11-1-2009

Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores

Shira R. Solomon
CNA Education, Solomons@cna.org

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 448-462

1538 – 9472/09/$95.00

Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores

Shira R. Solomon
CNA Education

Shlomo S. Sawilowsky
Wayne State University

The purpose of this article is to provide an empirical comparison of rank-based normalization methods for standardized test scores. A series of Monte Carlo simulations were performed to compare the Blom, Tukey, Van der Waerden and Rankit approximations in terms of achieving the T score's specified mean and standard deviation and unit normal skewness and kurtosis. All four normalization methods were accurate on the mean but were variably inaccurate on the standard deviation. Overall, deviation from the target moments was pronounced for the even moments but slight for the odd moments. Rankit emerged as the most accurate method among all sample sizes and distributions, thus it should be the default selection for score normalization in the social and behavioral sciences. However, small samples and skewed distributions degrade the performance of all methods, and practitioners should take these conditions into account when making decisions based on standardized test scores.

Key words: Normalization; normalizing transformations; T scores; test scoring; ranking methods; Rankit; Blom; Tukey; Van der Waerden; Monte Carlo.

Introduction

Standardization and normalization are two ways of defining the frame of reference for a distribution of test scores. Both types of score conversions, or transformations, mathematically modify raw score values (Osborne, 2002). The defining feature of standard scores is that they use standard deviations to describe scores' distance from the mean, thereby creating equal units of measure within a given score distribution. Standard scores may be modified to change the scale's number system (Angoff, 1984), but unless distributions of standard scores are normalized, they will retain the shape of the original score distribution. Therefore, standardization may enable effective analysis of individual scores within a single test, but normalization is needed for meaningful comparisons between tests.

Shira R. Solomon is a Research Analyst. Email: solomons@cna.org. Shlomo S. Sawilowsky is Professor of Evaluation and Research. Email: shlomo@wayne.edu.

The Problem of Non-continuous Data in Educational and Psychological Testing

Knowledge, intellectual ability, and personality are psychological objects that can only be measured indirectly, not by direct observation (Dunn-Rankin, 1983). The scales that describe them are hierarchical—they result in higher or lower scores—but these scores do not express exact quantities of test-takers' proficiency or attitudes. Ordinal test items such as Likert scales result in raw scores that are meaningless without purposeful statistical interpretation (Nanna & Sawilowsky, 1998). Measures with unevenly spaced increments interfere with the interpretation of test scores against performance benchmarks, the longitudinal linking of test editions, and the equating of parallel forms of large-scale tests (Aiken, 1987). They also threaten the robustness and power of the parametric statistical procedures that are conventionally used to analyze standardized test scores (Friedman, 1937; Sawilowsky & Blair, 1992).

Statisticians have been transforming ordinal data into a continuous scale since Fisher and Yates tabled the normal deviates in 1938. Wimberly (1975) favored rank-based

transformations to other normalizing transformations such as those based on logarithms, exponents, or roots for their superior accuracy among random scores of different variables. Rank-based transformations not only attempt to equate the means and homogenize the variance of test score distributions, they also aim to create conformity in the third and fourth moments, skewness and kurtosis. Central tendency and variability have clear implications for test score distributions.

The most prominent of the rank-based normalization procedures, based on their inclusion in widely used statistical software (e.g., SPSS, 2006) are those attributed to Van der Waerden, Blom, Bliss (the Rankit procedure), and Tukey. Van der Waerden's formula (1952, 1953a, 1953b; Lehmann, 1975) was thought to improve on percentiles by computing quantiles (equal unit portions under the normal curve corresponding with the number of observations in a sample) not strictly on the basis of ranks, but according to the rank of a given score value relative to the sample size (Conover, 1980). Blom's formula (1958) responds to the curvilinear relationship between a score's rank in a sample and its normal deviate. Because "Blom conjectured that α always lies in the interval (0·33, 0·50)," explained Harter, "he suggested the use of α = 3/8 as a compromise value" (1961, p.154). Bliss, Greenwood, and White (1956) credited Ipsen and Jerne (1944) with coining the term "rankit," but Bliss is credited with developing the technique as it is now used. Bliss, et al. refined this approximation in their study of the effects of different insecticides and fungicides on the flavor of apples. Its design drew on Scheffé's advancements in paired comparison research, which sought to account for magnitude and direction of preference, in addition to preference itself. Tukey may have proposed his formula, which he characterized as "simple and surely an adequate approximation to what is claimed to be optimum" (1962, p.22), as a refinement of Blom's.

These procedures have been explored to various degrees in the context of hypothesis testing, where the focus is necessarily on their properties in the tails of a distribution. In the context of standardized testing, however, the body of the distribution—that is, the 95% of the curve that lies between the tails—is the focus. Practitioners need to know how accurately each method normalizes non-theoretical score distributions. Solomon (2008) produced the first empirical comparison of the Van der Waerden, Blom, Tukey, and Rankit methods as they apply to standardized testing. This study sought to demonstrate their accuracy under a variety of sample size and distributional conditions.

Table 1: Chronology of Rank-Based Normalization Procedure Development

Procedure	Year	Formula
Van der Waerden	1952	r* / (n + 1)
Blom	1954	(r - 3/8) / (n + 1/4)
Rankit	1956	(r - 1/2) / n
Tukey	1962	(r - 1/3) / (n + 1/3)

*where r is the rank, ranging from 1 to n

Blom, Tukey, Van der Waerden, and Rankit each contribute a formula that approximates a normal distribution, given a set of raw scores or non-normalized standard scores. However, the formulas themselves had not been systematically compared for their first four moments' accuracy in terms of normally distributed data. Nor had they been compared in the harsher glare of non-normal distributions, which are prevalent in the fields of education and psychology (Micceri, 1989). Small samples are also common in real data and are known to have different statistical properties than large samples (Conover, 1980). In general, real data can be assumed to behave differently than data that is based on theoretical distributions, even if these are non-normal (Stigler, 1977).

