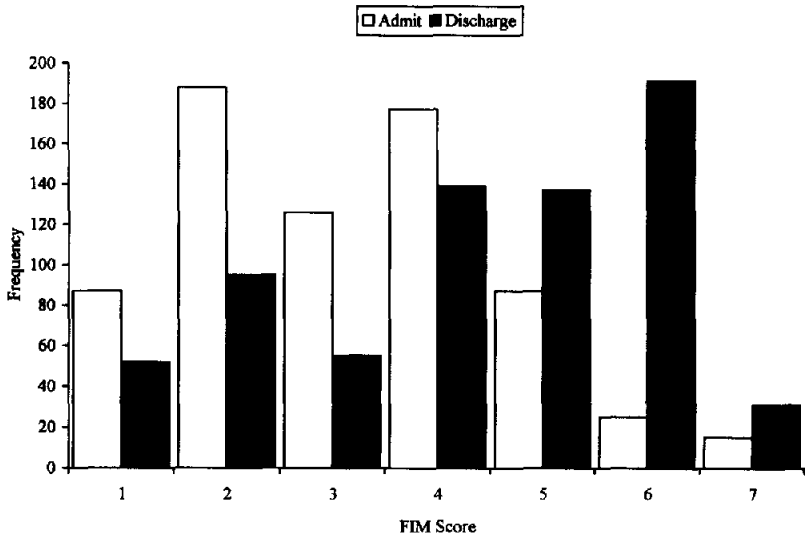


*Figure 4.* Distribution of Functional Independence Measure (FIM) Item 5 (i.e., dressing the lower body) scores.

tional therapists. The items on this subscale include FIM 1, 2, 3, 4, 5, 10, and 11. According to Keith et al. (1987), FIM 2 is a self-care skill item addressing grooming, including "oral care, hair care, washing hands and face, shaving, applying make-up" (p. 13). FIM 10 and 11 are mobility skills. FIM 10 refers to transfers on and off the toilet, and FIM 11 concerns transferring to a "tub or shower stall" (p. 13).

A Fortran 77 program was written for the MS-DOS compatible Pentium PC (with Intel validated floating point unit) accessing International Mathematical & Statistical Library (1987) subroutines to sample with replacement for sample sizes $n_1 = n_2 = (10,10)$, (20,20), (30,30), (40,40), and (60,60) from each respective FIM admit and FIM discharge score distribution. The independent samples $t$ test and its nonparametric counterpart, the Wilcoxon rank-sum test, were calculated for each FIM distribution at each sample size.

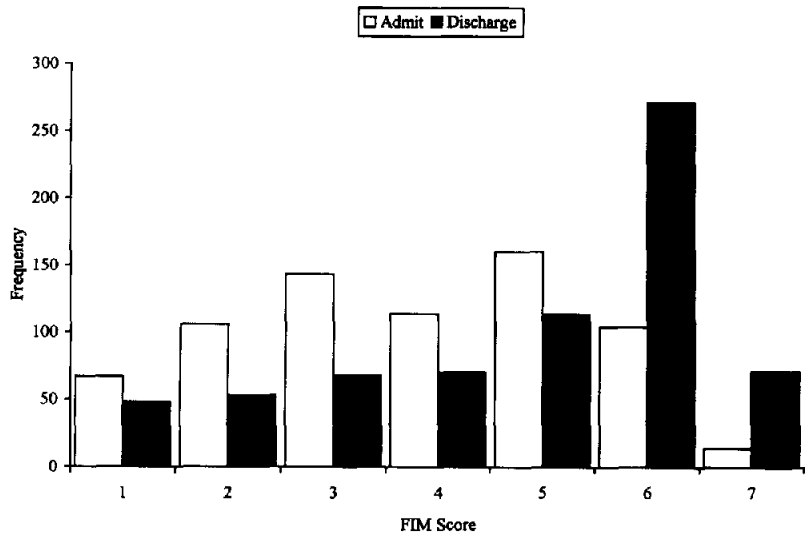


*Figure 5.* Distribution of Functional Independence Measure (FIM) Item 6 (i.e., perineal care) scores.

