3

Statistics, Data, and Special Educational Decisions

Basic Links to Realities

James M. Kauffman and John Willis Lloyd
University of Virginia

Should we identify a specific student as eligible for special education? What could be the consequence for our students if we use some specific teaching method with all of them? How might adopting this method affect the achievement of students who are struggling and those who are not? What might happen to our highest performing students if we employ a new method? To what extent are we willing to give up achievement gains in our high-performing students or our low-performing students for gains of average students? Do we have to choose between maximum improvement of the average test score and maximum gains in the highest or the lowest test scores? What should we anticipate if we adopt an effective prevention program? Answers to these and many other perplexing questions are not easy, and part of the reason that they are difficult is that educators too often ignore the mathematical and statistical aspects of them.

Not every important question can be fully understood as a mathematical equation, but all important questions have mathematical-statistical foundations. Thus, mathematics and statistics can help us reach better-informed conclusions. Controversies about classification, identification, effectiveness of services, and virtually all other issues reveal fundamental mathematical-statistical concepts that we need to understand better if we are to make better choices. In fact, in our judgment it is impossible to make morally defensible decisions without considering what we know about the mathematics or statistics involved.

Educators are not alone in failing to examine statistical realities. Tversky and Kahneman (1971) related an example in which specialists in mathematical psychology ignored problems with small sample sizes in making recommendations about experiments. Sometimes people overlook mathematical or statistical problems because we prefer stories to statistics (Kida, 2006). We may be more willing to accept and repeat a good yarn that fits our biases than to struggle with understanding and explaining mathematical or statistical ideas. Our judgments are usually based on intuitive understandings that are biased in well-known ways, especially when the situations we must judge are complicated and not transparent (Tversky & Kahneman, 1974).

Consider some examples. Most of us are more likely to discuss testing problems created by the No Child Left Behind Act (better known as NCLB) by telling stories than by explaining the statistical-mathematical realities involved in the law, even though the statistics of it might be nonsensical or absurd (Ho, 2008; Rothstein, Jacobsen, & Wilder, 2006). We would rather say how NCLB is absurd by comparing it to waving to Ray Charles than working through its statistical inadequacies (see Kauffman, 2005b). But the mathematical-statistical realities involved in education dog every decision and every policy with which we are concerned.

When states establish cut-off scores on tests for advancement, graduation, or other judgments of students’ achievement, certain mathematical-statistical realities apply. When we make judgments about disabilities based on test scores or performance or response to intervention (RtI), these statistical-mathematical principles apply. When people are selected for various jobs—including the job of teaching—based on measures of their performance, these concepts must be considered (Gladwell, 2008). In the economic difficulties beginning in late 2008 in which someone had to decide which individuals or institutions to help (e.g., with refinancing of their home mortgages or with institutional problems of debit and credit), statistical realities applied.

In short, we simply cannot escape statistical-mathematical issues, regardless of the social, economic, psychological, or educational program involved. And although good stories are often helpful, the success or failure of our policies ultimately depends on the math behind those policies, not just on what people like or believe. The mathematics or statistics behind most policies are not terribly complicated. The realities we discuss are well within the grasp of most adults.
We explain how, no matter how much we might think we are making decisions that are not affected by mathematics, we educators are actually making decisions that have important mathematical or statistical components. It is important to recognize the realities that constrain us so that our decisions do not merely become the object of humor. In the popular vernacular, we need to reject the thinking—in our profession endeavors—associated with Garrison Keillor’s entertaining tales of Lake Wobegon (Mr. Keillor’s imaginary home town, of which he speaks on the show, A Prairie Home Companion). As most of us know, in Lake Wobegon all of the children are above average. Most people recognize that this is statistically impossible, so it is funny. We do not laugh, however, when our fellow educators contend that it is possible for every child to be at or above grade level or that all students will be high achievers, although these suggestions are equally as absurd as Keillor’s description of Lake Wobegon. We will begin to address educational problems successfully when we see the tragicomedy in statements about education that ignore inconvenient statistical realities (Kauffman, 2010a, 2010b). To make the best decisions we can in the real world, we have to confront inconvenient truths, whether they are obvious or not.

The reason we focus on the mathematics or statistics underlying measurement and measurement theory is that measurement and comparisons of individuals within statistical distributions are an important part of identification of individuals needing special education and of assessing their progress. Actually, the mathematical realities to which we refer apply to all statistical distributions of measurements, regardless of whether they are measures of educational phenomena or measures of something else.

Particular measurement devices are critical in considering statistical data. For example, a test or other instrument could, in fact, be found to have admirable statistical properties, yet be useless or worse in helping anyone make good decisions about the classification or education of students. However, in our discussion the particular instrument is not the issue. The mathematics of statistical distributions is the issue.

We begin with measurement theory because measurement and related statistics are the primary bases for most decisions and for many controversies in education. Subsequently, we discuss some examples of problems that we face in making sound decisions. These include problems in improving education overall, identifying students for special programs, and preventing undesirable characteristics.

