The Theil–Sen Slope for High-Stakes Decisions from Progress Monitoring

Kimberly J. Vannest, Richard I. Parker, John L. Davis,
Denise A. Soares, and Stacey L. Smith
Texas A&M University

ABSTRACT: More and more, schools are considering the use of progress monitoring data for high-stakes decisions such as special education eligibility, program changes to more restrictive environments, and major changes in educational goals. Those high-stakes types of data-based decisions will need methodological defensibility. Current practice for summarizing progress monitoring data is to use a hand-fit trend lines (for practitioner use) or linear regression (for research). This study critically examines both approaches and compares them to a new nonparametric slope called the Theil–Sen. A field test with 372 published data series compared hand-fit, linear regression, and Theil–Sen slopes against evaluative criteria of power and precision, meeting data assumptions, and agreement with visual judgments. Results indicate promise for Theil–Sen slope in defensible high-stakes decision making.

■ Progress monitoring (PM) is the repeated measurement of behavioral or academic performance over time by equivalent probes or observation protocols. Students with emotional and behavioral disorders (EBDs) are frequently progress monitored in both academic and behavioral domains. Both special education and general education instructional environments see progress monitoring as an increasingly used evaluation tool (National Center on Student Progress Monitoring; National Center on Response to Intervention). The history of PM use in assessing individual client or student interventions extends at least 50 years, from early applied behavior analysis in the 1960s through precision teaching in the 1970s and curriculum-based measurement in the early 1980s (Baer, Wolf, & Risley, 1968; Bolger, 1965; Bijou & Baer, 1961; Estes, 1956; Sidman, 1960; Skinner, 1966)

The various forms and applications of PM have in common: (a) frequent or periodic assessment, (b) equivalence of probes or observation protocols from one to the next, (c) brevity of probes or observations, (d) relevance of the measured or observed performance over a medium-to-long period of time, and (e) graphic and statistical summary of scores as "improvement rate," "performance level," or slope. The analysis of performance in this way was historically an instructional decision-making tool where frequent reevaluation and nonpermanent decisions were the hallmarks. For example, progress monitoring the oral reading fluency of a student with EBD in a self-contained classroom may produce data used to change instructional methods from one unit to the next, or implementation of a self-monitoring intervention may be maintained based on homework completion or on-task data used to monitor progress in a behavioral goal or objective.

The response-to-intervention (RtI) model of this decade has increased the interest in PM. In particular, PM is increasingly relied on for the main data source in judgments of intervention responsiveness, decisions to intensify interventions, referral for special education services, and even eligibility for those services (Berkeley, Bender, Peaster, & Saunders, 2009; Fuchs, Deshler, & Reschly, 2004; Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Gerber, 2005; Hale, 2006; Kavale & Forness, 2000; Kavale, Holdnack, & Mostert, 2006; Naglieri & Crockett, 2005). School administrators and general education teachers are prompted to join school psychologists and special educators in understanding and carrying out PM (Hintze & Stecker, 2006; Stecker, Lembke, & Sáenz, 2007). One indication of broader PM impact is the recent establishment of a National Center for Student Progress Monitoring to evaluate assessment practices and materials for PM and to disseminate training materials (http://www.studentprogress.org). Familiarity with the practice of PM is expected for all educators with students at risk, as RtI is a general education as much as a special education process.

There is plenty of disagreement about whether PM data should (Gresham, 2002; Restori, Gresham, & Cook, 2008) or should not (Hale, Naglieri, Kaufman, & Kavale, 2004; Reynolds & Shaywitz, 2009; Wodrich & Schmitt, 2006) serve as the sole or primary basis for specific learning disabilities (SLD) special education referral or eligibility, and these issues are not limited to the field of learning disabilities. Three-tier models of behavioral interventions are implemented in schools across the United States, and PM data are used to judge both the effectiveness of programming and to determine the presence or absence of a disability. Whether sole or in tandem with other assessments, the use of PM data in decision making is of major concern if the measurement properties of the data are not clearly understood. Practitioners may be adopting a practice without full knowledge of the measurement issues or complications in its use. The design and implementation of PM might need to be considered in the context of the level of decisions, and this is not a simple process to be undertaken at the classroom level, perhaps not even at the school level.