A series of Monte Carlo simulations drew samples of different sizes from eight unique, empirically established population distributions. These eight distributions, though extensive in their representation of real achievement and psychometric test scores, do not represent all possible distributions that could occur in educational and psychological testing or in social and behavioral science investigations

more generally. Nor do the sample sizes represent every possible increment. However, both the sample size increments and the range of distributional types are assumed to be sufficient for the purpose of outlining the absolute and comparative accuracy of these normalizing transformations in real settings. Although the interpretation of results need not be restricted to educational and psychological data, similar distributional types may be most often found in these domains.

For normally distributed variables, the standardization process begins with the *Z* score transformation, which produces a mean of 0 and a standard deviation of 1 (Walker & Lev, 1969; Mehrens & Lehmann, 1980; Hinkle, Wiersma, & Jurs, 2003). *Z* scores are produced by dividing the deviation score (the difference between raw scores and the mean of their distribution) by the standard deviation: $Z = (X - \mu) / \sigma$. However, *Z* scores can be difficult to interpret due to decimals and negative numbers. Because 95% of the scores fall between -3 and +3, small changes in decimals may imply large changes in performance. Also, because half the scores are negative, it may appear to the uninitiated that half of the examinees obtained an extremely poor outcome.

Linear versus Area Transformations

Linear scaling remedies these problems by multiplying standard scores by a number large enough to render decimal places trivial, then adding a number large enough to eliminate negative numbers. Although standard scores may be assigned any mean and standard deviation through linear scaling, the *T* score scale ($S_T = 10Z + 50$) has dominated the scoring systems of social and behavioral science tests for a century (Cronbach, 1976; Kline, 2000; McCall, 1939). In the case of a normally distributed variable, the resulting *T*-scaled standard scores would have a mean of 50 and a standard deviation of 10. In the context of standardized testing, however, *T* scores refer not to *T*-scaled standard scores but to *T*-scaled normal scores. In the *T* score formula, *Z* refers to a score's location on a unit normal distribution—its normal deviate—not its place within the testing population.

Scaling standard scores of achievement and psychometric tests has limited value. Most educational and psychological measurements are ordinal (Lester & Bishop, 2000), but standard scores can only be obtained for continuous data because they require computation of the mean. Furthermore, linear transformations retain the shape of the original distribution. If a variable's original distribution is Gaussian, its transformed distribution will also be normal. If an observed distribution manifests substantial skew, excessive or too little kurtosis, or multimodality, these non-Gaussian features will be maintained in the transformed distribution.

This is problematic for a wide range of practitioners because it is common practice for educators to compare or combine scores on separate tests and for testing companies to reference new versions of their tests to earlier versions. Standard scores such as *Z* will not suffice for these purposes because they do not account for differing score distributions between tests. Comparing scores from a symmetric distribution with those from a negatively skewed distribution, for example, will give more weight to the scores at the lower range of the skewed curve than to those at the lower range of the symmetric curve (Horst, 1931). Normalizing transformations are used to avoid biasing test score interpretation due to heteroscedastic and asymmetrical score distributions.

Non-normality Observed

According to Nunnally (1978), "test scores are seldom normally distributed" (p.160). Micceri (1989) demonstrated the extent of this phenomenon in the social and behavioral sciences by evaluating the distributional characteristics of 440 real data sets collected from the fields of education and psychology. Standardized scores from national, statewide, and districtwide test scores accounted for 40% of them. Sources included the Comprehensive Test of Basic Skills (CTBS), the California Achievement Tests, the Comprehensive Assessment Program, the Stanford Reading tests, the Scholastic Aptitude Tests (SATs), the College Board subject area tests, the American College Tests (ACTs), the Graduate Record Examinations (GREs), Florida Teacher Certification Examinations for adults, and

Florida State Assessment Program test scores for 3rd, 5th, 8th, 10th, and 11th grades.

Micceri summarized the tail weights, asymmetry, modality, and digit preferences for the ability measures, psychometric measures, criterion/mastery measures, and gain scores. Over the 440 data sets, Micceri found that only 19 (4.3%) approximated the normal distribution. No achievement measure's scores exhibited symmetry, smoothness, unimodality, or tail weights that were similar to the Gaussian distribution. Underscoring the conclusion that normality is virtually nonexistent in educational and psychological data, none of the 440 data sets passed the Kolmogorov-Smirnov test of normality at alpha = .01, including the 19 that were relatively symmetric with light tails. The data collected from this study highlight the prevalence of non-normality in real social and behavioral science data sets.

Furthermore, it is unlikely that the central limit theorem will rehabilitate the demonstrated prevalence of non-normal data sets in applied settings. Although sample means may increasingly approximate the normal distribution as sample sizes increase (Student, 1908), it is wrong to assume that the original population of scores is normally distributed. According to Friedman (1937), "this is especially apt to be the case with social and economic data, where the normal distribution is likely to be the exception rather than the rule" (p.675).