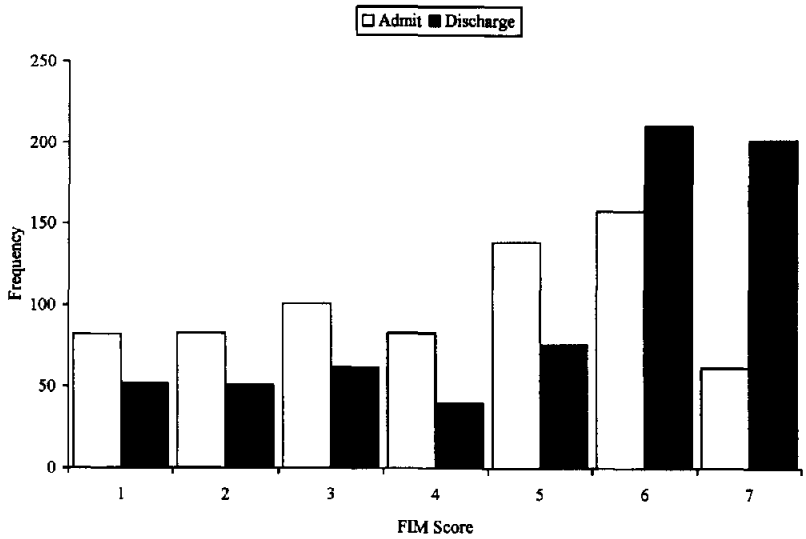


*Figure 6.* Distribution of Functional Independence Measure (FIM) Item 7 (i.e., bladder management) scores.

(Note that the Wilcoxon signed-rank test is inappropriate, despite the correlated nature of pretest–posttest scores, because the admit and discharge data sets were independently sampled. That is, pairs of admit–discharge data were not sampled; rather, scores for a particular pretest sample were obtained by random selection from the distribution of admit scores, and scores representing the posttest were independently sampled from the discharge scores. We randomly sampled, with replacement, 1,000,000 scores from each data set and calculated Spearman's rho; we repeated this process 1,000 times. The long run average of the correlation of scores from the pretest distribution with those independently sampled from the posttest distribution was −0.00083.)

Results of each respective test were recorded. The number of replications per experiment was 10,000. Then, the proportion of rejections was calculated, which is the power for each statistic. The one-tailed power of the independent samples *t* test and the Wil-

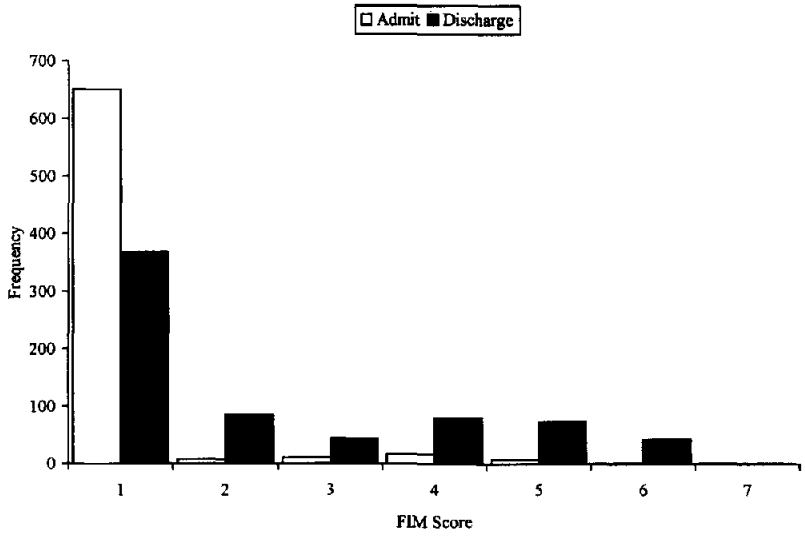


*Figure 7.* Distribution of Functional Independence Measure (FIM) Item 13 (i.e., wheelchair propulsion skills) scores.