Some consider PM evidence to be equal or superior to an individual psychoeducational battery (Bolt, 2005; Brown-Chidsey & Steege, 2005; Holdnack & Weiss, 2006; Restori et al., 2008). While the primary discourse is found in academics, the field of EBD would be wise to evaluate the discussion for two reasons: (a) Students with EBD have some of the most problematic academic performances of any students with a disability, and (b) the rhetoric and processes occurring in the related field of SLD are likely to cross over into EBD. Recent research conducted on oral reading fluency's "words correct per minute" has examined the psychometric soundness of PM data for making decisions at the individual student level (Christ, 2006; Christ & Ardoin, 2009; Francis et al., 2008; Jenkins, Zumeta, Dupree, & Johnson, 2005; Pŏncy, Skinner, & Axtell, 2005). This recent research stands in contrast to earlier CBM research (Deno, Fuchs, Marston, & Shin, 2001; Fuchs et al., 2004; Good & Jefferson, 1998; Marston, 1989), which projected classical test theory (CTT) estimates of alternate form reliability into progress monitoring applications, inferring greater psychometric strength than what actually existed. Likewise, the measurement properties of PM data in behavior are relatively unknown.

The fulcrum of PM data as an adequate measure for high-stakes decisions hinges in part on the equivalence and sensitivity of sets of similar tests or probes. But the judgment of adequacy will also depend on how PM data are summarized. Are summations of strength and precision available from the PM data? "Lower-stakes decisions," such as minor instructional and management adjustments within a classroom, have a lower threshold requirement for "precision"; therefore, access to the data summaries requires efficiency and accessibility for the teacher that may be more important than defensibility. However, for high- and even medium-stakes decisions, precision and accuracy are key standards, translated for each decision to the known likelihood of making an error. The identification of EBD or defensibility of instructional and management practices related to free and appropriate public education and least restrictive environment are decisions that require more knowledge about the data used in those decisions.

One possible option for strength and precision while maintaining accessibility and efficiency is the method employed to summarize slope and trend consistency (trendedness). These methods have not been well examined. The purpose of this paper is to critically examine two related PM summaries: "improvement rate," or slope of a trend line, and "trend consistency," or trendedness. The paper examines the most common summaries: linear regression's slope (ordinary least squares) and its R^2 trendedness counterpart method and the hand-fit method (Tukey tri-split slope); each is then compared to the relatively unknown Theil–Sen slope and its related Kendall's tau trendedness index.

Advantages of Progress Monitoring

Progress monitoring occurs repeatedly over time, yielding time series data, which provide a dynamic summary of learning or growth, notably trendline slope, or rate of improvement over time. A dynamic progress summary has at least five advantages over a single test score or two-shot, pre/post assessments for all students, but particularly for students with and at risk of EBD. These five advantages include more timely evaluation feedback because of its frequency; incremental judgments on student progress, updated regularly; (equivalent) comprehensive progress monitoring probes that measure both new learning and maintenance of earlier learning; goal-setting and a graphic display of monitoring progress toward that goal; and a dynamic rate of improvement summary index with proven predictive strength in research, adding value to predictions from static performance level alone (Ardoin & Christ, 2009; Christ & Coolong-Chaffin, 2007).

The first four of these five advantages (frequent incremental judgments, comprehensiveness, final expectations and graphic display, and improvement rate) are integral to the technique (inherent in the practice of PM). However, they do not reflect how *well* PM does these things. How accurate are the periodic judgments of progress, the goal achievement summaries, and the computed rates of improvement from the trendline slope? The answer to these and similar questions depends on quality of measurement, quality of the statistical summary, and accuracy of its interpretation. Only recently have these issues of quality been addressed for individual student decisions from PM data (Christ & Coolong-Chaffin, 2007).

This paper contributes to the literature by raising new questions about the statistical summary of PM data, which could be very valuable to our work in the field of EBD. Three trendline slopes—a hand-fit slope (Tukey's tri-split median-based slope; Tukey, 1977), the linear regression (LR) slope, and the Theil–Sen (or Kendall's slope; (Sen, 1968; Theil, 1950)—are compared for use in evaluating PM data. Neither Koenig's (1972) nor White's (1972, 1974) hand-fit, "quarter intersect" or "split middle" slopes are included, though they have served special educators well for over 30 years because they are surpassed in precision by Tukey's tri-split slope (Johnstone & Velleman, 1985; Parker, Stein, & Tindal, 1992), which can be hand-drawn onto graphed data, as can the LR slope and Theil–Sen or Kendall's slope.