There has been considerable empirical evidence that raw and standardized test scores are non-normally distributed in the social and behavioral sciences. In addition to Micceri (1989), numerous authors have raised concerns regarding the assumption of normally distributed data (Pearson, 1895; Wilson & Hilferty, 1929; Allport, 1934; Simon, 1955; Tukey & McLaughlin, 1963; Andrews et al., 1972; Pearson & Please, 1975; Stigler, 1977; Bradley, 1978; Tapia & Thompson, 1978; Tan, 1982; Sawilowsky & Blair, 1992). The prevalence of non-normal distributions in education, psychology, and related disciplines calls for a closer look at transformation procedures in the domain of achievement and psychometric test scoring.

The Importance of T Scores for the Interpretation of Standardized Tests

Standardized test scores are notoriously difficult to interpret (Chang, 2006; Kolen and Brennan, 2004; Micceri, 1990; Petersen, Kolen, and Hoover, 1989). Most test-takers, parents, and even many educators, would be at a loss to explain exactly what a score of 39, 73, or 428 means in conventional terms, such as pass/fail, percentage of questions answered correctly, or performance relative to other test-takers. Despite the opaqueness of *T* scores relative to these conventional criteria, they have the advantage of being the most familiar normal score scale, thus facilitating score interpretation. Most normal score systems are assigned means and standard deviations that correspond with the *T* score. For example, the College Entrance Board's *Scholastic Aptitude Test* (SAT) Verbal and Mathematical sections are scaled to a mean of 500 and a standard deviation of 100. *T* scores fall between 20 and 80 and SAT scores fall between 200 and 800. The *T* score scale facilitates the interpretation of test scores from any number of different metrics, few of which would be familiar to a test taker, teacher, or administrator, and gives them a common framework.

The importance of transforming normal scores to a scale that preserves a mean of 50 and a standard deviation of 10 calls for an empirical comparison of normalizing transformations. This study experimentally demonstrates the relative accuracy of the Blom, Tukey, Van der Waerden, and Rankit approximations for the purpose of normalizing test scores. It compares their accuracy in terms of achieving the *T* score's specified mean and standard deviation and unit normal skewness and kurtosis, among small and large sample sizes in an array of real, non-normal distributions.

Methodology

A Fortran program was written to compute normal scores using the four rank-based normalization formulas under investigation. Fortran was chosen for its large processing capacity and speed of execution. This is important for Monte Carlo simulations, which typically require from thousands to millions of iterations.

Normal scores were computed for each successive iteration of randomly sampled raw scores drawn from various real data sets. The resulting normal scores were then scaled to the *T*. The first four moments of the distribution were calculated from these *T* scores for each of the 14 sample sizes in each of the eight populations. Absolute values were computed by subtracting *T* score means from 50, standard deviations from 10, skewness values from 0, and kurtosis values from 3. These absolute values were sorted into like bins and ranked in order of proximity to the target moments. The values and ranks were averaged over the results from 10,000 simulations and reported in complete tables (Solomon, 2008). Average root mean square (RMS) values and ranks were also computed and reported for the target moments. This paper summarizes the values and ranks for absolute deviation values and RMS, or magnitude of deviation. Together, deviation values and magnitude of deviation describe the accuracy and stability of the Blom, Tukey, Van der Waerden, and Rankit approximations in attaining the first four moments of the normal distribution.

Sample Sizes and Iterations

Simulations were conducted on samples of size n = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200, 500, and 1,000 that were randomly selected from each of the eight Micceri (1989) data sets. Ten-thousand (10,000) iterations were performed to break any ties up to three decimal places.

Achievement and Psychometric Distributions

Micceri (1989) computed three indices of symmetry/asymmetry and two indices of tail weight for each of the 440 large data sets he examined (for 70% of which, n $\geq$ 1,000), grouped by data type: achievement/ability (accounting for 231 of the measures), psychometric (125), criterion/mastery (35), and gain scores (49). Eight distributions were identified based on symmetry, tail weight contamination, propensity scores, and modality. Sawilowsky, Blair, and Micceri (1990) translated these results into a Fortran subroutine using achievement and psychometric measures that best represented the distributional characteristics described by Micceri (1989).

The following five distributions were drawn from achievement measures: Smooth Symmetric, Discrete Mass at Zero, Extreme Asymmetric – Growth, Digit Preference, and Multimodal Lumpy. Mass at Zero with Gap, Extreme Asymmetric – Decay, and Extreme Bimodal were drawn from psychometric measures. All eight achievement and psychometric distributions are nonnormal. These distributions are described in Table 2 and graphically depicted in Figure 1.

Results

The purpose of this study was to compare the accuracy of the Blom, Tukey, Van der Waerden, and Rankit approximations in attaining the target moments of the normal distribution. Tables 3, 4, and 5 present these results. Table 3 summarizes the major findings according to moment, sample size, and distribution. It presents values and simplified ranks for the accuracy of the four normalizing methods on the first measure, deviation from target moment. For example, the *T* score's target standard deviation is 10. Therefore, two methods that produce a standard deviation of 9.8 or 10.2 would have the same absolute deviation value: 0.2. The highest ranked method for each condition is the most accurate, having the smallest absolute deviation value over 10,000 Monte Carlo repetitions. It is possible to assign ranks on the mean despite the accuracy of all four normalization methods because differences begin to appear at six decimal places. However, all numbers are rounded to the third decimal place in the tables.

Table 3 shows that rank-based normalizing methods are less accurate on the standard deviation than on the mean, skewness, or kurtosis. Furthermore, the standard deviation has more immediate relevance to the interpretation of test scores than the higher moments. For these reasons, Tables 4 and 5 and Figures 2 and 3 restrict their focus to the methods' performance on the standard deviation. Table 4 summarizes the methods' proximity to the target standard deviation by distribution type. Table 5 reports proximity for all eight distributions.