**Measurement and Measurement Theory**

“Theory” in mathematics and statistics has a meaning quite different from its popular meaning. In popular language, theory is often assumed to mean speculation, a concept that scientists and mathematicians would probably refer to as hypothesis. However, theory in mathematics, as in all sciences, refers to a body of organized knowledge that is based on reliable observation and that allows accurate prediction. Thus, “number theory,” “set theory,” “probability theory,” “measurement theory,” and so on do not connote speculation about what might occur. Rather, they refer to well-established relationships and predictions that can be tested and confirmed or disconfirmed.

Measurement is required for any scientific endeavor. In fact, measurement at some level is required for any kind of evaluation, even if the evaluation is qualitative and the simple conclusion of the evaluation is that \( a = b \) or that \( a > b \) or \( b > a \). Measurement is an inseparable part of accountability. Suggestions that measurement and accountability should be uncoupled are not coherent (cf. Kauffman & Konold, 2007). Granted, we may complain about particular measurement instruments, and unhappiness might be justified, but accountability requires measurement. Our discussion therefore assumes that the instrument used to measure is actually valid and reliable.

Most attempts to evaluate various individuals, groups, or programs involve more complex or sophisticated measurement than simply \( a = b \) or \( a \neq b \). As we learned in our fundamental courses, measurements may represent nominal, ordinal, interval, or ratio data.

1. Sometimes, the variables measured are *categorical* (female-male; qualified or not qualified; Asian, Black, or White). Even when the data are categorical or nominal, people often apply basic mathematical concepts. For example, they might say that event or condition A occurred more often than event or condition B.

2. Usually, however, educational measurement involves scores that statisticians consider at least *ordinal*. For example, the familiar categories primary, elementary, and secondary are aligned in an order. That is, they at least imply a rank order, like first, second, third, and so on.

3. Often, educational measures are assumed to be *interval* data, so that the steps between scores are assumed to be equal, as in the numbers 1, 2, 3, 4, and so on. In fact, most test scores are assumed to be interval data.

4. Only a few scores used by educators qualify as *ratio* data, a level of measurement in which the ratio of two scores is meaningful (for example, a change from 5 to 10 is equivalent to a change from 20 to 40). Some scores on curriculum-based measures, for example, qualify as ratio data.

Measurement theory helps us make sense of data. The organized body of knowledge that comprises the mathematics of statistical tests allows us to estimate with considerable precision the probability of obtaining various measurements and differences and to determine the confidence we should place in a given measurement. Measurement theory also tells us that all measurement contains error, that measurement with absolute precision (no error whatsoever) is not possible, and that measurement always produces a statistical distribution of values, including error.
“Error” in measurement does not mean “mistake.” It is simply unexplained variation, differences in measurement that have not been attributed to known factors such as the conditions in effect when the measurements were made, differences in the measurement system, and so forth. When measures have a relatively low proportion of error, they are better (i.e., more trustworthy), but they can never be completely free from error.

Most measurements of educational performance produce a score distribution that lies atop a continuous statistical distribution of outcomes (Kauffman & Konold, 2007; Konold & Kauffman, 2009). Continuous distributions are those in which what is measured varies from a little to a lot with fine gradations or increments being possible (cf. Kauffman & Hallahan, 2005). There are no natural, inherent, or obvious breaks in a continuous distribution. Height and weight are examples of continuous distributions, as are speed and rate. In education, we usually consider such things as intelligence and academic achievement to be continuous distributions.

The reason we say a distribution of obtained scores “lies atop” rather than “is” the distribution is that measurement of human performance always produces a discrete statistical distribution—one with “saw teeth” rather than one that is completely smooth. That is, the distribution can also be represented by a bar graph or histogram, and the “curve” representing it is really the histogram smoothed out. Another way of stating this is that regardless of the degree of precision of measurement (e.g., tenths, hundredths, thousandths…), there is a break in the scale when moving from one measured point to another (a sort of “saw tooth” or bar on a graph). Often, however, the data resemble a continuous, normal distribution, in that a graph of the scores approximates the symmetry of a bell by modeling the underlying continuous scale that is inherent to the variable being measured (hence, “bell-shaped curve” or “bell curve”). Not all distributions or curves are normal; some are lopsided (i.e., skewed) and some are more peaked (leptokurtotic) or flatter (platykurtotic) than normal. Some distributions are different from normal in both skew and kurtosis. And some distributions of larger populations may have “lumps” or subdistributions consisting of smaller groups.

Normal distributions do not really exist in nature but are mathematically idealized concepts. However, “many concepts in mathematics and science that are never quite true give good practical results nevertheless, and so does the normal distribution” (Hays, 1963, p. 218). Good, practical results are what we are after, and the math underlying the idea of the normal curve and other approximations of nature are often highly useful in making sound decisions about people. Moreover, all continuous distributions, whether normal or not, have immutable properties known to statisticians as moments.

