DOCUMENT RESUME

ED 363 637 TM 020 657

AUTHOR Thompson, Bruce
TITLE GRE Percentile Ranks Cannot Be Added or Averaged: A Position Paper Exploring the Scaling Characteristics of Percentile Ranks, and the Ethical and Legal Culpabilities Created by Adding Percentile Ranks in Making "High-Stakes" Admission Decisions.
PUB DATE Nov 93
NOTE 40p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 12, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Admission; *College Entrance Examinations; *Decision Making; Ethics; Heuristics; Higher Education; Legal Problems; *Legal Responsibility; Measurement Techniques; *Scaling; Scores; Testing Programs; *Test Use
IDENTIFIERS Graduate Record Examinations; *High Stakes Tests; Percentile Ranking; *Percentile Ranks

ABSTRACT

The nature of percentile ranks scores is explored using concrete heuristic examples. It is explained why arithmetic operations require measurement on equal-interval scales, and that percentile ranks are not measured on equal-interval scales and therefore may not be added or averaged. The consequences of inappropriately adding percentile ranks are explored from the perspectives of various textbook authors. The discussion is couched in the context of the legal requirements for high-stakes testing. The information presented herein is not new or previously unknown, rather, the distinguishing characteristics of the position paper hopefully include clarity and concreteness of presentation. (Five figures and two tables illustrate the discussion, and an appendix provides additional commentary. Contains 31 references.) (Author)

perctile.wp1 8/19/93

GRE PERCENTILE RANKS CANNOT BE ADDED OR AVERAGED: A POSITION PAPER EXPLORING THE SCALING CHARACTERISTICS OF PERCENTILE RANKS, AND THE ETHICAL AND LEGAL CULPABILITIES CREATED BY ADDING PERCENTILE RANKS IN MAKING "HIGH-STAKES" ADMISSION DECISIONS

Bruce Thompson

Texas A&M University
and
Baylor College of Medicine

Bruce Thompson is a Fellow of the American Psychological Association's Evaluation, Measurement and Statistics Division, a Charter Fellow of the American Psychological Society, a former President of the American Counseling Association's Association for Assessment in Counseling division, and the representative of the National Council on Measurement in Education to the Joint Committee on Standards for Program Evaluation. He is an Executive Editor of the Journal of Experimental Education, since 1985 has been the Book Review Editor for Educational and Psychological Measurement, and is a former editor of Measurement and Evaluation in Counseling and Development. He currently sits on three additional journal editorial boards. He has authored or edited five books/monographs, has published more than 100 articles, and has presented 170 papers at refereed professional meetings. He formerly was a university Research Professor at the University of New Orleans in the LSU System, and Adjunct Professor of Biometry at LSU Medical Center. He currently is a Professor in the Department of Educational Psychology at Texas A&M University, a Distinguished Research Fellow, and an Adjunct Professor of Community Medicine at Baylor College of Medicine.

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, November 12, 1993. The helpful comments of Jerome T. Kapes on a previous version of this paper are gratefully acknowledged.

ABSTRACT

The nature of percentile ranks scores is explored using concrete heuristic examples. It is explained why arithmetic operations require measurement on equal-interval scales, and that percentile ranks are not measured on equal-interval scales and therefore may not be added or averaged. The consequences of inappropriately adding percentile ranks are explored from the perspectives of various textbook authors. The discussion is couched in the context of the legal requirements for high-stakes testing. The information presented herein is not new or previously unknown, rather the distinguishing characteristics of the position paper hopefully include clarity and concreteness of presentation.

"In its broadest sense," wrote Stevens (1951, p. 1, emphasis added), "*measurement* is the assignment of numerals to objects or events according to rules." For example, when we assign numerals to objects based on the correspondence of their heights to a stick marked off in units, we are engaged in measurement. Although all measurement results in the assignment of numerals, it is absolutely essential to remember that numerals do not always contain the same amount of *information* or meaning. It is the nature of the rules we employ during measurement that determines how much information a resulting set of numerals will contain.

The present paper is an essay about measurement. It is argued that percentile rank scores are *ordinally* scaled numerals that should not be added. Although some authors use the term "percentile" and "percentile rank" interchangeably, I will use the term "percentile rank" to refer to the scores defined thusly by Mehrens and Lehmann (1991, p. 231, emphasis in original):

> A *percentile rank* gives a person's relative position or the percentage of students' scores falling below his obtained score. For example, let us assume that John has a raw score of 76 on an English test composed of 100 items. If 98 percent of the scores in the distribution fall below a score of 76, the percentile rank of the score of 76 is 98...

To make the discussion concrete, we will presume that a hypothetical university has adopted an admissions policy requiring that GRE percentiles be added together to create a "criterion

score". The admissions policy might invoke an algorithm, for students holding only baccalaureate degrees, in the form:

CRITSCOR = (50 x Undergraduate GPA) + GRE Verbal %tile + GRE Quantitative %tile

Hypothetically, the university might require that a student must have a criterion score of 220 in order to be admitted without "exceptional" further justification.

The remainder of the paper is divided into six major sections. First, why the mathematics of addition require measurement on "equal-interval" scale is explained (readers with more background may skip this section without serious loss of continuity). Second, three measurement levels of scale are explained, including some which will not and one which will allow meaningful addition. Third, the nature of percentile ranks is explored. Fourth, various scholars' views on adding (or performing other arithmetic operations with) percentile ranks are elaborated. Fifth, the legal requirements for "high stakes" testing are explored. Finally, by way of summary, the reasons why a university might adopt and enforce a graduate admissions policy that inappropriately requires the addition of GRE percentile rank scores are briefly considered. Although the discussion is couched in terms of the addition of GRE percentile ranks, the conclusions fit equally well as regards the addition of percentile ranks from all sorts of measures.

Why Addition Requires Equal-Interval Measurement

Long, long ago, people measured the heights of their horses using the lengths of their hands, because of the convenient

availability of our hands. However, it soon came to be recognized that, because different people have hands of different lengths, all horses measured as being x hands tall were not necessarily equally tall. So that the information contained in the numerals would be more useful, it was recognized that it would be advantageous to employ measuring units of a fixed, standard length.

Measuring height with units of varying lengths is somewhat related to creating rulers marked off in non-standardized units. For example, we might wish to create units of height called, "thumbs", analogously to the original "hands" with which the heights of horses were traditionally measured. Ruler A presented in Figure 1 was created by marking the ruler using the thumb width of 10 different people of different ages who, not surprisingly, had somewhat different widths of thumbs. Ruler B was created by marking the ruler in *equal-interval* units representing the approximate average thumb width of adults.