Table 2: Basic Characteristics of Eight Non-normal Distributions

	Range	Mean	Median	Variance	Skewness	Kurtosis
	Achievement					
1. Smooth Symmetric	$0 \le x \le 27$	13.19	13.00	24.11	0.01	2.66
2. Discrete Mass at Zero	$0 \le x \le 27$	12.92	13.00	19.54	-0.03	3.31
3. Extreme Asymmetric – Growth	$4 \le x \le 30$	24.50	27.00	33.52	-1.33	4.11
4. Digit Preference	$420 \le x \le 635$	536.95	535.00	1416.77	-0.07	2.76
5. Multimodal Lumpy	$0 \le x \le 43$	21.15	18.00	141.61	0.19	1.80
	Psychometric					
	Range	Mean	Median	Variance	Skewness	Kurtosis
6. Mass at Zero w/Gap	$0 \le x \le 16$	1.85	0	14.44	1.65	3.98
7. Extreme Asymmetric – Decay	$10 \le x \le 30$	13.67	11.00	33.06	1.64	4.52
8. Extreme Bimodal	$0 \le x \le 5$	2.97	4.00	2.86	-0.80	1.30

Figure 1: Appearance of Five Achievement and Three Psychometric Distributions (Sawilowsky & Fahoome, 2003)

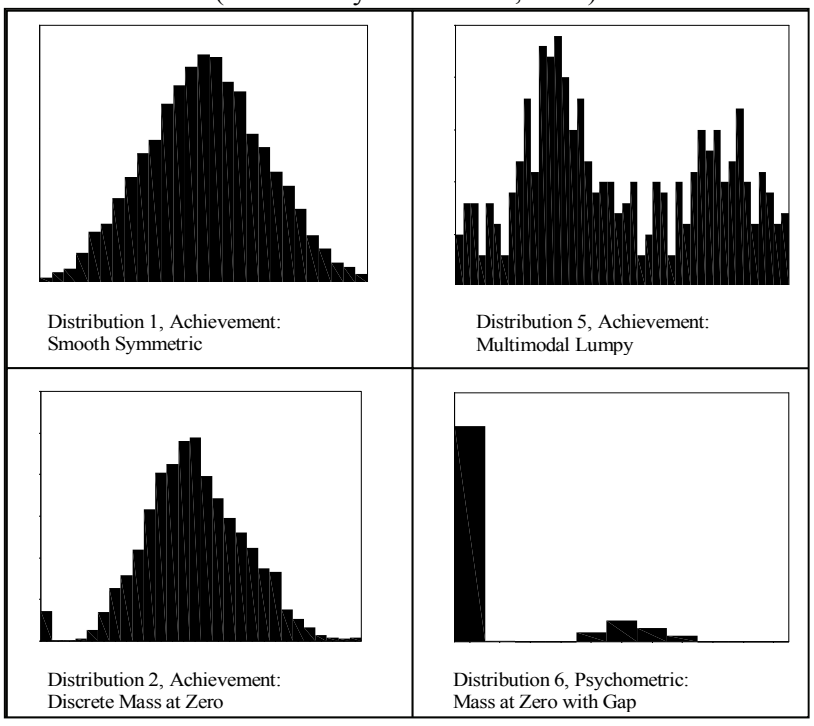


Figure 1 (Continued): Appearance of Five Achievement and Three Psychometric Distributions (Sawilowsky & Fahoome, 2003)

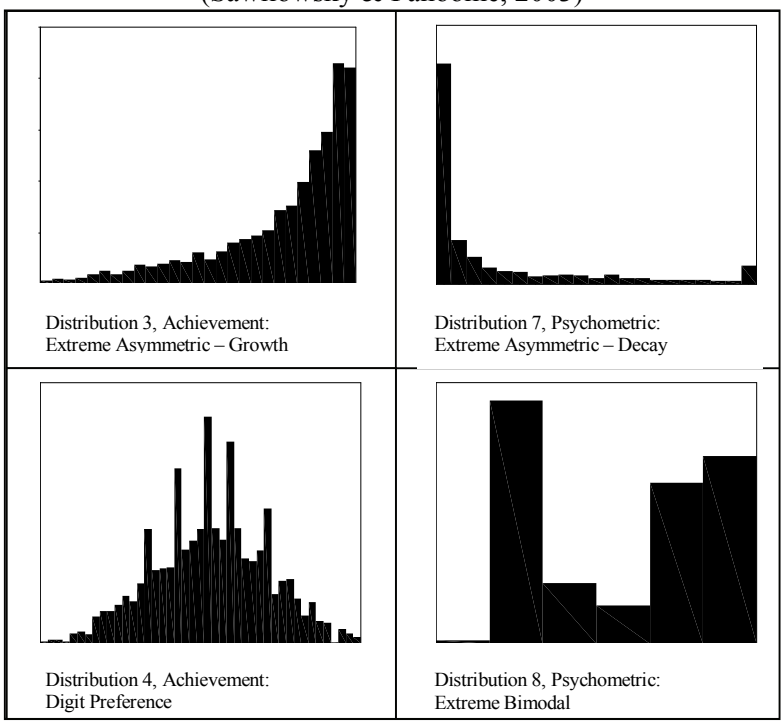


Table 3: Deviation from Target, Summarized by Moment, Sample Size and Distribution