Measurement theory includes four statistical moments: central tendency, variability, skew, and kurtosis. These are well-established realities that apply to all groups—including all groups of averages for schools, districts, states, nations, and subgroups of students. Any criterion used to categorize individuals (e.g., a student did or did not meet a criterion used to sort students into those needing or not needing additional instruction) are actually derived from the measurement of continuous distributions.

These realities lead inevitably to the issue of a distribution’s extremes and to the matter of marginalization. Marginalizing often means trivializing, isolating, or disenfranchising people. We do not mean to trivialize the serious issue of marginalizing people in the sense of disenfranchising or isolating them socially. Nevertheless, our topic here is the mathematical-statistical aspects of what might in some sense be called statistical marginalization. Whenever possible, we refer to this phenomenon with other terminology to avoid any possible confusion.

We use the term extreme score (EXT) to refer to a subset of the population relatively far from the central tendency in a statistical distribution. For example, a tested IQ of 35 is far below the mean IQ (100), so a score of 35 would be considered an EXT—far below average. Remember, too, that the scores we consider to indicate an Intellectual Disability (ID) is a relatively small group compared to those judged not to indicate ID and that ID may have its own distribution and its own EXTs, both those far above and far below whatever IQ is the mean for ID.

We use the term close to criterion (CTC) to refer to scores that are in close proximity to a cut point and are very close to (just below or just above) the cut point. Suppose that the cut point for what we call ID is an intelligence quotient (IQ) of 70. Then an IQ of 67 or 72, for example, would also be considered a CTC score. Or, we might say that gifted IQ starts at 135. But, then, a student with an IQ of 134 would be considered to have a score that is CTC too, as we use the term, and one with an IQ of 176 would have a score that is EXT. That is, a score that we consider a CTC is located in the proximity of a line of demarcation on a particular measure and an EXT is very clearly within the smaller subset defined by that cut point. In education, the criterion for whatever we measure may be either low or high in the distribution. That is, it could be in either tail (extreme high or extreme low) of the distribution. Obviously, in any given distribution we could have two criteria, one high and one low (we might refer to CTC+ or CTC– or to EXT+ or EXT–).

We might use another example of these concepts: those who are CTC financially or EXT on a measure of financial worth. Financial CTC– includes are those who are near but just below a criterion of low financial assets, and those with EXT– are clearly in a group we might consider “poor.” Financial CTC+ includes those who are near but a little over a criterion of great financial assets, and those with EXT+ are clearly in a group we might consider “rich.” So people could be considered poor if their net worth is $10,000 (if that is the cut point) or below, but those whose net worth is $10,500 would be considered CTC+ because they’re very close to and just over the cut point for “poor,” and a net worth of $5,000 would be way beyond the cut point (EXT–).
We might point out, too, that the designation of EXT and CTC are arbitrary but useful concepts. Although statistical moments may be well established and immutable phenomena, the designation of a distribution’s tails, or EXTs, is a matter of choice, not a mathematical function. Moreover, CTC is also a matter of judgment. We do recognize the fact that the designation of tail of the curve, EXT, or CTC are arbitrary or judgmental, although the facts of a distribution and its moments are not. And although these are arbitrary designations, they are extremely useful in the sense that voting age, driving age, senior citizen, low birth weight, and many other arbitrary decisions about continuously distributed variables are useful, if not unavoidable, if social justice is to be realized.

**Statistical Effects of Better Education for All Students**

Better education (or better educational outcomes) for all students is in our judgment a good idea. We believe that we have a moral obligation to give every student the best education we can. At the same time, we realize that we may have to face the prospect of trade-offs in helping one group versus another. Moral and ethical decisions are not easy, and they are sometimes complicated by knowing more about the likely outcomes of our actions. Nevertheless, making decisions without considering available knowledge is not wise either, in our opinion.

To say that we seek better education for students begs the question, “Better than what?” The word “better” is a comparative, so there is a mathematical foundation for it. Something is better than something else. Although we do not have an absolute answer for what constitutes a better education, we think special education has specified the fundamental process for moving toward it: In concert with parents, identify the most important goals for education of students; implement with fidelity the most effective educational practices for achieving those goals; monitor students’ progress toward meeting the goals; adjust instruction on the basis of students’ progress.

Our efforts to educate all students to the best of our ability must be informed by mathematical-statistical realities or we will almost certainly fail to make the best decisions we can or even to know what we are doing. For example, we need to know whether giving equal effort to improving every student’s performance is likely to increase or decrease population variance—the “spreadoutness” of scores. The issue is really whether equally improved education for all students will have the effect of increasing or decreasing the homogeneity of the population. We need to ask ourselves, “Given equal educational effort, will differences between high and low achievers be reduced or become greater?”