Figure 1

Two Rulers and Their Use in Measuring Sets of Books

```
-----1--2---3----4---5--6---7---8--9--10  Ruler A
xx Book #1 xx                             Sideview of Book #1
yyyyyyyy Book #2 yyyyyyyy                 Sideview of Book #2
xx Book #1 xxxx Book #1 xx                2 Copies of Book #1
----1----2----3----4----5----6----7----8  Ruler B
```

Ruler A does have some utility. For example, as seen in Figure 1, if we measure the thicknesses of two book in these units of "thumbs", starting from the leftmost side of Ruler A, we are

entirely correct in assuming that numerals of 3 and 6, respectively, would indicate that the second book is thicker than the first one. Unfortunately, what we cannot say is that the second book is 3 (6 - 3) "thumbs" thicker than the second book. This is because non-standardized units are not meaningfully additive. As Siegel (1956, p. 19) put it:

> The variables involved must have been measured in *at least* an interval scale, so that it is possible to use the operations of arithmetic (adding..., findings means, etc.) on the scores.

It seems counterintuitive to many persons, even to some educated people with terminal degrees serving on faculty at world-class universities, that some numbers simply cannot be added [or subtracted, since subtraction actually is addition, via the addition of negative numbers]. The problems inherent in adding numerals from nonstandardized measurement can be difficult to see, because of paradigm influences.

As defined by Gage (1963, p. 95), "Paradigms are models, patterns, or schemata. Paradigms are not the theories; they are rather ways of thinking...." But even highly educated scientists usually do not consciously recognize the influence of their paradigms. As Lincoln and Guba (1985, pp. 19-20) note:

> If it is difficult for a fish to understand water because it has spent all its life in it, so it is difficult for scientists... to understand what their basic axioms or assumptions might be and what impact

those axioms and assumptions have upon everyday thinking and lifestyle.

Even though researchers are usually unaware of paradigm influences, paradigms are nevertheless potent influences in that they tell us what we need to think about, and also the things about which we need not think. As Patton (1975, p. 9) suggests,

> Paradigms are normative, they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of a paradigm that constitutes both its strength and its weaknesses--its strength in that it makes action possible; its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm.

Most of us have paradigms about numbers that were unconsciously formulated, typically in the primary grades of elementary school. When we are given several numerals, we are used to presuming that we can add them up. Few of us were ever admonished that we can only add numbers when the numerals represent data derived using an equal interval measurement ruler. In fact, few of us consciously recognize that addition itself does presume equal-interval measurement.

Our two rulers measuring "thumbs" can be used to illustrate that addition does require equal-interval measurement. If we stack two copies of Book #1 on top of each other, and measure the stack, a ruler will have been used legitimately for comparative purposes

only if the resulting measurement honors our perception of reality. Ruler A tells us that Book #1 is 3 "thumbs" tall, while Ruler A indicates that Book #2 is 6 "thumbs" tall. If Ruler A is an equal-interval measurement, then we can add numerals from the measurement protocol and obtain results that correspond with the reality being measured.

The facts that Book #2 is 6 "thumbs" tall according to Ruler A, while Book #1 is 3 thumbs tall, suggest that adding a second copy of Book #1 on top of a first copy of Book #1 should yield a stack that is exactly the same height as one copy of Book #2, since 3 + 3 = 6. Unfortunately, as we can see from Figure 1, two copies of Book #1 create a stack slightly larger than Book #2 alone. The failure of our measurement to yield results that correspond to perceived reality alerts us to the fact that our addition is not sensible, and since the problem is not in the process of addition itself (we have correctly followed the rules of addition), the problem must then be with the numerals from our measurement.

Of course, we can always add up numbers/numerals, but even if we correctly implement the mathematical process, our answer will only make sense or be *meaningful* if our numbers have sufficient information to make the process of addition itself sensible for our numerals. The fact that we can implement a mathematical process while correctly following its procedural rules <u>does not</u> imply that the resulting answer will necessarily be sensible. Even fewer people see this now than used to, because computers will promptly mathematically manipulate any numbers we give them, and persons

associating omniscience with computers presume that computers would certainly be helpful enough to know vicariously how much information is in our numbers and would doubtless kindly warn us whenever our numerals cannot *meaningfully* be added/subtracted.

Of course, physical scientists may most tend to assume that all numerals are intervally scaled, because most of the phenomenon that they investigate are indeed measured with "rulers" that are demarcated with equal intervals (e.g., centimeters, grams). But things get more complicated when psychological constructs, such as academic achievement or aptitude, are measured, as with the Graduate Record Exam (GRE).

The Levels of Scale

Stevens (1946, 1951, 1968) proposed the concept of *levels of scale* to help us describe how much information is in given sets of numerals, and therefore what mathematical operations we may sensibly perform on our numerals. Three of the levels of scale will be described here.

The first level of scale is the *nominal* (or what some call the "categorical") level of scale. Variables measured at this level of scale contain the information (1) that people in the same category of the variable (e.g., male) are considered identical with respect to the variable being measured (e.g., gender), and (2) that people in different categories of the variable are different with respect to the variable being measured (e.g., males are different than females with respect to the variable, gender).

When we measure gender, we may assign any numerals we wish, as

long as we do not distort the two pieces of information available to us. Thus, we may assign all males a -3 and all females +1, or vice versa, or we may assign +2 and .967, respectively. However, whatever number we pick for males, for example, we must assign all males that number. Furthermore, whatever number we pick to assign to all the males, though we have infinitely other legitimate choices, we may not assign that particular number to the females.

Say that our sample of subjects included William, Dean, Robert, Dan and Sallie. Say that we choose to assign +1 to males and +2 to females. We could apply the algorithm for the mean, by adding up the scores from our measurement, and then dividing by the number of people. Though the mathematics are possible, the resulting answer makes no sense. It is not meaningful to say that the mean amount of gender in our sample was 1.2.

We can see the senselessness of the mean for nominal data when we recognize that we can change the meaningless result by changing our arbitrary measurement system. We could have just as reasonably assigned +.5 to females and +100 to males. The mean for the same data becomes 80.1. But the mean itself has no meaning when the measurement scale is not delineated by equal intervals. Certainly we could not correctly conclude that the same 5 people now have more of the variable gender because the mean score on gender is now larger. Instead, we would not compute a mean for nominal data, because we cannot sensibly add numbers on a variable measured at the nominal level of scale.

We might assign 0 to females and +1 to males, and use the

algorithm for the mean to obtain a result of .8. The mean of a dichotomously scaled nominal variable coded 0 and +1 does correctly advise us of the proportion of people in the group scored +1, but we should not conclude even here that the mean itself has become sensible. Rather, we should recognize that counting the number of people in a given group is possible for this level of scale, that we can sensibly determine the proportion of people in a given group, and that for dichotomous variables scored in this manner we can coincidentally determine the proportion of people in the group scored +1 if we employ the algorithm for the mean. But this result derives its meaning as a proportion, and not as a mean per se.

Data measured at the *ordinal* (or the "ranked") level of scale contain the information (1) that people in the same category of the variable are considered identical with respect to the variable being measured, (2) that people in different categories of the variable are different with respect to the variable being measured, and (3) that the people we have measured have a certain meaningful ordering with respect to the variable being measured. For example, presume that William, Dean, Robert, Dan and Sallie went to the same high school, and that they graduated valedictorian, saluditorian, 3rd, 4th and 5th in class rank, respectively.