	Moment							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Mean	4	0.000	1	0.000	2	0.000	3	0.000
Standard Dev	2	1.142	3	1.186	4	1.603	1	1.119
Skewness	2	0.192	2	0.192	1	0.191	2	0.192
Kurtosis	2	0.947	3	0.941	4	0.952	1	0.930
	Sample Size							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
$5 \leq 50$	2	0.609	3	0.628	4	0.769	1	0.603
$100 \leq 1000$	2	0.435	3	0.423	4	0.447	1	0.416
	Distribution							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Smooth Sym	2	0.393	3	0.411	4	0.531	1	0.391
Discr Mass Zero	2	0.404	3	0.421	4	0.539	1	0.403
Asym – Growth	2	0.453	3	0.470	4	0.583	1	0.452
Digit Preference	2	0.390	3	0.408	4	0.527	1	0.370
MM Lumpy	2	0.412	3	0.396	4	0.510	1	0.376
MZ w/Gap	2	1.129	3	1.126	4	1.204	1	1.113
Asym – Decay	2	0.726	3	0.739	4	0.835	1	0.725
Extr Bimodal	2	0.655	3	0.669	4	0.765	1	0.654

Proximity to target includes deviation values, at the top of the Tables 4 and 5, and RMS values, at the bottom. RMS is an important second measure of accuracy because it indicates how consistently the methods perform. By standardizing the linear distance of each observed moment from its target, RMS denotes within-method magnitude of deviation. Respectively, the two accuracy measures, deviation value and magnitude of deviation, describe each method's average distance from the target value and how much its performance varies over the course of 10,000 random events.

Predictive Patterns of the Deviation Range

Figure 2 plots the range of deviation values for each distribution against a power curve among small samples. Curve fitting is only possible for the deviation range on the second and fourth moments, standard deviation and kurtosis. The first and third moments, mean and skewness, either contain zeros, which make transformations impossible, or lack sufficient variability to make curve fitting worthwhile.

To evaluate trends at larger sample sizes, the small-sample regression models are fitted a second time with the addition of four sample sizes: n = 100, n = 200, n = 500, and n = 1000. To whatever extent predictive patterns are established when $n \leq 50$, those regression slopes either improve in fit or continue to hold when sample sizes increase. Figure 3 shows that inclusion of larger sample sizes causes the Smooth Symmetric power curve to remain intact and the Digit Preference power curve to improve in fit.

Table 4: Proximity to Target Standard Deviation for Achievement and Psychometric Distributions

Deviation Value

	Blom		Tukey		Van der Waerden		Rankit	
	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$
Achievement	0.736	0.205	0.824	0.122	1.413	0.231	0.735	0.089
Psychometric	2.263	1.382	2.332	1.390	2.802	1.455	2.260	1.374

Magnitude of Deviation (RMS)

	Blom		Tukey		Van der Waerden		Rankit	
	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$
Achievement	0.018	0.001	0.017	0.001	0.017	0.001	0.009	0.001
Psychometric	0.542	0.096	0.540	0.096	0.536	0.096	0.497	0.088

Table 5: Proximity to Target Standard Deviation for Small and Large Samples

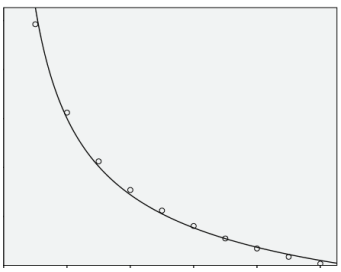
Deviation Value

	Blom		Tukey		Van der Waerden		Rankit	
	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000
Smooth Sym	0.720	0.077	0.808	0.089	1.401	0.202	0.719	0.047
Discr MZ	0.736	0.082	0.823	0.094	1.414	0.208	0.734	0.073
Asym – Gro	0.829	0.247	0.914	0.260	1.489	0.356	0.827	0.237
Digit Pref	0.702	0.072	0.790	0.084	1.385	0.195	0.700	0.043
MM Lumpy	0.696	0.547	0.785	0.085	1.378	0.196	0.695	0.044
MZ w/Gap	3.651	2.804	3.711	2.815	4.117	2.896	3.647	2.795
Asym – Dec	1.668	0.420	1.743	0.425	2.244	0.458	1.666	0.417
Extr Bimod	1.469	0.921	1.543	0.931	2.045	1.011	1.467	0.912

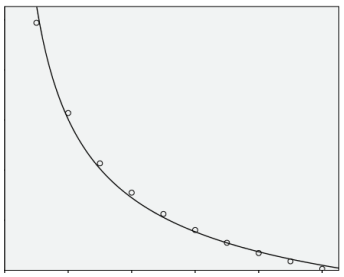
Magnitude of Deviation (RMS)

	Blom		Tukey		Van der Waerden		Rankit	
	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000
Smooth Sym	0.003	0.000	0.003	0.000	0.003	0.000	0.003	0.000
Discr MZ	0.015	0.000	0.015	0.000	0.014	0.000	0.003	0.000
Asym – Gro	0.043	0.003	0.042	0.003	0.042	0.003	0.035	0.003
Digit Pref	0.013	0.000	0.014	0.000	0.013	0.000	0.003	0.000
MM Lumpy	0.013	0.000	0.013	0.000	0.013	0.000	0.002	0.000
MZ w/Gap	1.081	0.225	1.077	0.225	1.069	0.225	0.993	0.226
Asym – Dec	0.310	0.031	0.309	0.031	0.307	0.031	0.290	0.031
Extr Bimod	0.236	0.031	0.235	0.031	0.232	0.031	0.208	0.007