The effect of equalizing inputs may be small, but it is important, even if it is small. To the extent that the population becomes more homogeneous by even a little, the effect of better education for all students will be consistent with the desire to close the gap between high and low achievers. To the extent that the population becomes less homogeneous by even a little, the effect is inconsistent with the desire to close the gap or distance between high and low achievers.

We also need to consider what might happen to the distribution’s skew and kurtosis if we improve education for all students equally. If the distribution has a negative skew (i.e., the left tail representing lower achievement is elongated), then more than half of the population will assuredly fall above the arithmetic average (mean); the opposite is true, too, so that if the distribution has a positive skew, then it is entirely predictable that less than half of the population will be above the mean (arithmetic average). For example, it is possible that, depending on how teacher effort is distributed across individuals, more low performers may advance toward the mean than high performers gain in distance from it, and the distribution of scores may become more leptokurtotic (the standard deviation may shrink). This seems a likely outcome based on measurement theory if we give disproportionate effort to increasing the low achievers’ compared to high achievers’ scores. Alternatively, it is possible that the distribution of teacher effort may result in gains for all students but that high performers may gain more in absolute performance than do low performers, thereby expanding the range of scores (i.e., the distribution may become more platykurtotic, having a larger standard deviation). This seems likely if we give equal effort to improving the achievement of students regardless of their current level of performance.

We may want to believe that no tradeoffs are required in education and that we need never give up one thing for another. However, this does not seem likely on the basis of a mathematical analysis of the problem. Gerber (2005) has shown how mathematical realities related to economic theory inevitably limit a teacher’s distribution of effort when teaching more than a single student. He has shown how teachers must make choices about the dispersion of their efforts to achieve maximum mean achievement gains when teaching groups. The reality is that teachers cannot be all things to all students, even if they expend optimum effort. Mathematical realities dictate that in any given instructional circumstance they must give up achievement gains of one student when they seek to increase the achievement of another. If a teacher is disbursing 100 units of teaching effort and she or he needs to reallocate units to a struggling student or group, then the teacher must take units from some other student or group to allocate them to the student(s) to whom she or he delivers extra help. This is not because teachers are reckless, but simply because their efforts are constrained by the mathematical realities of finite capacity. Gerber’s analysis does mean, however, that “differentiated instruction” offered in larger groups cannot include the concentrated teacher attention that can be given in smaller groups or in individual instruction.

Policy makers must remember these aspects of statistical distributions if they are to enact policies that achieve their intended effects. To the extent that the improvement of instruction for all children is the goal, such that each child’s
potential for learning reaches its maximum, population variance (difference among students) is likely to increase (see Kauffman, 2002, for discussion of the effects of differences in rate of learning). The late physicist Richard Feynman (1985) said, “For instance, in education you increase differences. If someone’s good at something, you try to develop his ability, which results in differences, or inequalities” (p. 281). The basic mathematical-statistical reality is that if you help all students equally to get better at something and they already differ in their ability to do it—whether their differences are innate or due to experience such as exposure to or practice of what you are teaching them—then the differences among them become greater.

To the extent that all students reaching an established criterion of performance is the desired end, population variance (difference among students) is likely to decrease. Devoting units of educational effort to raise the scores of low achievers up to some specified standard is likely to come at the expense of devoting those units to promoting the performance of high achievers (see Kauffman & Konold, 2007; Konold & Kauffman, 2009). In any case, consideration of the effects of programs intended to improve educational progress overall should include the entire statistical distribution, not simply assess differences in means or gaps based on cut scores (Ho, 2008). Indeed, increasing what most students can do (the average, whether defined as mean or median) is desirable, but an important consideration is also whether this increases population variance and what happens at the extremes of the distribution (the outliers). Just as having a few wealthy software engineers move into a middle-class neighborhood would raise the average per-household income, boosting the scores of a few very high-achieving students would raise the average score for a local school.

These are problems involving social justice as well as statistics, but our judgments must be made in the light of what we can predict on the basis of measurement theory about effects on all parts of a statistical distribution of outcomes. Ho (2008) suggests that the old injunction to measure twice and cut once should be rephrased for thinking about cut points in education; cut once, measure everywhere, he advises.

Effects on Extreme Scores (EXT) and Scores Close to Criterion (CTC)

The effects of education on students whose scores are EXT or CTC (i.e., those under the tails of a distribution or close to a cut point) are important for at least two reasons. First, on tests of psychological and educational variables, the students whose scores are EXT or CTC are likely to be considered exceptional children, either because they are identified as having disabilities or because they are identified as gifted. Second, these students can contribute significantly to the disproportional representation of a given group in any activity for which performance is measured and either exceptionally high or exceptionally low performance is the selection criterion. Taking an example from athletics, it could be that different ethnic groups have about the same average running ability, but if, say, those who are EXT in ability to run very fast are predominantly from a particular ethnic group, then that ethnic group is going to be disproportionately represented among fast runners (see Gladwell, 1997). In education, small differences in students with EXT scores might result in serious disproportionality among those identified as exceptional in any given category even though the averages of the groups in question are not significantly different. Gladwell (1997) points out that many people misinterpret the meanings of the central tendency and other aspects of statistical distributions and, therefore, misinterpret the meanings of disproportionality in any area of performance. It is exceptionally important, therefore, that we better understand what happens when we establish cut points in statistical distributions for special education programs.