For this variable we may again assign any numerals we wish, as long as our measurement model does not distort the information we believe is available, given our model of reality. The measurement model and the model of reality must always correspond. For example, we can assign the subjects the numerals 1, 2, 3, 4, and 5,

respectively. But we could also order the categories from the opposite end of ranking, and assign the numerals 5, 4, 3, 2, and 1; this would only mean that a larger number would now reflect a higher high school grade point average.

However, since we have no information available about how far apart our subjects were with respect to their GPA's, we could just as legitimately assign the numerals 1, 1.5, 8, 9, and 700, respectively. Of course, if William and Dean had had exactly the same average, we would have assigned them both the same numeral, as we would have done with nominally-scaled variables as well.

Since our class rank numerals are not based on numerals assigned using a "ruler" marked off with equally-spaced intervals, if all we are given is class rank information, we cannot add our numerals or compute a meaningful mean with our numerals. We could do the computations, but the result would not be sensible.

The third level of scale is the *interval* (or continuous) level of scale. Here we know that (1) that people in the same category of the variable are considered identical with respect to the variable being measured, (2) that people in different categories of the variable are different with respect to the variable being measured, (3) that the people we have measured have a certain meaningful ordering with respect to the variable being measured, and (4) that there are meaningful distances of the categories from each other that we have ascertained with some "ruler" that has been demarcated with equally-spaced units. Of course, the last condition does not mean that all the people are equally spaced (e.g., 58" tall, 59"

tall, and 60" tall), but only says that every unit on our measurement "ruler" is equally spaced (e.g., every inch on a yardstick represents exactly the same distance).

Again, we can measure our variable in any way we wish, as long as we honor all the information available, and do not distort the relationship between our presumptive model of reality and our measurement model or protocol. We can measure in inches, in centimeters, or in "thumbs" using Ruler B in Figure 1, as we chose. But if William and Dean are exactly equally tall, among other things we must assign them the same numeral on the variable of height. And if Dan are Sallie are equally tall, and if Dan stands on top of Sallie's head and together they are exactly as tall as William, among other things we must assign William a numeral that is the sum of the numerals we assigned to Dan and to Sallie.

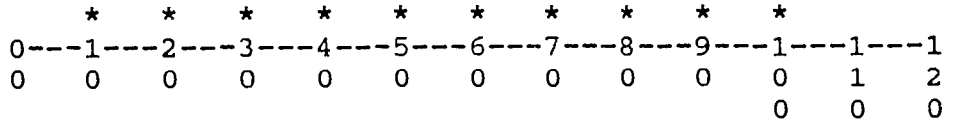
The Ordinal Nature of Percentile Ranks

Percentile Ranks for Flat ("*Rectangular*") Score Distributions

The calculation of percentile ranks will first be illustrated with respect to what statisticians call a "rectangular" score distribution. This is a symmetrical score distribution with an equal number of persons in each score interval. Figure 2 presents the histogram for a heuristic example involving the number of right answers of 10 people on a 120 item spelling test.

Figure 2

Histogram of Scores of 10 People on a 120-Item Spelling Test



Note. Each asterisk represents one person, as follows: William, 100 rights answers; Dean, 90; Robert, 80; Dan, 70; Sallie, 60; Jane, 50; Jon, 40; Mike, 30; Larry, 20; Bruce, 10.

The percentile rank for William's score of 100 is 95, because 90% of the 10 people scored lower than his score, and because conventionally half the number of people in a given score interval is then added to that percentage, e.g., 90 + (.5 x 1) = 95. The percentile ranks for the remaining test takers are, respectively: Dean, 85; Robert, 75; Dan, 65; Sallie, 55; Jane, 45; Jon, 35; Mike, 25; Larry, 15; Bruce, 5.

For *any* distribution, percentile ranks and raw or related scale scores (e.g., GRE or IQ scales) *order* people in exactly the same order. Thus, for example, the person with the highest GRE will always have the highest percentile score, and so forth.

Futhermore, for a rectangular score distribution, and only for a rectangular score distribution with scores that are equi-distant, percentile ranks and raw or related scores contain the same information, but are merely *scaled differently* (i.e., may have different means and/or standard deviations). Thus, the Pearson correlation coefficient between rectangularly-distributed raw or scale scores and percentile ranks will always be +1.0.

Percentile Ranks for *Non-Rectangular* Score Distributions

Although the conversion of non-rectangularly distributed (e.g., normally-distributed GRE) scores into percentile ranks does not change the ordering of people, the conversion does usually change everything else characteristic of the initial score distribution, including the *shape* of the distribution. Kirk (1984,

p. 221) notes this necessity using the language of the statistician:

> We can see from the figure that the transformation of scores to percentile ranks has altered four characteristics of the distribution: (1) central tendency, (2) dispersion, (3) skewness, and (4) kurtosis. The only characteristic that isn't changed by the transformation [of non-rectangularly distributed scores] is the rank order of scores within the distribution.

The reason why percentile ranks always change the shape of a non-rectangular distribution of original scores is because distributions of percentile ranks are <u>always</u> rectangular, by definition. As Eichelberger (1989, p. 137) notes, "Percentiles [always] form a rectangular distribution." Each centile will always have an equal number of subjects at that given percentile rank, specifically, 1/100th of the people in the score set. Thus, since percentile ranks are always rectangularly distributed, when any non-rectangularly-distributed set of scores is changed into percentile ranks, the shapes of the two distributions will always differ.

Table 1 presents a data set that can be used to illustrate these dynamics. The Table 1 data are normally distributed. The table presents the ID numbers of the 100 subjects, and the number of right answers they got on a verbal ability test. The table alsc presents the <u>Z</u> ($\overline{X}$ = 0, <u>SD</u> = <u>V</u> = 1) and the GRE scale score ($\overline{X}$ =

500, SD = 100) equivalents of these raw scores. Finally, the table presents the percentile ranks for a given set of raw, Z and GRE scores, e.g., in this score set the percentile rank of 8 right answers, a Z score of -2.6, and a GRE score of 240, is 1.

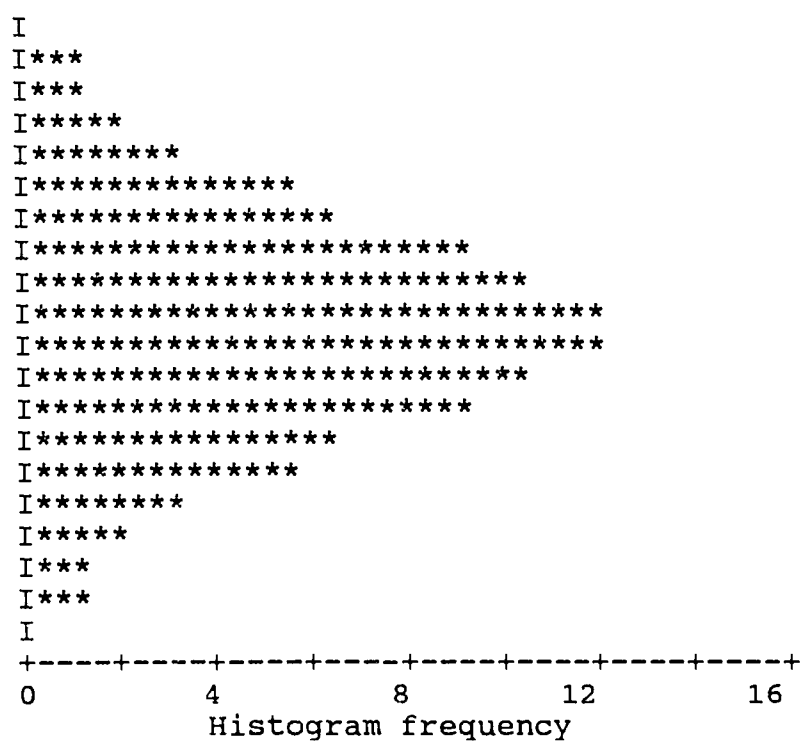
INSERT TABLE 1 ABOUT HERE.