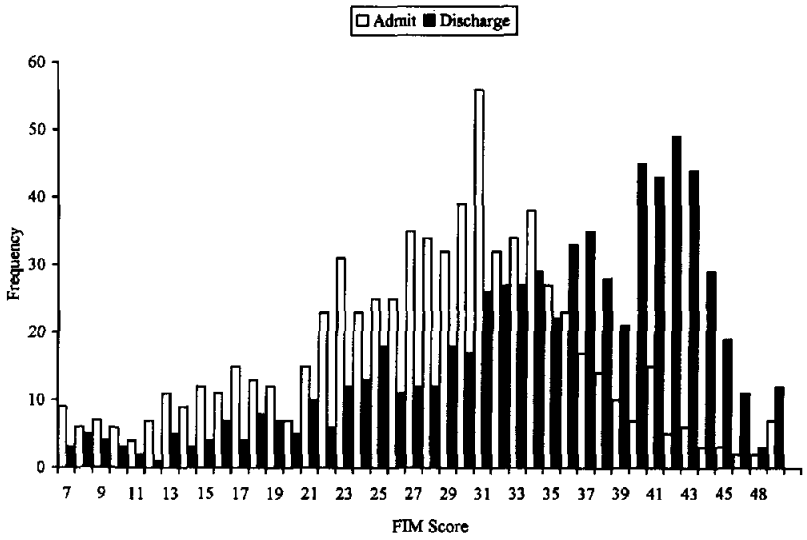


*Figure 8.* Distribution of Functional Independence Measure (FIM) Occupational Therapy subscale composite score.

coxon rank-sum test were compared at the 0.01, 0.05, and 0.10 alpha levels.

### *Results*

Figures 1–8 depict the distributions for FIM 1, 3, 4, 5, 6, 7, 13, and the OT composite scores. Power comparisons for the *t* and Wilcoxon statistic are presented in Table 1. Results for FIM 1 (eating) indicate that the Wilcoxon statistic held modest power advantages over the *t* test at each sample size with the largest difference at sample size $n_1 = n_2 = 60$, with a power advantage of about 0.1. Results for FIM 3 (bathing) are similar, as the Wilcoxon held modest advantages over the *t* test at every sample size. Results for FIM 4 (dressing upper body), 5 (dressing lower body), and 6 (toileting) scores are congruent with the first two distributions. That is, the Wilcoxon held moderate power advantages over the *t* test for each distribution and sample size. The results for FIM 7 (bladder management) scores indicate the Wilcoxon holds a power advantage of 0.12 over the *t* test at sample size (30,30)

Table 1
*Power Comparisons for the t and Wilcoxon Statistic*

| | | FIM data set | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *n*1, *n*2 | Statistic | 1 | 3 | 4 | 5 | 6 | 7 | 13 | OT |
| 10, 10 | *t* | .144 | .318 | .176 | .313 | .209 | .171 | .348 | .162 |
| | W | .149 | .340 | .198 | .314 | .235 | .201 | .385 | .238 |
| 20, 20 | *t* | .229 | .547 | .288 | .546 | .358 | .287 | .779 | .259 |
| | W | .259 | .608 | .349 | .560 | .426 | .364 | .790 | .429 |
| 30, 30 | *t* | .318 | .728 | .399 | .726 | .486 | .396 | .936 | .359 |
| | W | .366 | .788 | .489 | .742 | .580 | .512 | .939 | .588 |
| 40, 40 | *t* | .394 | .833 | .508 | .841 | .602 | .499 | .984 | .460 |
| | W | .462 | .885 | .622 | .856 | .706 | .634 | .983 | .719 |
| 60, 60 | *t* | .544 | .951 | .677 | .952 | .782 | .672 | .999 | .606 |
| | W | .641 | .974 | .792 | .960 | .873 | .809 | .994 | .873 |

*Note.* One tailed power of the *t* and Wilcoxon rank–sum (W) statistics for data sampled from various Functional Independence Measure (FIM) data sets and sample sizes (*n*1, *n*2), nominal $\alpha = 0.05$. OT = Occupational Therapy composite score.

and 0.14 at sample size (40,40) and (60,60). The results for FIM 13 indicate the Wilcoxon statistic maintains a small advantage over the $t$ test until sample size reaches (40,40).

Table 1 also contains the results for the OT composite of seven selected FIM scores (FIM 1–5, which are self-care skills, and 10 and 11, which are mobility skills). Results for this distribution are remarkable, with the Wilcoxon displaying significant increases in power over the $t$ test at sample sizes (30,30), (40,40), and (60,60). At sample size (30,30) there was a difference in power of 0.23, 0.26 more power at sample size (40,40), and 0.27 more statistical power at sample size (60,60).