By definition, the tails of a distribution (or EXT scores) involve relatively low numbers and small percentages of a sample or population (see also Ho, 2008). Thus, if a given effort to improve education results in even small changes in EXT in a distribution, it can have substantial effects on the proportion of minority members included, whether the minority is defined by race, gender, ability, or any other specified characteristic that may be unrelated to the average (mean or median) of the variable being measured. Consequently, decisions that involve (or affect) either tail of a distribution without consideration of (or changing) the central tendency may have profound moral and practical implications. When educators should focus on the central tendency and when on the EXT or tails and when on both are matters of great importance, but confusing the meaning of typical and EXT certainly will lead to tragicomic decisions (see Kauffman, 2010b).

Unavoidability of Lines and Margins

Qualification of an individual for any special program requires establishing criteria for participation. Sometimes, the criteria themselves are categorical (e.g., male or female, is or is not a citizen of a particular state), but categories are often established based on arbitrary lines drawn in a continuous distribution (e.g., age of designation as “senior,” voting age, financial assets required for a home loan, pecuniary assets excluding someone from an assistance program, test scores required for admission, height required for admission, and so on). Moreover, the line in a continuous distribution, regardless of where it is drawn, can be changed at will (i.e., it is arbitrary, although it may be defensible). So, for example, we could say that a discrepancy of 20% difference between a student’s performance in reading and the local norm or expectation defines need for additional assistance in reading. But, what of 21%, or even 20.5%, or, say, 19%? We could, in fact, question any CTC as well as the distribution, regardless of where it is drawn, can be changed at will (i.e., it is arbitrary, although it may be defensible).
but those who are going to do something to help students have to face the mathematical realities of categories and line-drawing (Kauffman, 2009). That is, they have to make decisions about the qualification of particular individuals for particular programs. Not every individual can be included in every education program that is offered, nor is such universal, all-pervasive inclusion defensible on logical or moral grounds. So, what should be the criterion, regardless of how it is determined, and what do we do about the near misses (CTCs)? We have no ready answer. Our point is that we cannot avoid the question. Finding a new instrument never solves the basic problem.

Inevitably, lines drawn in continuous statistical distributions create margins—measurements that approach but do not quite reach the designated criterion and those that surpass it by only a little (CTC+ and CTC–). Neither drawing the line nor creating the margin can be avoided if a subset of the larger group is to be designated. The matter is just this simple—no line, no special program. And the matter of margins is equally simple—every line has margins, which include regions close to the line; one is close but does not quite reach the criterion, and the other is close but just a little beyond the line. The line’s margins can be determined by a variety of mathematical computations, but those whose scores are CTC are always an arbitrarily designated and necessary subgroup.

**Prevention**

Kauffman (1999, 2004, 2005a, 2007, 2010a) and Kauffman and Landrum (2009) have described how drawing lines and dealing with the line’s margins are part of the problem of early intervention and prevention. The line is the cut point established for deciding which students qualify and which students do not. The margins are the close calls either way, a little above or a little below the cut point (CTC+ or CTC–). Margins may also be thought of as regions of statistical error, the plus-or-minus distance from the score that statisticians call a standard error. The margins or close calls also tell us something about the kind of mistake we might make, a *false positive* or a *false negative*. A false positive means that we identified a student when we should not have. A false negative means that we should have identified a student but did not. Both are mistakes in identification. In a false positive, we got the wrong one; in a false negative, we failed to get the right one.

Consider an example involving special education. We could identify a student as eligible for special education due to an emotional or behavioral disorder when, actually, the student does not have this disorder. His identification was a mistake. He was falsely identified. His case would then be considered a false positive. Or we could miss identifying a student who we should have identified. We find out later that although he was not identified, he should have been. We made a mistake. In this case, our mistake was nonidentification. His case would then be considered a false negative.

Let us think about disabilities and false positives and false negatives further. Actually, false positives may be found at any location in the distribution that is designated as the smaller subset for which services are intended (that is, it is possible for a student to be tested and found to be extremely discrepant from the norm, yet to be falsely identified). The opposite kind of error, false negatives may be found at any location in the distribution designated as the larger subset for which services are not intended (that is, it is possible for a student to be tested and found to be typical or even superior in relation to the norm, yet be missed and remain unidentified). People do not make mistakes only about students who are close to a cut point (CTC). But when people make a mistake, they usually make it about a close call. To the extent that the measurement tool is valid and reliable for the variable in question, the false positives will be concentrated within the margin of the criterion further from the central tendency and the false negatives will be concentrated within the margin of the criterion closer to the median (or other measure of central tendency)—that is, the highest probability of error is in cases close to the cut point (CTC).