Table 2 presents a conversion table for these data. The example presented in the note to Table 2 also makes dramatically clear that percentile ranks, unlike the GRE scores for which they are derived, are not equal interval scores.

INSERT TABLE 2 ABOUT HERE.

Figure 3 presents a histogram of the GRE scores for these 100 subjects. The figure indicates that the distribution is not rectangular. In fact, this particular set of scores is very nearly normally distributed (Bump, 1991). Therefore, the conversion of these 100 GRE scores into percentile ranks will honor the ordering of the GRE scores, but will rescale the scores and will also change the shape of the score distribution.

Figure 3
Histogram of GRE Scores (n = 100)



Note. One asterisk equals approximately .40 occurrences.

The fact that percentile ranks are always rectangularly distributed means that percentile ranks are *not linear transformations* of non-rectangularly distributed raw or scale scores. Adding (or subtracting) +2 to all scores, or multiplying (or dividing) all scores by -3, are examples of linear transformations. There are several ways to illustrate that the conversion of non-rectangular scores to percentiles ranks is not linear.

First, an index of linear relationship, the Pearson correlation coefficient, can be computed for each possible pair of scores. Scores that are linear transformations of each other will

be perfectly correlated; scores that are not linear transformations will not be perfectly correlated.

Table 3 presents the correlation coefficients for the Table 1 data. The correlation coefficients between pairs of the variables, number of right answers, Z score of number of right answers, and GRE score equivalent of number of right answers, are all perfectly correlated, as expected. The percentile rank equivalents of these scores are less than perfectly correlated with these three scores, because the percentile ranks are not linear transformations of raw or scale scores. Furthermore, the percentile rank scores have exactly the same correlation coefficients (all +.9775) with the three scores, again because the three scores are only linear transformations of each other.

Table 3

Correlation Coefficients for the Table 1 Data

(n=100)

	VRIGHT	VERBZ	GREV	VPER
VRIGHT	--			
VERBZ	**1.0000**	--		
GREV	**1.0000**	**1.0000**	--	
VPER	.9775	.9775	.9775	--

Note. **Bolded** entries involve pairs of variables that are linear transformations of each other.

Second, the non-linear transformation *required* to convert normally-distributed scores into rectangularly-distributed scores can be illustrated graphically. Figure 4 presents a scattergram plot of the 100 GRE scores and the equivalent percentile ranks. The fact that the plotted data do not yield a straight line indicates that one set of scores is not a linear transformation of the other.

INSERT FIGURE 4 ABOUT HERE.

The fact that the correlation coefficient ($\underline{r}$ = +.9775, $\underline{r}^2$ = .9555) between the GRE scores and the associated percentile rank scores is so high might suggest to a naive reader that there are differences in the two sets of scores, but that these differences are minor and that focusing on them might merely represent the traditional statistician's nit-picking.

However, it is very important to note that most of the 100 scores are in the middle of the distribution, by definition, given that the GRE scores are normally distributed. This is reflected in Figure 4 by the large number of 3's and 4's in the middle of the plot. This is the area within the plot that most closely fits a linear pattern, as indicated by the regression line plotted in Figure 4.

The worst fit to the linear regression line plotted in Figure 4 is for the more extreme scores. When one converts normally-distributed scores into rectangularly-distributed scores, *differential (i.e., non-linear) conversion* is applied at different points along the continuum of the nonrectangularly-distributed raw or scale scores. As Thorndike, Cunningham, Thorndike and Hagen (1991, p. 65, emphasis added) explain:

> Thus, percentile units are typically and systematically unequal relative to raw score units. The difference between being first or second in a group of 100 is many times as great as the

> difference between being 50th and 51st. Equal percentile differences do not, in general, represent equal differences in amount of the trait in question. Any interpretation of percentile ranks must take into account the fact that such a scale has been *pulled out* at both ends and *squeezed* in the middle.

Blommers and Forsyth (1977, p. 64, emphasis added) describe the non-equal-interval nature of percentiles ranks in this context, by noting that:

> ...deciles, or for that matter quartiles or [per]centiles, cannot be regarded as units in the usual sense. The actual score distances between them *fluctuate* [at various points along the original score continuum].

Kirk (1984, pp. 221-222, emphasis added) emphasizes the same point, noting that

> the interpretation of a 10-point difference between percentile ranks *depends* on where the difference is on the 0-100 scale.

Ahmann and Glock (1981, p. 221, emphasis added) describe the situation thusly:

> Percentile norms have "*rubber units*"--units of varying sizes. The extent to which the units have been "*rubberized*" depends on the nature of the distribution of the raw scores. If that distribution

is normal or nearly normal, the amount of distortion is large....

Hinkle, Wiersma and Jurs (1979, p. 28, emphasis added) note that

> ...in the center of the distribution, the use of percentile scores tends to *exaggerate* small, nearly nonexistent differences. On the other hand, in the tails of the distribution, the use of percentile scores tends to *underestimate* actual differences.

Asher (1976, p. 91, emphasis added) characterizes these features of percentile ranks as distortions:

> The percentile rank scoring system systematically *distorts* differences between scores near the center of a distribution, making them appear larger, while also *distorting* differences between scores at either extreme by making them appear much smaller than when they are scored in most other measurement systems. In general, when only percentile rank scores are available for a variable, it is good practice to use a conversions table to change them back to raw scores.

Ahmann and Glock (1981, p. 221, emphasis added) offer a similar view:

> Certainly the percentile ranks are *hiding* large differences between raw scores when they occur at either the high or low extremity of the raw-score

> distribution, and also are *enlarging* small differences between raw scores when they occur near the center of the distribution.

Thus, Crocker and Algina (1986, p. 441) note that the nonlinear conversion implicit in conversion to percentile ranks can cause people to misinterpret these scores:

> Most misinterpretations arise when test users fail to recognize that the percentile rank scale is a nonlinear transformation of the raw score scale. Simply put, this means that at different regions on the raw score scale, a gain of 1 point may correspond to gains of different magnitudes on the percentile rank scale.

Figure 5 presents the analogous scattergram for the 26 people below the 25 percentile for the Table 1 data. The fact that there is the least linearity in score conversion at the extremes of the scores distribution is indicated by the less linear pattern present in this plot. Cunningham (1986, p. 68) notes that most of the raw and scale score distributions we encounter induce distortions in conversions to percentile rank scores:

> Because most distributions that we encounter are likely to approximate a normal rather than a rectangular distribution, the intervals between percentiles will be quite different from the intervals between raw scores....