Results for the 0.10 and 0.01 alpha level were similar. To conserve space, we do not present them here. A complete set of tables are available from Michael J. Nanna. A copy of the Fortran program is available from Shlomo S. Sawilowsky, or may be downloaded from the World Wide Web at edstat2.coe.wayne.edu or 141.217.33.243.

Two secondary findings in this study replicated results found by Blair (1981) and Sawilowsky and Blair (1992) but have not been publicized. First, many statistics textbook authors stated that a condition conducive to the use of nonparametric tests is when sample size is small. Ostensibly, this will allow the Central Limit Theorem to rehabilitate the $t$ test when sample sizes ($n_1 = n_2$) reach at least 30. Yet, the Wilcoxon test achieves optimal power advantages over the $t$ test as the sample size increases, indicating that it should be used for large sample sizes as well. Second, the greatest power advantages of the Wilcoxon test occurred with data sampled from the OT subscale. The OT subscale is relatively continuous (i.e., 42 scale points, with a minimum OT score of 7 and a maximum OT score of 49) in comparison with the 7-point Likert scaling of the other FIM scales. Thus, it is a misconception that the Wilcoxon test is preferable over the $t$ test only when analyzing ranked data.

### *Discussion*

A shortcoming of previous Monte Carlo studies conducted to address the robustness and comparative power of the $t$ test and Wilcoxon test is that they have relied on simulated populations (usually continuous distributions of mathematical interest) and simulated treatment effects (such as simple shifts in location or changes in both location and scale). However, the practical applications of these studies to clinical evaluation and measurement had been unclear and there remained questions as to whether the results of these studies could be applied to data gathered in realistic settings. Applied data are often ordinal level Likert scale data and certainly are not represented by the Gaussian distribution.

The Wilcoxon rank-sum test outperformed the $t$ test for almost every sample size and alpha level examined. Although the superiority in power the Wilcoxon test held over the $t$ test was modest in many situations (less than 0.10), in some instances the Wilcoxon test achieved power advantages over the $t$ test that were substantial, as high as 0.27. This compares favorably with the results of Sawilowsky and Blair (1992), who found power differences of about 0.26 in favor of the Wilcoxon rank-sum test over the $t$ test for a continuous extreme asymmetric real data set discussed by Micceri (1989).

Although the $t$ test was relatively robust with respect to Type I errors for the conditions studied, it was never more powerful than the Wilcoxon statistic. It has generally been assumed that because the parametric independent samples $t$ test is robust with respect to Type II errors (i.e., the power obtained under normality is preserved under nonnormality) it must therefore be more powerful than nonparametric counterparts under normality (e.g., Boneau, 1960; Glass, Peckham, & Sanders, 1972). The importance of this misconception is exacerbated by findings of Bradley (1977, 1978), Blair (1981), Blair and Higgins (1980a, 1980b), Micceri (1989), Pearson and Please (1975), Still and White (1981), and Tan (1982), among many others, who have shown that normality is the exception rather than the norm in applied research, such as the FIM data sets in this study.

What do the results of this study mean to the applied researcher? For real data sets such as those examined here, consider the following two examples.[2] If the researcher was trying to detect a treatment of magnitude effect size = $0.20\sigma$, 16 participants are necessary ($n = 8$ per group) for a power level of 0.10 for the $t$ test. To achieve power equal to the typical advantage held by the Wilcoxon test when the treatment magnitude is $0.20\sigma$, we require 52 more participants (68 participants with $n = 34$ per group). As a second example, if the treatment magnitude is $0.50\sigma$,

[2] As pointed out by an anonymous reviewer, these illustrations assume the data set has zero mean and unit variance.

a power level of .50 for the $t$ test requires 44 participants ($n = 22$ per group). To obtain power equal to the maximum advantage found in this study by the Wilcoxon test, an additional 48 participants would be required for the $t$ test (92 participants, $n = 46$ per group). The cost of increasing sample size is well known; also, in many research contexts large sample sizes are simply not available.