The reason that false positives and false negatives are related to prevention is that in prevention we are trying to keep something from happening—catch it before it happens and prevent it. So, we have to think about how we would identify cases before they get to a certain point, either before they occur at all (primary prevention) or before they get as bad as they are by the time we usually catch them (secondary or tertiary prevention). So, prevention requires that we think about the kind of mistake we might make (false positive or false negative) if we try to prevent something—unless we believe we can avoid mistakes altogether (not likely, being human).

The following comments apply to secondary and tertiary prevention only, not to primary prevention for which no identification of a subset of the population in question is necessary (see Kauffman, 1999, 2005a; Simeonsson, 1991, for discussion of differences among primary, secondary, and tertiary prevention). When we consider what is required for prevention and early intervention in the light of statistical distributions, it is obvious not only that more students must be identified as needing help but that the risk of false positive identification will increase, given the same means of identification. Prevention cannot occur without encountering greater risk of false positives because of the mathematics of distributions, not because of preferences for or aversions of certain risks or because of ideologies (Kauffman, 2010a). The reality is simply that as the criterion for identification is moved from more extreme or obvious to less extreme or obvious cases a greater number of cases will be included and more cases will be found at the margins of the cut point (i.e., more will be CTC+ or CTC–, depending on whether a lower or higher score is considered a sign of danger). This is true whether the distribution is for age of onset, duration of the problem before intervention, or intensity or degree of disability. For example, if we say that students qualify as having an emotional or behavioral disorder when their score is 2.0 standard deviations above the mean on a measure of
misbehavior, then if we move the criterion to 1.75 standard deviations from the mean (i.e., we move it closer to the central tendency), we’re going to have a greater number of students (the area designated under the curve is greater).

And, because more individuals will be identified as eligible for the program, the risk of making a mistake known as a false positive will be greater. The only exceptions to this reality require (a) a significantly improved means of identification and the finding that a smaller percentage of the population actually has the condition for which identification is necessary than was previously thought to be the case or (b) the finding, using current methods of identification, that the prevalence of the condition is, indeed, lower than believed.

Noting the mathematical realities involved in the risk of false positives does not discount the importance of early intervention and prevention in special education. In fact, given the realities and risks, we would argue for these efforts (e.g., Kauffman & Brigham, 2009; Lloyd & deBettencourt, 1986). However, it is important to acknowledge the mathematical realities that are involved—that there will by necessity be an increase in the number of students identified as needing or likely needing special services. Moreover, the same effects will be seen regardless of the scale used for assessment and regardless of whether the issue is prevention of school failure, prevention of child abuse, or prevention of any other undesirable phenomenon.

**Effects of Kurtosis and Skew on Extreme Scores (EXT) and Those Near a Criterion (CTC)**

Kurtosis and skewness alter the effects of decision rules, especially for values or scores near the cut point or criterion (CTC). Educators might make decisions on the basis of a cut score (e.g., all individuals with values below a given value are eligible for a program) or on the basis of a percentage (e.g., the lowest given percentage of values indicate eligibility for the program). Here, the statistics become somewhat more abstract, but think about what the shape of a distribution (lopsided one way or the other or more peaked or more flattened) would mean for who gets identified for a special program.

Greater kurtosis means that more of the variance in a distribution is due to infrequent and extreme deviations from the mean than from the more frequent and less extreme deviations in a normal distribution. Thus, the distribution is more peaked than normal (and the standard deviation is smaller). If a particular absolute value below or above the mean is chosen as a cut point (say, a score of 80 on a given test) and the distribution of scores remains constant in its skewness, mean, and standard deviation but the distribution becomes more leptokurtotic, fewer people will have scores that are EXT or CTC according to our definition. However, given the same assumptions about the distribution but a percentage of the population (say, 20%) or a standard deviation (say, a standard deviation of ~1.25) is chosen as the rule for eligibility (or another decision) and the distribution becomes more leptokurtotic, then the same percentage will have EXT or CTC scores by our definition.

Also, with respect to those on the end of the distribution below the central tendency, a negative skew is more likely to result in fewer individuals with scores at the margins than is a positive skew, given that the criterion for designating those with EXT or CTC scores is a fixed value on the measurement scale, not a percentage of the population. Of course, the opposite is also true under similar assumptions; a positive skew is more likely to result in fewer EXT and CTC scores on the positive end of the distribution than is a negative skew. And, if having an EXT score is defined as a percentage of the population (say, 15%) or a standard deviation (say, 2.0), then the number or percentage of individuals having an EXT score cannot change, regardless of skew or kurtosis (i.e., absolute value can change, but not value relative to the mean). If, for example, educators decide that 10% of all students will receive a service (say, eye glasses), then the shape of the distribution of need-eye-glasses scores would be irrelevant; this is true no matter whether the cut point is 5% or 15% of students.