INSERT FIGURE 5 ABOUT HERE.

The fact that there is the least comparability or linearity in conversion at the extremes of the score distribution is troubling in the context of our admissions example, because this is exactly the region of the score distribution at which we are generally working in making decisions about admissions to graduate school for students who are on the decision margins. For example, given our hypothetical admissions policy, a student with Verbal and Quantitative GRE percentile ranks of 20 and 20 would have had to have had an undergraduate GPA of 3.6, and a student with a 3.4 average must have percentile ranks, for example, of 25 and 25, or 2 and 48, or 10 and 40.

Views on the Use of Percentile Ranks in Mathematical Operations

Textbook authors who speak to the issue are unanimous in their view that percentile ranks "are not equal interval scores" (Carey, 1988, pp. 383-384). Gronlund and Linn (1990, p. 349) offer yet another perspective focusing on *why* percentile ranks are not intervally scaled:

> A percentile difference of 10 near the middle of the scale (e.g., 45 to 55) represents a much smaller difference in test performance than the same percentile difference at the ends (e.g., 85 to 95), because a large number of pupils receive scores near the middle, whereas relatively few pupils have extremely high or low scores. Thus, a pupil whose

raw score in near average can surpass another 10 percent of the group by increasing the raw score just a few points. On the other hand, a pupil with a relatively high score will need to increase the raw score by a large number of points in order to surpass another 10 percent, simply because there are so few pupils at that level.

Because measurement scholars uniformly posit that percentile ranks are not intervally scaled, they are also uniform in their view that percentile ranks should <u>not</u> be added, averaged or otherwise mathematically manipulated. For example, Mehrens and Lehmann (1991, p. 232) note that:

> Percentile ranks have a disadvantage in that the size of the percentile units is not constant in terms of raw-score units[2]. [[2]Except in the unusual case where the raw-score distribution is rectangular.] For example, if the distribution is normal, the raw-score difference between the 90th and the 99th percentiles is much greater than the raw-score difference between the 50th and the 59th percentiles.... Of course, the ordinal nature of the percentile rank units means that one cannot treat them further statistically.

These considerations bear directly upon the merit of graduate admissions policies requiring that GRE percentile ranks be added, as in an algorithm in the form:

CRITSCOR = (50 x Undergraduate GPA) + GRE Verbal %tile +
GRE Quantitative %tile

Hinkle, Wiersma and Jurs (1979, p. 28, emphasis added), with respect to adding percentile ranks, note that:

> Percentiles are unequal units of measurement and hence should not be arithmetically manipulated. Thus, there is *no justification for summing* or combining them, averaging them or manipulating them in ways we would manipulate scores that are equally spaced on a scale.

Similarly, Hills (1981, p. 239, emphasis added) notes that:

> ...[I]t is *not sound to add* percentiles or to average them. If you take two scores and obtain their percentiles [percentile ranks] and then average the percentiles [percentile ranks], the result will not be the same as that found by averaging the scores first and then obtaining the percentile [rank] for the average.

Thorndike, Cunningham, Thorndike and Hagen (1991, p. 65, emphasis added) concur:

> One of the consequences of this inequality of units in percentile scale is that percentiles cannot be treated with many of the procedures of mathematics. For example, we *cannot add two percentile ranks* together and get a meaningful result. The sum or average of the percentiles of two raw scores will

not yield the same result as determining the percentile rank of the sum or average of the two scores directly.

Appendix A presents a supplementary cascade of quotations from scholars regarding the entirely dubious innovation of adding percentile ranks.

Percentile ranks present helpful information about scores, i.e., what percentage of test takers in a given normative group obtained a lower score on the test. But the utility of percentile ranks is limited. The question might then be posed as to why percentile ranks are used so often. One answer is that "percentile [rank] scores are useful because they are easily understood by the *layperson*" (Borg & Gall, 1989, p. 248, emphasis added). Wike (1985, p. 72, emphasis added) concurs, noting that:

> Percentiles [percentile ranks] are readily understood by real people like *bartenders, undertakers, used car salespersons, and bankers*. That is an advantage.

However, hopefully laypersons and bartenders never add percentile ranks, and nor would highly trained academics ever do so.

Legal Requirements for "High Stakes" Testing

Measurement specialists have come to call tests that have serious consequences for individuals or institutions, "high stakes" tests. Examples are licensure exams, or statewide exams that high school students must pass to graduate with a diploma. It could be argued that graduate school admissions decisions have "high stakes"

for applicants. And as Mehrens and Popham (1992, p. 265) have observed:

> When tests are used for high-stakes decisions, there is a strong possibility that individuals for whom an unfavorable decision is made will bring a legal suit against the developer and/or user of the test.

The notion of high stakes testing evolves out of the common law traditions of remedy. As a general principle of law, the remedies that are fashioned on behalf of injured plaintiffs are developed to put the plaintiffs as nearly as possible back in the conditions they would have enjoyed had they not been unlawfully injured. Of course, in certain cases punitive damages may also be added on top of these nonpunitive damages.

Various courts have interpreted or created law related to test use. For example, in *Groves et al. v. Alabama State Board of Education et al.* the court held that the ACT may not be used as an absolute criterion for admission into colleges of education. But in *United States v. LULAC* (1986) the court upheld the use of a standardized test in making college admissions decision, because here a relationship was shown between what was tested and what was required in the training program.

Courts have also generally held that judgment may be exercised in establishing criterion cut-off scores (e.g., *Tyler v. Vickery*, 1975). But courts have differed over what legal standards must be met in establishing criterion cut-off scores, and at times have found that cut scores are too high, even when evidence regarding

the validity of the cut scores is available (e.g., *Richardson v. Lamar County Board of Education*, 1989). It appears that at a minimum the test user must be able to offer some reasons for the selection of cut scores, although this judgment may rest only on "analyzing the test results to locate a logical 'break-point' in the distribution of scores" (Byham, 1983, p. 107).

One very important source of law relevant to test use is the U.S. Constitution's 14th Amendment *due process* requirements. Basically, tests must be used in ways that are substantively and procedurally fair. It might be expected that these requirements would pose serious dangers to those about the business of adding percentile ranks when making admissions decisions.

In general, policy must not lead to arbitrary and capricious judgments that negatively impact people. Policy requiring the adding of percentile ranks are fraught with arbitrariness, because one is inherently invoking "rubberized" measurement scales in making these judgments, as emphasized earlier. Such "rubberized" scales will arbitrarily penalize one applicant with a given profile of scores, while arbitrarily benefiting another test taker whose profile is in a different portion of the distribution.

A separate, and equally worrisome, aspect of arbitrariness is introduced by the mere process of adding percentile ranks. Percentile ranks are *inherently confusing* to persons without advanced measurement training. Given paradigm influences, these decision-makers will therefore presume that all numbers can always be added to obtain a meaningful result, and will incorrectly think

that the criterion scores they derive are meaningful.