Introduction and Illustration of Three Slopes

Pearson's R^2 tells the percentage of score variance explained by linear time, whereas Kendall's tau tells the percent of scores which improve over time. The LR trend line is iteratively fit to data to minimize the distances (squared) of all data points from the line. The LR slope ("b") is calculated as rise/run. The Tukey line is drawn from intersects of the median time and median score values in the first and last third of the data series. The Tukey slope can be calculated directly from score medians or from the drawn line (as rise/run). An optional variation of the Tukey line is its adjustment up or down to split the data into two equal parts. This variation changes its y-intercept, but not its slope. The Theil–Sen slope is usually computer calculated. The process for Theil–Sen begins with computing the "mini-slopes" of all possible pairs of data points in the time series. The quantity of these mini-slopes equals the number of pair-wise data comparisons, calculated as (N × [N − 1] / 2). So for eight data points, there are (8 × 7) / 2 = 28 mini-slopes. The median value of these mini-slopes is the Theil–Sen slope.

These three slope techniques are applied to four graphs in Figure 1a–d as a demonstration. These graphs were chosen to demonstrate typical differences and similarities among the slopes. They are from three articles in leading journals (Fantuzzo, Polite, & Grayson, 1990; Kern & Bambara, 2002; Swanson, Kozleski, & Stegink, 1987). Their slopes and trendedness indices are given in Table 1.

The first graph (Figure 1a) presents data with a visually apparent trend, nearly linear except for local high variability around Time three to five. The three quite similar trendline slopes (points improved per week) range from 9.21 to 12.84. The graph does not show high "bounce," or outliers at the ends of the trend line where their influence on slope would be greatest. The similarity of LR and Theil–Sen slopes here is no accident; as seen later, they correlate closely over a large sample. The two trendedness indices, R^2 and tau, are quite close, despite their different algorithms. Pearson's R^2 tells the percentage of score variance explained by linear time, whereas Kendall's tau tells the percent of Scores which improve over time. R^2 is distance-based (measured from a trend line), whereas tau is based solely on rank order (not involving any line). The Theil–Sen trend line is an "add-on," not part of the calculation of tau.

Figure 1b shows a less clear straight line data pattern, perhaps even a curve. Or perhaps the data show two different slopes for the halves of the series. Complex patterns such as these are not unusual. They are "ill-conforming" and thus violate the LR assumptions of linearity and constant variance, but data like these are no problem for tau.

Figure 1b shows all three slopes as negative, but Theil–Sen and Tukey have a ratio of almost 1.6:1. Tukey is heavily influenced by the two first extreme scores because its algorithm is pegged to the median of the first three data-points. The Theil–Sen is not so bound, and its results are quite close to the LR slope. The LR slope is slightly steeper due to