**Alternative Measurements Used for Eligibility for Special Education**

Ordinarily, a single line of demarcation is used to indicate eligibility for a special education program. For example, special education in any given categorical area is usually a matter of determining eligibility on the basis of a single criterion (although one that may, as in the case of learning disabilities [LD] defined as a discrepancy between ability and achievement, be based on multiple test scores). In fact, for a sample of students considered for eligibility for special education services as students with learning disabilities, there would be a set of discrepancies between ability and achievement. That set of scores will, of course, have statistical moments. Some discrepancies will be above the mean and others below it. If the distribution is roughly normal, about 50% will be above the mean. Moreover, the cut score on this distribution will be arbitrary and subject to all of the statistical phenomena we have discussed to this point.

Alternative procedures for identifying students as eligible for special education (e.g., response to intervention or RtI) face the same problems—continuous distributions (in the case of RtI, responsiveness to given instruction) and an arbitrary criterion for selection. That is, all of the following aspects of RtI are continuously distributed: (a) nature or quality of instruction, (b) length of time given instruction has been used, (c) responsiveness to instruction, (d) starting levels (i.e., intercepts), and (e) rates of progress. Thus, in no way does RtI overcome the statistical realities that apply to other measurements or methods of identification. The possible advantages of RtI may be found in (a) the benefits of employing primary prevention (i.e., using curricula and teaching methods for all students in general education that demonstrably yield lower failure rates) and (b) the validity
James M. Kauffman and John Wills Lloyd

and reliability of its measurement, not its avoidance of mathematical or statistical criteria related to the nature of distributions.

Other Statistical Issues and Problems

Other statistical matters that must be considered include error and subdistributions. Ordinarily, an error term (or standard error) is calculated for an entire distribution and assumed to apply to the extreme scores under a distribution as well as to scores near the central tendency. In some cases, however, it is extremely difficult and perhaps risky to assume that error is equally probable near the central tendency and several standard deviations from it.

Moreover, if a subdistribution (of LD, for example) is fully contained within the larger statistical distribution with a smooth tail that is characteristic of idealized or hypothetical curves, then all the effects we discuss apply to the larger distribution. However the subdistribution may also contribute a “hump” or “bump” to the tail of a larger distribution, such that the usual statistical assumptions or mathematical calculations must be altered somewhat or are indeterminate (we might call such a bump or lump a perturbation of the curve’s tail). For example, there might be a “hump” in the lower end of the distribution of adaptive behavior. Moving away from the central tendency of the larger distribution, the left tail might go up before it goes down again toward the baseline. In that case, moving a cut point for prevention toward the central tendency of the larger distribution will still result in more cases being identified, although depending on just where the cut point is moved fewer cases than would be expected without the perturbation would be identified. A greater number of cases will be identified simply because moving a cut point toward the central tendency of a distribution always, without fail, regardless of any perturbations, includes more individuals—a larger area, a larger percentage of the sample represented by the distribution—regardless of the shape of the distribution. Given that the cut point is moved from within the perturbation’s downward slope and toward the central tendency of the larger distribution’s, the risk of false positives could decline and would rise again only if the cut point were moved up the slope of the larger distribution.

The fact is that for many variables, we are not certain just what the statistical distribution is. The “normal” distribution gives us only an idealized model. It may be a very useful model and may be the best guess we can make, but if we actually obtained the data and sketched a distribution of them, we may be surprised at what we find.

Discussion

Although we have described neither all of the statistical-mathematical realities that apply to the data of education or any other endeavor nor all of the problems or issues in education to which these might apply, we have examined some of the cornerstone statistics and some of the primary issues involved in education. Certainly, it is important to apply statistical-mathematical realities to the distributions of actual data, not only to hypothetical distributions or models. Nevertheless, we make two observations: (a) the principles we described apply to actual data as well as to hypothetical distributions and (b) sometimes a hypothetical distribution or an approximation of existing data is the only or the best model we can obtain.

As special educators, we are particularly concerned about EXT and CTC scores and the effects that particular interventions may have on the shapes of statistical distributions, especially students whose scores are EXT or CTC. What happens to the extremes or tails of distributions can have a profound effect on public perceptions as well as the individuals who fall far from the central tendency on any given measure. Although two publications of the National Research Council on science in research on education contain many important ideas, they include no discussion of topics important to research in special education—statistical distributions and EXT or CTC scores in them (Shavelson & Towne, 2002; Towne, Wise, & Winters, 2004). We are disappointed that these publications do not address scientific issues that are essential to special education (see also Kauffman, 2011).