Academic departments housing people who know that percentile ranks are "rubberized" will make admission decisions that pay less attention to criterion scores, when such flexibility is allowed. Other departments housing less enlightened faculty will blithely add percentile ranks, and pay attention to their results when making their admissions judgments. This process is systematically structured to yield admissions results that are largely random across various admissions committees.

Of course, it might be argued that the degree of arbitrariness is minimal. For example, GRE scale scores and percentile ranks are highly correlated, as indicated in Table 3. It might be suggested that percentile ranks are "rubberized", but that maybe they're not "rubberized" a whole lot.

The problem with this argument is that the distortions introduced by adding percentile ranks are entirely *gratuitous*. There simply is no reason to introduce these distortions. It would be one thing to accept the cost of a distortion if some greater good offset this penalty. But there is no such benefit establishing any balance against the distortions that percentile ranks unavoidably create. The only reason for using percentile ranks is ignorance, and it is questionable whether a defense of ignorance will be viable.

Furthermore, it is simply *embarrassing* to have an admissions policy that requires adding percentile ranks, since it is clear from the previous discussion that there is no psychometric basis

for adding percentile ranks. And it would be the ultimate embarrassment to see one's colleagues with measurement training called by the droves to be qualified as expert witnesses and then have to testify under oath that their university's admissions policy is nonsensical. It would be especially embarrassing if one had to acknowledge that these problems were all described before the admissions policy was then adopted anyway.

The Etiology of Bad Academic Policy Formulation

Many naively believe that the strength of the academy is its dedication to knowledge. But because members of university communities so highly value knowledge, these same folk are also very hesitant to give up prior claims to knowledge or insight, and may be hesitant to seek contrary views or to perceive and respond to reasoned objection. Thus, discovering that the earth revolves around the sun can lead to excommunication, whether or not this new knowledge is important and true. When I have vested my career in a certain set of beliefs, it requires extraordinary character to say that the beliefs I represented as knowledge for many years were wrong.

Fortunately, the academy does have a special strength. That strength is not the intrinsic wisdom of those who people its faculty, and are purely human. Rather, the academy's strength is its fundamental dedication to protecting the free exchange of ideas. In the atmosphere of free discussion, the truth will usually ultimately out.

References

Ahmann, J.S., & Glock, M.D. (1981). Evaluating student progress: Principles of tests and measurements. Boston: Allyn and Bacon.

Asher, J.W. (1976). Educational research and evaluation methods. Boston: Little, Brown and Company.

Blommers, P.J., & Forsyth, R.A. (1977). Elementary statistical methods in psychology and education (2nd ed.). Boston: Houghton Mifflin.

Borg, W.R., & Gall, M.D. (1989). Educational research: An introduction (5th ed.). New York: Longman.

Bump, W. (1991, January). The normal curve takes many forms: A review of skewness and kurtosis. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 342 790)

Byham, W.C. (1983). Review of legal cases and opinions dealing with assessment centers and content validity. Pittsburgh: Development Dimensions International.

Carey, L. M. (1988). Measuring and evaluating school learning. Boston: Allyn and Bacon.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cunningham, G.K. (1986). Educational and psychological measurement. New York: Macmillan.

Eichelberger, R. T. (1989). Disciplined inquiry: Understanding and doing educational research. New York: Longman.

Gage, N.L. (1963). Paradigms for research on teaching. In N.L. Gage

(Ed.), Handbook of research on teaching (pp. 94-141). Chicago: Rand McNally.

Glass, G.V, & Hopkins, K.D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Groves et al. v. Alabama State Board of Education et al., Civil Action No. 88-T-730-N (Oct. 3, 1991).

Hills, J.R. (1981). Measurement and evaluation in the classroom (2nd ed.). Columbus, OH: Merrill.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1979). Applied statistics for the behavioral sciences. Chicago: Rand McNally.

Kirk, R.E. (1984). Elementary statistics (2nd ed.). Monterey, CA: Brooks/Cole.

Lincoln, Y.S., & Guba, E.G. (1985). Naturalistic inquiry. Beverly Hills: SAGE.

Mehrens, W.A., & Lehmann, I.J. (1991). Measurement and evaluation in education and psychology. Fort Worth: Holt, Rinehart and Winston.

Mehrens, W.A., & Popham, W.J. (1992). How to evaluate the legal defensibility of high-stakes tests. Applied Measurement in Education, 5(3), 265-283.

Moore, G.W. (1983). Developing and evaluating educational research. Boston: Little, Brown and Company.

Patton, M.Q. (1975). Alternative evaluation research paradigm.

Grand Forks: University of North Dakota Press.

Richardson v. Lamar County Board of Education et al., Civil Action No. 87-T-568-N (1989); U.S. Court of Appeals, 11th Cir., Nos. 90-7002, 90-7336, July 17, 1991.

Siegel, S. (1956). Non-parametric statistics for the behavioural sciences. New York: McGraw-Hill.

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

Stevens, S. (1951). Mathematics, measurement, and psychophysics. In S. Stevens (Ed.), Handbook of experimental psychology (pp. 1-49). New York: Wiley.

Stevens, S. (1968). Measurement, statistics, and the schemapiric view. Science, 161, 849-856.

Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan.

Tyler v. Vickery, 517 F.2d 1089 (5th Cir. 1975), *cert. denied*, 426 U.S. 940 (1976).

United States v. LULAC, 793 F.2d 636, 640 (5th Cir. 1986).

Wike, E.L. (1985). Numbers: A primer of data analysis. Columbus, OH: Merrill.