Figure 2: Power Curves Fitted to the Deviation Range of the Standard Deviation at 10 Small Sample Sizes

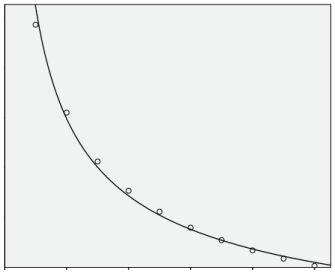


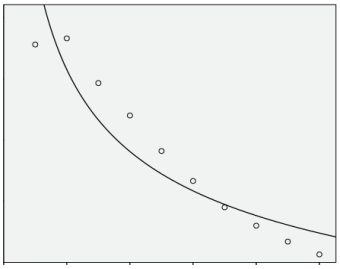
Smooth Symmetric

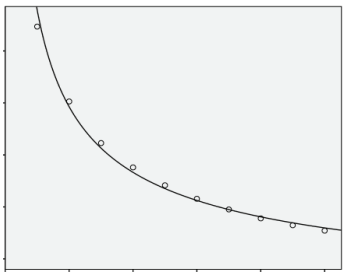


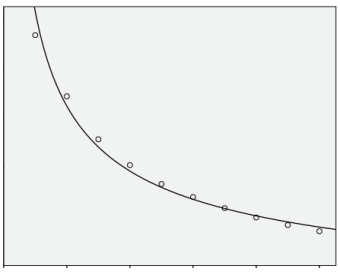
Multimodal Lumpy

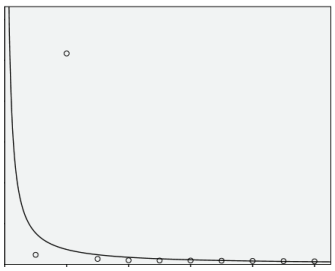
Figure 2 (Continued): Power Curves Fitted to the Deviation Range of the Standard Deviation at 10 Small Sample Sizes


Discrete Mass at Zero


Mass at Zero with Gap


Extreme Asymmetric Growth


Extreme Asymmetric Decay


Digit Preference

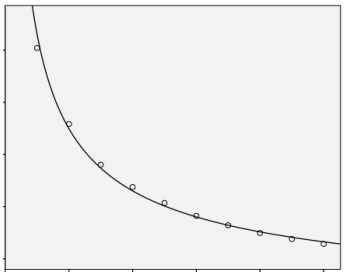
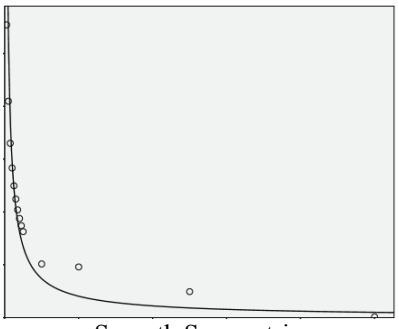
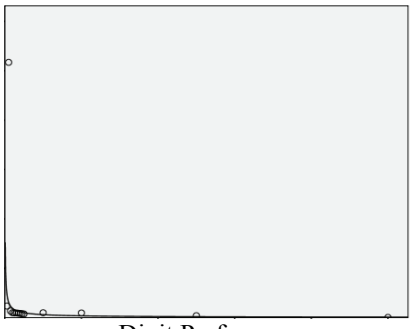

Extreme Bimodal

Figure 3: Power Curves Fitted to the Deviation Range of the Standard Deviation with Inclusion of Four Large Sample Sizes


Smooth Symmetric


Digit Preference

Conclusion

Table 3 shows that Rankit outperforms the other methods across moments at small and large sample sizes and with all eight distributions. Blom and Tukey consistently place second and third, and Van der Waerden performs the worst.

Mean, Skewness, and Kurtosis

All four rank-based normalization methods attain the target value of 50 for the mean. Differences appear in the numerical results only after the third decimal place, and are therefore meaningless in terms of practical application. These differences are reflected in the deviation ranks in Table 3. The four methods' average deviation from the target skewness value of the normal distribution is 0.192 (Table 3). Normalization methods should not be selected on the basis of their deviation from target skewness values because the deviation quantities are small and the differences between them are negligible.

Deviation values for kurtosis show greater deviation from target than those for skewness but less than those for standard deviation. The average deviation value for kurtosis across all sample sizes and distributions is 0.943 (Table 3). Moderate flatness or peakedness might reflect something about the test instrument or the population, but it is not clear how kurtosis could affect decisions made about test scores.

Standard Deviation: Deviation from Target Standard Deviation.

None of the Normalization methods attains the target standard deviation on either accuracy measure. Rankit is the most accurate method, averaging a distance of 1.119 from the target T score standard deviation of 10 (Table 3). This means that the practitioner who uses Rankit to normalize test scores without reference to sample size or distribution can expect to obtain an estimated standard deviation between 8.881 and 11.119. If Z = 2, the T score would fall between 67.762 or 72.238, for a range of 4.476. Adding in the test instrument's standard error compounds the problem. An instrument with a standard error of three (± 3) would expand the true score range by six points, to 10.476. Rounding to the nearest whole number, this means that the test-taker's standardized test score falls somewhere between 65 and 75. Even a standard error half this size would lead to a true score range of 7.476. Thus, a standard deviation that is off target by 1.119 would combine with a standard error of ± 1.5 to increase the true score range by 249%, from a theorized range of three to an actual range of seven and a half. As the standard error increases, the estimated difference between the theorized and actual score range diminishes. At a standard error of three, Rankit produces a standard deviation that causes the true score range to be 175% greater than the presumed score range.