In addition to contributing to the literature concerning the comparative power properties of the $t$ test and Wilcoxon rank-sum test, these results suggest an increase in the ability to effectively assess functional improvements as measured by the FIM. There is increasing attention being given to identifying and quantifying functional outcome—a situation necessitated by third-party payers (e.g., insurance organizations, governmental agencies) developing stricter criteria for funding and reimbursement. Moreover, there has been considerable focus on the field of rehabilitation medicine because traditionally there has not been strict criteria or standardized measurement instruments for documenting change in a patient's functional status from admit to discharge. Therefore, not only is it important to develop treatment modalities that facilitate a patient's recovery in a most cost-efficient fashion and to develop reliable measures to assess functional status, but it is also important to choose more efficient ways of detecting changes in functional outcome. These concerns apply throughout behavioral and social science research.

## References


Ashworth, B. (1964). Preliminary trial of carisoprodol in multiple sclerosis. *Practitioner, 192,* 540–542.

Baldrige, R. B. (1993). Functional assessment of measurement. *Neurology Report, 17*(4), 3–10.

Batavia, A. I. (1988). *The payment of medical rehabilitation services. Current mechanisms and potential models.* Chicago: American Hospital Association.

Bevan, M. F., Denton, J. Q., & Myers, J. L. (1974). The robustness of the $F$ test to violations of continuity and form of treatment populations. *British Journal of Mathematical and Statistical Psychology, 27,* 199–204.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research, 51,* 499–507.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of Wilcoxon's rank-sum statistic to that of student's $t$ statistic under various non-normal distributions. *Journal of Educational Statistics, 5,* 309–335.

Blair, R. C., & Higgins, J. J. (1980b). The power of $t$ and Wilcoxon statistics. *Evaluation Review, 4,* 645–656.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples $t$ test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin, 97,* 119–128.

Blair, R. C., Higgins, J. J., & Smitely, W. D. S. (1980). On the relative power of the $\cup$ and $t$ tests. *British Journal of Mathematical and Statistical Psychology, 33,* 114–120.

Bohannon, R. W., & Smith, M. B. (1987). Interrater reliability of a modified Ashworth Scale of muscle spasticity. *Physical Therapy, 67,* 206–207.

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

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31*(4), 147–150.

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

Bridge, P. D., & Sawilowsky, S. (1997). Revisiting the $t$ test on ranks as an alternative to the Wilcoxon rank-sum test. *Perceptual and Motor Skills, 85,* 399–402.

Brynes, M. B., & Powers, F. F. (1989). FIM: Its use in identifying rehabilitation needs in the head injured patient. *Journal of Neuroscience Nursing, 21,* 61–63.

Chernoff, H., & Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics, 29,* 972–999.

Dinnerstein, A. J., Lowenthal, M., & Dexter, M. (1965). Evaluation of a rating scale of ability in activities in daily living. *Archives of Physical Medicine and Rehabilitation, 46,* 579–584.

Dixon, W. J. (1954). Power under normality of several nonparametric tests. *Annals of Mathematical Statistics, 25,* 610–614.

Findley, T. W. (1991). Research in physical medicine and rehabilitation: IX. Primary data analysis. *American Journal of Physical Medicine & Rehabilitation, 70* (Suppl.), S84–S93.

Gibbons, J. D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's $t$, and alternative $t$ tests for means of normal distributions. *Journal of Experimental Education, 59,* 258–267.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Con-

sequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42,* 237–288.

Granger, C. V., Cotter, A. C., Hamilton, B. B., Fiedler, R. C., & Hens, M. M. (1990). Functional assessment scales: A study of persons with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation, 71,* 870–875.

Granger, C. V., Hamilton, B. B., Keith, R. A., Zielezny, M., & Sherwin, F. S. (1986). Advances in functional assessment for medical rehabilitation. *Topics in Geriatric Rehabilitation, 1,* 59–74.

Hamilton, B. B., Laughlin, J. A., Granger, C. V., & Kayton, R. M. (1991). Interrater agreement of the seven level Functional Independence Measure (FIM). *Archives of Physical Medicine and Rehabilitation, 72,* 790.

Harada, N., Sofaer, S., & Kominski, G. (1993). Functional status outcomes in rehabilitation: Implications for prospective payment. *Medical Care, 31,* 345–357.

Harvey, R. F., & Jellinek, H. M. (1981). Functional performance assessment: A program approach. *Archives of Physical Medicine and Rehabilitation, 62,* 456–460.