Table 1
Hypothetical Data for 100 Subjects

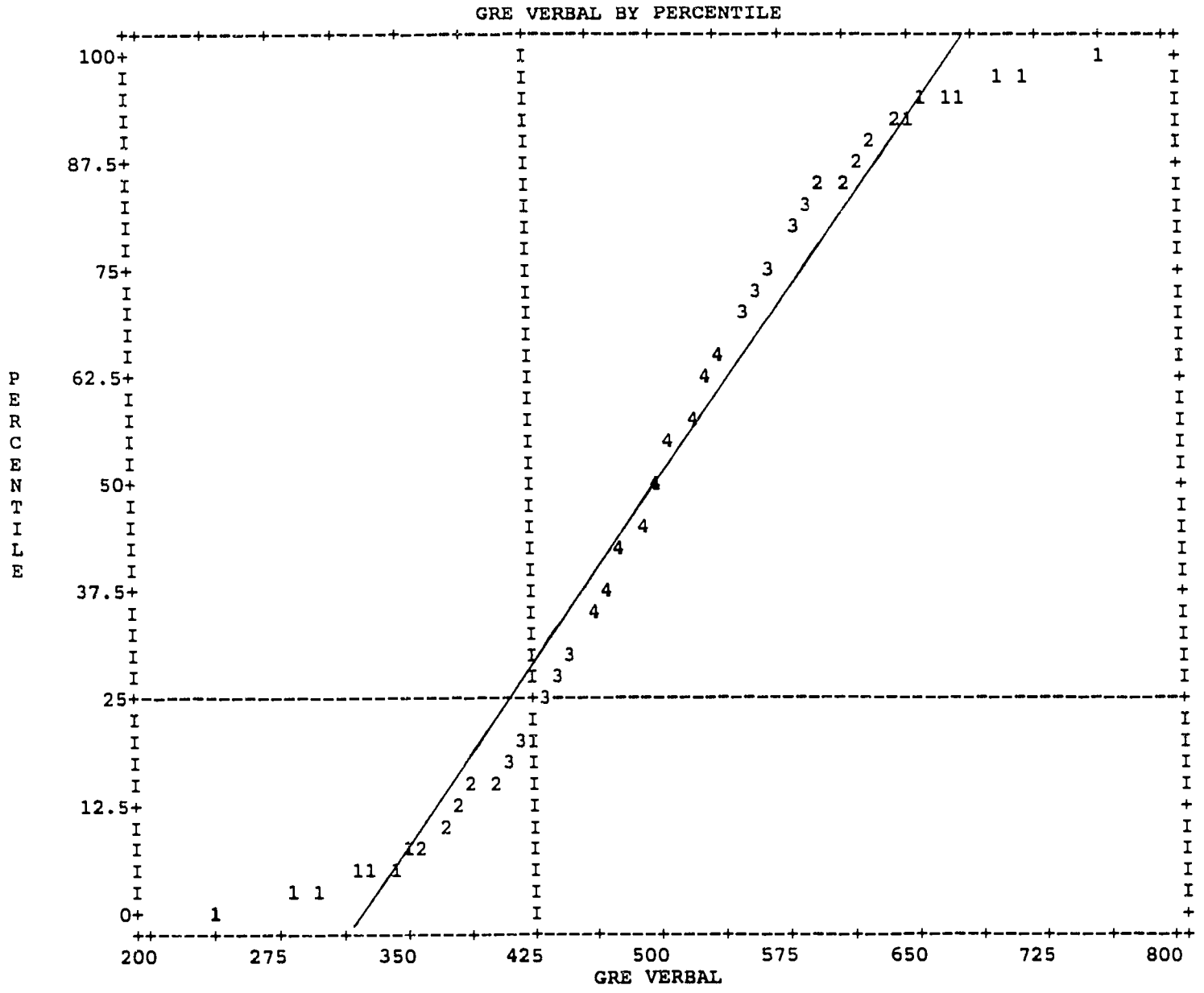
ID	VRIGHT	VERBZ	GREV	VPER	ID	VRIGHT	VERBZ	GREV	VPER
1	8	-2.6	240	1	51	112	.0	500	50
2	24	-2.2	280	2	52	112	.0	500	50
3	32	-2.0	300	3	53	116	.1	510	54
4	40	-1.8	320	4	54	116	.1	510	54
5	44	-1.7	330	5	55	116	.1	510	54
6	48	-1.6	340	6	56	116	.1	510	54
7	52	-1.5	350	7	57	120	.2	520	58
8	56	-1.4	360	8	58	120	.2	520	58
9	56	-1.4	360	8	59	120	.2	520	58
10	60	-1.3	370	10	60	120	.2	520	58
11	60	-1.3	370	10	61	124	.3	530	62
12	64	-1.2	380	12	62	124	.3	530	62
13	64	-1.2	380	12	63	124	.3	530	62
14	68	-1.1	390	14	64	124	.3	530	62
15	68	-1.1	390	14	65	128	.4	540	66
16	72	-1.0	400	16	66	128	.4	540	66
17	72	-1.0	400	16	67	128	.4	540	66
18	76	-.9	410	18	68	128	.4	540	66
19	76	-.9	410	18	69	132	.5	550	70
20	76	-.9	410	18	70	132	.5	550	70
21	80	-.8	420	21	71	132	.5	550	70
22	80	-.8	420	21	72	136	.6	560	73
23	80	-.8	420	21	73	136	.6	560	73
24	84	-.7	430	24	74	136	.6	560	73
25	84	-.7	430	24	75	140	.7	570	76
26	84	-.7	430	24	76	140	.7	570	76
27	88	-.6	440	27	77	140	.7	570	76
28	88	-.6	440	27	78	144	.8	580	79
29	88	-.6	440	27	79	144	.8	580	79
30	92	-.5	450	30	80	144	.8	580	79
31	92	-.5	450	30	81	148	.9	590	82
32	92	-.5	450	30	82	148	.9	590	82
33	96	-.4	460	34	83	148	.9	590	82
34	96	-.4	460	34	84	152	1.0	600	84
35	96	-.4	460	34	85	152	1.0	600	84
36	96	-.4	460	34	86	156	1.1	610	86
37	100	-.3	470	38	87	156	1.1	610	86
38	100	-.3	470	38	88	160	1.2	620	88
39	100	-.3	470	38	89	160	1.2	620	88
40	100	-.3	470	38	90	164	1.3	630	90
41	104	-.2	480	42	91	164	1.3	630	90
42	104	-.2	480	42	92	168	1.4	640	92
43	104	-.2	480	42	93	168	1.4	640	92
44	104	-.2	480	42	94	172	1.5	650	93
45	108	-.1	490	46	95	176	1.6	660	94
46	108	-.1	490	46	96	180	1.7	670	95
47	108	-.1	490	46	97	184	1.8	680	96
48	108	-.1	490	46	98	192	2.0	700	97
49	112	.0	500	50	99	200	2.2	720	98
50	112	.0	500	50	100	216	2.6	760	99

Table 2
Score Conversion Table for the Table 1 Data

GRE	Δ GRE	%tile	Δ %tile
240		1	
280	40	2	1
300	20	3	1
320	20	4	1
330	10	5	1
340	10	6	1
350	10	7	1
360	10	8	1
370	10	10	2
380	10	12	2
390	10	14	2
400	10	16	2
410	10	18	2
420	10	21	3
430	10	24	3
440	10	27	3
450	10	30	3
460	10	34	4
470	10	38	4
480	10	42	4
490	10	46	4
500	10	50	4
510	10	54	4
520	10	58	4
530	10	62	4
540	10	66	4
550	10	70	4
560	10	73	3
570	10	76	3
580	10	79	3
590	10	82	3
600	10	84	2
610	10	86	2
620	10	88	2
630	10	90	2
640	10	92	2
650	10	93	1
660	10	94	1
670	10	95	1
680	10	96	1
700	20	97	1
720	20	98	1
760	40	99	1