Van der Waerden is the least accurate method, averaging a distance of 1.603 from the target *T* score standard deviation (Table 3). Using Van der Waerden to normalize a test score ($Z = 2$) without reference to sample size or distribution produces a rounded true score range of 64 to 76 at a standard error of ± 3. At a standard error of ± 1.5, the test-taker's *T* score would fall between 65 and 75, the same range that Rankit produced at twice the standard error. Van der Waerden's inaccuracy on the standard deviation causes the true score range to increase over the expected score range by 207% at a standard error of ± 3 and 314% at a standard error of ± 1.5.

As with Rankit, smaller standard errors correspond with greater relative inaccuracy of the true score range. The more reliable a test instrument is, the less precise are the *T* scores, regardless of the normalization method used. This is illustrated in Table 6, which presents the percentage increase to the true score range based on each method's overall distance from the standard deviation across all sample sizes and distributions.

The inaccuracy of the rank-based normalization methods on the standard deviation becomes more pronounced in the context of sample size and distribution type (Table 4). All four methods are more accurate among large samples and achievement distributions and less accurate among small samples and psychometric distributions. Rankit's worst average deviation value, among psychometric distributions at small sample sizes, is 25 times higher than its best among achievement distributions at large sample sizes.

Table 6: Increase of True Score Range over Expected Score Range by Standard Error

Standard Error	% Increase			
	Rankit	Blom	Tukey	Van der Waerden
± 0.5	548%	557%	574%	741%
± 1.0	324%	328%	337%	421%
± 1.5	249%	252%	258%	314%
± 2.0	212%	214%	219%	260%
± 2.5	190%	191%	195%	228%
± 3.0	175%	176%	179%	207%

Van der Waerden's worst deviation value — again, among psychometric distributions at small sample sizes — is 12 times higher than its best. Normalization performance is so heavily influenced by sample size and distribution type that Van der Waerden, which is the worst overall performer, produces much more accurate standard deviations under the best sample size and distributional conditions than Rankit does under the worst distributional conditions. Under these circumstances, Rankit's worst deviation value is 10 times higher than Van der Waerden's best deviation value.

Table 5 illustrates this phenomenon even more starkly. The overall best method, Rankit, has its least accurate deviation value, 3.647, among small samples of the psychometric distribution, Mass at Zero with Gap. Van der Waerden attains its most accurate deviation value, 0.195, among large samples of the Digit Preference achievement distribution. The best method's worst deviation value on any distribution is 19 times higher than the worst method's best value. This pattern holds independently for sample size and distribution. Van der Waerden's best deviation values are superior to Rankit's worst among small and large samples. Sample size exerts a strong enough influence to reverse the standing of the best- and worst-performing methods on every distribution. All four methods perform best with Digit Preference and Multimodal Lumpy and worst with Mass at Zero with Gap.

Separately, the influence of sample size and distribution can make the worst normalization method outperform the best one. Together, their influence can distort the standard deviation enough to render the *T* score distribution, and the test results, meaningless. In the best case scenario, Rankit would be used among large samples of the Digit Preference distribution, where it is off target by 0.043 (Table 5). With a *Z* score of 2 and a standard error of ± 2, this leads to a true score range of 4.172, only 4% greater than the expected score range. In the worst case scenario, Van der Waerden could be used among small samples of the Mass at Zero with Gap distribution, where it is off target by 4.117. With the same *Z* score and standard error, this combination produces a true score range of 20.468, or 512% greater than the expected score range. Clearly, a true score range of 20 is psychometrically unacceptable.Telling a parent that her child scored somewhere between a 60 and an 80 is equally pointless.

Magnitude of Deviation on the Standard Deviation

Returning to the second accuracy measure, magnitude of deviation, Table 4 shows how consistently the methods perform on the standard deviation.[1] Among achievement distributions, they exhibit virtually no variability with large samples (RMS = 0.001) and slight variability with small samples (average RMS = 0.015). Among psychometric distributions, the pattern is the same but the magnitude of deviation is greater for both large and small samples (average RMS = 0.094 and 0.529, respectively). As expected, small samples and psychometric distributions aggravate the instability of each method's performance and exacerbate the differences between them. Average magnitude of deviation for small samples is nearly six times greater than larger samples. Average magnitude of deviation for psychometric distributions is 39 times greater than achievement distributions. Table 5 provides RMS values for all eight distributions. It is notable that Extreme Asymmetric – Growth, which is highly skewed, presents the highest RMS value among achievement distributions, although it is still lower than the psychometric distributions.

The Blom, Tukey, Van der Waerden, and Rankit approximations display considerable inaccuracy on the standard deviation, which has practical implications for test scoring and interpretation. Overestimation or underestimation of the standard deviation can bias comparisons of test-takers and tests. Therefore, practitioners should consider both sample size and distribution when selecting a normalizing procedure.

Small samples and skewed distributions aggravate the inaccuracy of all ranking methods, and these conditions are common in achievement and psychometric test data. However, substantial differences between methods are found among large samples and relatively symmetrical distributions as well. Therefore, scores from large samples should be plotted to observe population variance, in addition to propensity scores, tail weight, modality, and symmetry. Practitioners including analysts, educators, and administrators should also be advised that most test scores are less accurate than they appear. Caution should be exercised when making decisions based on standardized test performance.