We are also disappointed that an article by Gladwell (2008) includes neither discussion of statistical extremes in groups of students being taught nor cut points in statistical distributions of teaching finesse. Although Gladwell discusses the dramatic differences between success in college football and success in professional football, his discussion of education does not include differences between teaching children with disabilities and those without disabilities. As Kauffman (2010b) notes, we could make the assumption that no student is truly special for instructional purposes. Like award-winning general education teacher Esquith (2007), who makes no mention whatever of students with disabilities, we could assume that if the teacher teaches well, then children with disabilities are just like everyone else in the class, so no one needs to mention them or what is required to teach them. Our suspicion is that good teaching of modal students and those who are exceptional for instructional purposes may be substantially different.

One implication of our description of statistical and mathematical realities is that educators and those who make policy decisions about education should take these into account in discussions of the data they have and their descriptions of the data-based outcomes they desire. Although it is possible for anyone to propose programs or policies based on wishes that are not reality-based, as discussed by Kauffman and others (e.g., Kauffman, 2010a, 2011; Kauffman & Konold, 2007; Konold & Kauffman, 2009), such actions are, in our opinion, ill-advised.

If someone is serious about secondary or tertiary prevention of a difficulty (regardless of what it may be), then the statistical and mathematical certainties are that, given the identification procedures now practiced, (a) more
individuals, not fewer, must be involved in the preventative intervention and (b) greater, not lesser, risk of false positive identification will be encountered. These certainties will apply unless (a) the accuracy of identification is improved over current methods and the actual prevalence of the condition is found to be lower than previously believed, (b) the actual prevalence of the condition is found to be lower, using the current method of identification, than previously believed, or (c) the distribution is found to have a perturbation and movement of the cut point is related to it in the way we have discussed. Thus, if someone argues that students are now often mistakenly identified for a service or intervention (say, LD) and that such misidentification should be reduced but that we need to practice prevention of LD, then it is incumbent upon that person to show how identification of LD can be made sufficiently more accurate by using alternative identification method B relative to identification using current method A and that the true prevalence LD is lower than previously thought—or, simply, that the prevalence of LD using current identification method A is considerably lower than previously thought. In the absence of such demonstration, one must assume that identification method A = identification method B in accuracy, that the true prevalence of EBD or LD is as presumed, and therefore that prevention requires risking more false positives. However, a perturbation in the distribution and movement of the cut point required for prevention could be demonstrated to occur as discussed above, resulting in a lowered risk of false positives until the cut point reached the region in which the perturbation gives way to the distribution for the larger population.

An important point here is accuracy of identification, such that both false positives and false negatives are reduced. That is, it might be possible to devise an alternative identification method B that reduces identification for special education, yet results in increases in false negatives. In that case, method B might appear, at least for a time, to result in prevention while reducing the number of students identified, but in the long run it will become obvious that nothing has been prevented, only that more students have gone without identification and treatment.

Another implication is that some current efforts to reform education are seriously misguided. For example, a major goal of the No Child Left Behind (NCLB) law is eliminating gaps between the mean test scores of students with and without disabilities. Besides recognizing the fact that this is logically impossible (cf. Ho, 2008; Kauffman, 2005b, 2010b; Kauffman & Konold, 2007; Konold & Kauffman, 2009; Rothstein et al., 2006), educators need to consider the consequences of differentially disbiasing instructional resources on the nature of statistical distributions of outcomes (Gerber, 2005; Kauffman, 1990, 2002). Would the result of investing very heavily in raising the scores of those below the mean cause a distribution to change its shape in predictable and desirable ways? Or would the result be change in undesirable ways? For example, would most of the scores on the low side of the normal distribution simply be pushed up to the mean, or would the dedication of resources to students below the mean result in a concomitant reduction in resources, and an associated lowering of outcomes, for those who scored above the mean under previous conditions?

We argue that better thinking and more responsible educational reform are possible only when our thinking is constrained by the “box” of statistical and mathematical realities. Some outcomes that educational policies describe implicitly or explicitly as desirable (e.g., increasing mean achievement while reducing population variance in test scores; eliminating mean test score gaps between certain groups) may not be achievable in the real world. Recognizing the difference between the thinkable and the possible is something with which educators have struggled, but not always successfully. We emphasize our agreement with philosopher Susan Neiman, who stated, “This is important: Not everything that’s thinkable is genuinely possible, and distinguishing between the two is what allows us to distinguish between demands for utopia and for responsible social change” (2008, p. 142).

In other words, as suggested by the title of Thomas Kida’s (2006) book and a bumper sticker we have seen, “Don’t believe everything you think.” We advocate responsible social change, which gives more learning opportunities to individual students of every description. Ignoring statistical and mathematical realities in pursuing this goal is in our opinion inimical to achieving the responsible social change we desire.

References
Kauffman, J. M. (1990, April). What happens when special education works? The sociopolitical context of research in the 1990s. Invited address, Special Education Special Interest Group, American Educational Research Association Meeting, Boston, MA.