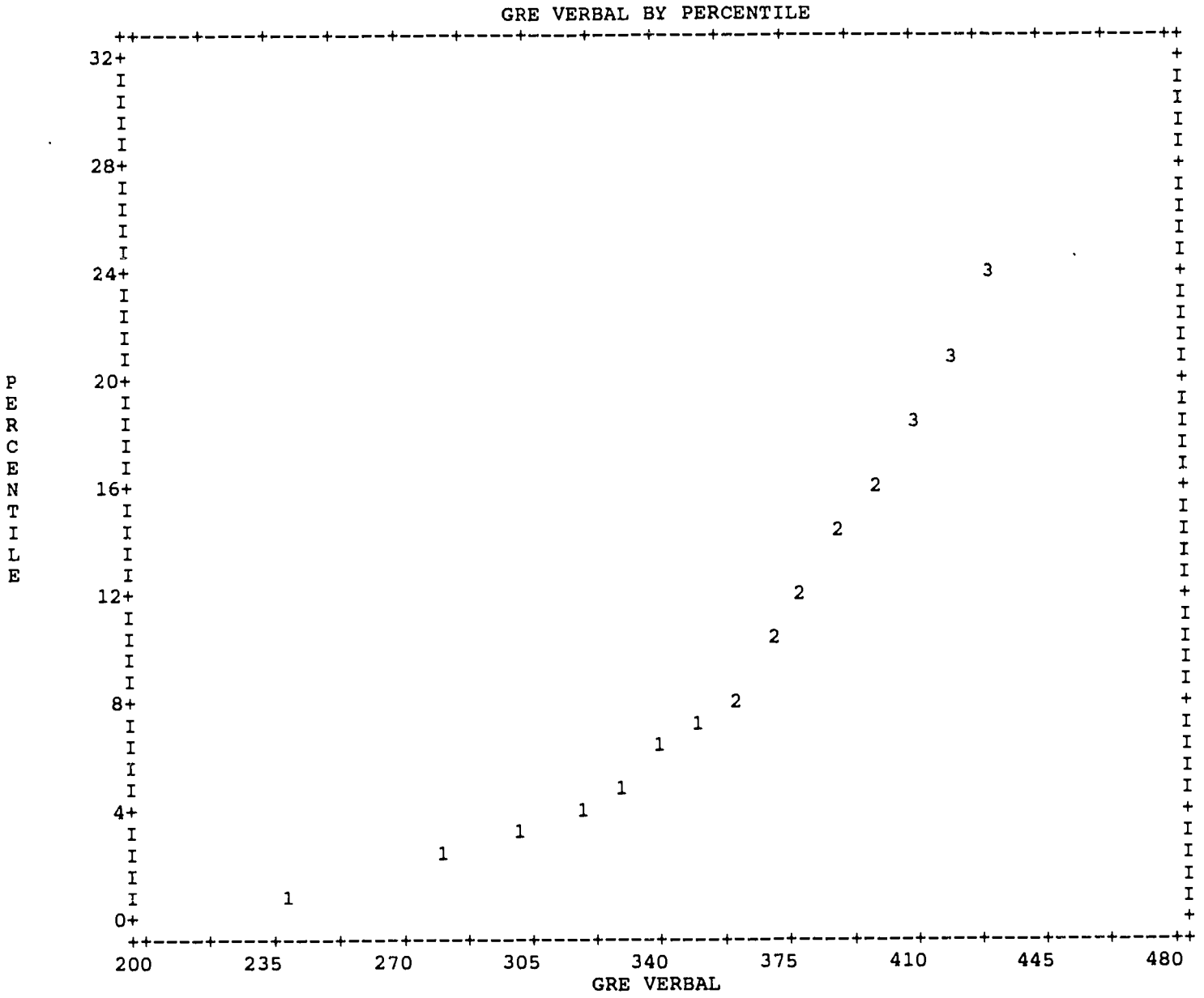
Note. The non-linear non-equal-interval nature of percentile ranks is indicated by the table. Going from a GRE score of 240 to 280, a 40 unit change in GRE score, results in a change of 1 percentile, while going from 500 to 540, also a 40 unit change in GRE score, results in a change of 16 percentile rank units!

Figure 4
Scattergram of GRE scores by Equivalent Percentile Ranks



Note. The numbers with the scattergram indicate the number of people at a given Cartesian coordinate, e.g., 1 person had a percentile rank of 1 and the associated GRE score of 240, while 4 people had GRE scores of 500 and the associated percentile rank of 50.

Figure 5
Scattergram of GRE scores by Equivalent Percentile Ranks
for the 26 Subjects Below the 25th %tile



APPENDIX A
Additional Quotations from Authors Noting That Percentile Ranks May Not Be Added or Averaged

"[Percentile rank] equivalents should not be used in data analyses involving descriptive or inferential statistics, however. The reason for not using equivalents in these analyses is that they have unequal units. For example..., if the mean of a test is 50 and its standard deviation is 10, a person with a score of 50 and a person with a score of 40 would be about 35 percentiles different from each other. However, two other persons with the same raw score difference of 10, but having raw scores of 40 and 30, would only be about 13 percentiles different from each other." (Borg & Gall, 1989, p. 340)

"Arithmetic and statistical computations of percentile rank scores cannot be meaningfully interpreted in some situations.... This can be seen from a simple example with data from Table 19.2; suppose group A consists of two examinees with raw scores 12 and 20, and group B consists of two examinees with raw scores 15 and 17. Both group A and group B have a raw score mean of 16, yet the means of their corresponding percentile rank scores are considerably different (40.5 for group A and 24.5 for group B)." (Crocker & Algina, 1989, pp. 441-442)

"For all the clarity and simplicity of percentile scores, they do not lend themselves to many statistical operations such as averaging and correlating scores. The difference in actual measured heights [in the example] between two men at the 50th and 52nd percentiles is very much smaller than the height difference between two men at the 97th and 99th percentiles.... Or, in IQ units P_{50} and P_{52} differ by less than one IQ point, whereas P_{97} differs from P_{99} by almost seven points. Standard scores avoid this problem and lend themselves readily to meaningful summary statistical calculations." (Glass & Hopkins, 1984, p. 66).

"The inequality of units requires special caution when using percentile ranks. First, a difference of several percentile points should be given greater weight at the extremes of the distribution than near the middle. In fact, small differences near the middle of the distribution generally can be disregarded. Second, percentile ranks should not be averaged arithmetically." (Gronlund & Linn, 1990, p. 349)

"The main limitation of percentile norms is that the percentile units are not equal on all sections of the scale.... Two implications of the unequal unit nature of percentiles should be remembered. One is that percentile ranks that are averaged arithmetically--by calculating a *mean score*--do not result in a meaningful value, hence should be discouraged.... The second implication is that small differences in percentile rank scores

near the middle of the distribution are not very meaningful." (Moore, 1983, p. 219)

"Researchers should be cautioned against the use of percentiles as variables in statistical analyses that require interval data [i.e., any analysis in which scores have to be added, subtracted, multiplied or divided as part of calculations] because the nonlinear transformation is likely to introduce distortions into the results. Even though most analysis procedures can be applied to data that deviate somewhat from being of an interval scale, the magnitude of the deviation from interval scale introduced by the use of percentiles [percentile ranks], coupled with the complete violation of the assumption of normality that accompanies their use [when parametric tests of statistical significance are conducted], could render the conclusions of such studies suspect." (Cunningham, 1986, p. 69)

"In short, percentile norms [ranks] are ordinal scales, not interval scales.... [Therefore...,] percentiles and percentile ranks as such cannot be treated arithmetically [e.g., added] and a meaningful product obtained." (Ahmann & Glock, 1981, p. 221)