This experiment demonstrates that Rankit is the most accurate method on the standard deviation when sample size and distribution are not taken into account; it is the most accurate method among both small and large samples; and it is the most accurate method among both achievement and psychometric distributions. Van der Waerden's approximation consistently performs the worst across sample sizes and distributions. In most cases, Blom's method comes in second place and Tukey's, third.

It would be useful to perform a more exhaustive empirical study of these ranking methods to better describe their patterns. It would also be of theoretical value to analyze the mathematical properties of their differences. More research can be done in both theoretical and applied domains. However, these results identify clear patterns that should guide the normalization of test scores in the social and behavioral sciences.

Note

[1]Curiously, the worst RMS values belong to Blom (Table 4), yet Blom achieves the second place deviation value on three out of four moments, among small and large samples and all eight distributions (Table 3). This suggests that Blom's approximation may achieve some technical precision at the expense of stability.

References

Aiken, L. R. (1987). Formulas for equating ratings on different scales. *Educational and Psychological Measurement*, *47*(1), 51-54.

Allport, F. M. (1934). The J-curve hypothesis of conforming behavior. *Journal of Social Psychology*, *5*, 141-183.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location survey and advances*. Princeton, NJ: Princeton University Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Bliss, C. I., Greenwood, M. L., & White, E. S. (1956). A Rankit analysis of paired comparisons for measuring the effect of sprays on flavor. *Biometrics*, *12*(4), 381-403. Retrieved March 26, 2007 from JSTOR database.

Blom, G. (1954). Transformation of the binomial, negative binomial, Poisson and χ^2 distributions. *Biometrika*, *41*(3/4), 302-316.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. NY: John Wiley & Sons.

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

Chang, S. W. (2006). Methods in scaling the basic competence test. *Educational and Psychological Measurement*, *66*, 907-929

Conover, W. J. (1980). *Practical nonparametric statistics*. NY: John Wiley & Sons.

Cronbach, L. J. (1976). *Essentials of psychological testing* (*3rd Ed.*). NY: Harper & Row.

Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale: Lawrence Erlbaum Associates.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675-701.

Gosset, W. S. ("Student") (1908). The probable error of a mean. *Biometrika*, *6*(1), 1-25.

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika*, *48*(1/2), 151-165. Retrieved August 3, 2007 from JSTOR database.

Horst, P. (1931). Obtaining comparable scores from distributions of dissimilar shape. *Journal of the American Statistical Association*, *26*(176), 455-460. Retrieved August 23, 2007 from JSTOR database.

Ipsen, J., & Jerne, N. (1944). Graphical evaluation of the distribution of small experimental series. *Acta Pathologica, Microbiologica et Immunologica Scandinavica*, *21*, 343-361.

Kline, P. (2000). *Handbook of psychological testing* (*2nd Ed.)*. London: Routledge.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (*2nd Ed.*). NY: Springer Science+Business Media.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden-Day.

Lester, P. E., & Bishop, L. K. (2000). *Handbook of tests and measurement in education and the social sciences* (*2nd Ed.*). Lanham, MD: Scarecrow Press.

McCall, W. A. (1939). *Measurement*. NY: MacMillan.

Mehrens, W. A., & Lehmann, I. J. (1980). *Standardized tests in education* (*3rd Ed.*). NY: Holt, Rinehart and Winston.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

Micceri, T. (1990). Proportions, pitfalls and pendulums. *Educational and Psychological Measurement*, *50*(4), 769-74.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, *3*(1), 55-67.

Nunnally, J. C. (1978). *Psychometric theory*. NY: McGraw-Hill.

Osborne, J. W. (2002). Normalizing data transformations. *ERIC Digest*, ED470204. Available online: www.eric.ed.gov

Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Societ*y, *Series A*, *186*, 343-414.

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

The Psychological Corporation. (1955). Methods of expressing test scores. *Test Service Bulletin*, *48*, 7-10.

Sawilowsky, S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika*, *55*: 729.

Sawilowsky, 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*(2), 352-360.

Sawilowsky, S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Oak Park: JMASM.

Solomon, S. R. (2008). *A comparison of ranking methods for normalizing scores*. Ph.D. dissertation, Wayne State University, United States - Michigan. Retrieved February 27, 2009, from Dissertations & Theses @ Wayne State University database. (Publication No. AAT 3303509).

SPSS (2006). *Statistical Package for the Social Sciences (SPSS)* 15.0 for Windows. Author.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*(6), 1055-1098.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departures from normality. *Communications in Statistics*, *A11*, 2485-2511.

Tapia, R. A., & Thompson, J. R. (1978). *Nonparametric probability density estimation*. Baltimore: Johns Hopkins University Press.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston, MA: Houghton Mifflin.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1-67. Retrieved August 3, 2007 from JSTOR database.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Indian Journal of Statistics*, *25*, 331-351.

Van der Waerden, B. L. (1952/1953a). Order tests for the two-sample problem and their power. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 55 (*Indagationes Mathematical 14*), 453-458, & 56 (*Indagationes Mathematicae 15*), 303-316.

Van der Waerden, B. L. (1953b). Testing a distribution function. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A),*56* (*Indagationes Mathematicae 15*), 201-207.

Walker, H. M., & Lev, J. (1969). *Elementary statistical methods* (*3rd Ed.*). NY: Holt, Rinehart and Winston.

Wilson, E. B., & Hilferty, M. M. (1929). Note on C. S. Peirce's experimental discussion of the law of errors. *Proceedings of the National Academy of Science*, *15*, 120-125.

Wimberley, R. C. (1975). A program for the T-score normal standardizing transformation. *Educational and Psychological Measurement*, *35*, 693-695.