Heeren, T., & D'Agostino, R. (1987). Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine, 6,* 79–90.

Hodges, J. C., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the *t* test. *Annals of Mathematical Statistics, 27,* 324–335.

Hsu, T. C., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the *F*-test. *American Educational Research Journal, 6,* 515–527.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology, 34,* 384–389.

International Mathematical & Statistical Library. (1987). *International Mathematical & Statistical Library: User's Manual.* Fortran subroutines for statistical analysis (version 1.0). Houston, TX: Author.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged: The Index of ADL. A standardized measure of the biological and psychosocial function. *Journal of the American Medical Association, 185,* 914–919.

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). The Functional Independence Measure: A new tool for rehabilitation. In M. G. Eisenberg & R. C. Grzesiak (Eds.), *Advances in clinical rehabilitation* (Vol. 1, pp. 6–18). New York: Springer.

Mahoney, F. I., & Barthel, D. W. (1965). Functional evaluation: The Barthel Index. *Maryland State Medical Journal, 14,* 61–65.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Neave, H. R., & Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics, 10,* 509–522.

O'Brien, P. C. (1988). Comparing two samples: Extensions of the *t,* rank-sum, and log-rank tests. *Journal of the American Statistical Association, 83,* 52–61.

Ottenbacher, K. J., Hsu, Y., Granger, C. V., & Fiedler, R. C. (1996). The reliability of the Functional Independence Measure: A quantitative review. *Archives of Physical Medicine & Rehabilitation, 77,* 1226–1232.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62,* 223–241.

Posten, H. O. (1982). Two-sample Wilcoxon power over the Pearson system and comparisons with the *t* test. *Journal of Statistical Computation and Simulation, 16,* 1–18.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric tests.* New York: Wiley.

Sarno, J. E., Sarno, M. T., & Levita, E. (1973). The Functional Life Scale. *Archives of Physical Medicine and Rehabilitation Medicine, 54,* 214–220.

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

Sawilowsky, S. S. (1993). Comments on using alternatives to normal theory statistics in social and behavioral science. *Canadian Psychology, 34,* 432–439.

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, 111,* 352–360.

Sawilowsky, S. S., & Brown, M. T. (1991). On using the *t* test on ranks as an alternative to the Wilcoxon test. *Perceptual and Motor Skills, 72,* 360–362.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples *t*-test under a prevalent psychometric measure distribution. *Journal of Consulting & Clinical Psychology, 60,* 240–243.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Schoening, H. A., & Inersen, I. A. (1968). Numerical scoring of self-care status: A study of the Kenny Self-Care Evaluation. *Archives of Physical Medicine and Rehabilitation, 49,* 221–229.

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

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

Still, A. W., & White, A. P. (1981). The approximate randomization test as an alternative to the *F* test in analysis of variance. *British Journal of Mathematical and Statistical Psychology, 34,* 243–252.

Stineman, M. G., Escarce, J. E., Goin, J. E., Hamilton, B. B., Granger, C. V., & Williams, S. V. (1994). A case-mix classification system for medical rehabilitation. *Medical Care, 32,* 366–379.

Stineman, M. G., Shea, J. A., Jette, A., Tassoni, C. J., Ottenbacher, K. J., Fiedler, R., & Granger, C. V. (1996). The Functional Independence Measure: Tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of Physical Medicine & Rehabilitation, 77,* 1101–1108.

Tan, W. Y. (1982). Sampling distributions and robustness of *t, F* and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics, 11,* 2485–2511.

Van der Brink, W. P., & Van der Brink, G. J. (1989). A comparison of the power of the *t* test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical and Statistical Psychology, 42,* 183–189.

Wilkerson, D. L., Batavia, A. I., & DeJong, G. (1992). Use of functional status measures for payment of medical rehabilitation services. *Archives of Physical Medicine & Rehabilitation, 73,* 1111–1120.

Zimmerman, D. W. (1987). Comparative power of student *t* test and Mann-Whitney ∪ test for unequal sample sizes and variances. *Journal of Experimental Education, 55,* 171–174.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34,* 390–400.

Received November 12, 1995
Revision received October 9, 1996
Accepted January 23, 1997 ■