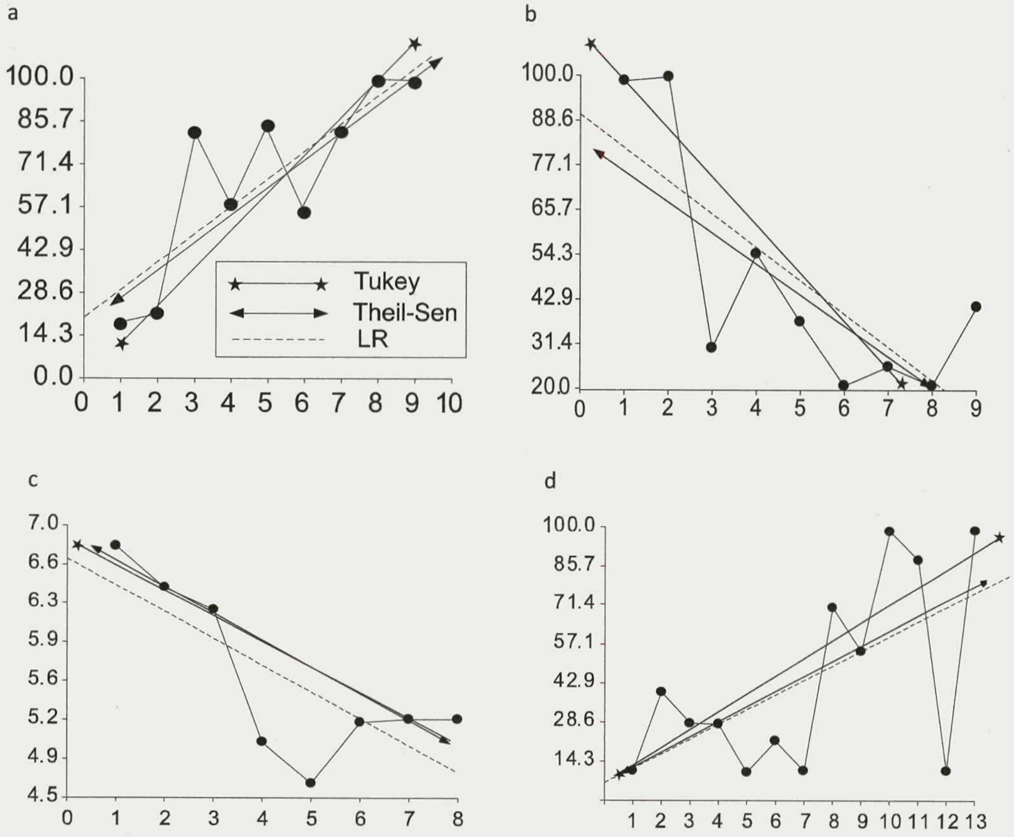

Figure 1. Graphs to demonstrate differences and similarities among three slopes.

the "pull" of the first two data points. LR slopes are heavily influenced by outlier scores at the beginning or end of a series. Median-based slopes, such as Theil–Sen and Tukey, are more robust or resistant to outliers than an LR slope, but both can fall victim to the inflexibility of a median in a small *N* series. Inflexibility with small *N* data is worst for Tukey, better for Theil–Sen.

Figure 1c shows even closer alignment of the three slopes; they are nearly equal. The LR slope is set lower, with a smaller y-intercept due to the influence of the most extreme scores at Times 4 and 5. When located at the middle of the data series, outliers have little effect on slope. Unlike slopes, the trendedness indices differ substantially, by a ratio of 1.6:1. Slopes and their comparisons convey different information than do trendedness indices. The lower tau is a signal that the straight line is not a good representation of improvement in about half of the data. A "local" (around Times 4 to 8)

TABLE 1
Slopes and Trendedness Indices for Four Published Data Series

		Slopes			Trendedness Indices	
	N	LR	Theil–Sen	Tukey	R^2	tau
Figure 1a	9	9.21	9.24	12.84	.69	.67
Figure 1b	9	−8.50	−7.77	−12.17	−.56	−.48
Figure 1c	8	−.25	−.25	−.24	−.59	−.49
Figure 1d	13	5.35	5.58	7.27	.37	.35

Notes. R^2 values cannot be negative. The signs were added for comparability with tau.

reversed or deterioration trend exists in the data, which is not captured by the Pearson R^2. The quality of an LR slope and R^2 summaries depend upon a straight line relationship, and visual analysis (at minimum) is needed to see whether a trend exists or not.

The fourth dataset, Figure 1d, was chosen to show typical variability in a longer series (13 data points). Tukey slopes become more realistic with this *N*, as the medians (for the first and last thirds of the data) have more flexibility, each based on four data points. Though the three slopes are similar, the Theil–Sen and LR slopes show slightly more attraction to the outlier at Time 12 than does the Tukey slope. The most noticeable difference, however, is between R^2 and tau, a size ratio of over 1.7:1. R^2 is based on the fit of a straight line, whereas tau considers improvements (and deterioration) in shorter local segments as well. Tau is sensitive to the generally downward data pattern visible in the first half of the data, around Times 2 to 7. Also, tau reflects only the order of data-points, whereas R^2 is sensitive to their distances.

This study is a first application of the Theil–Sen slope to PM data from individual students. Theil–Sen is compared to the strongest hand-fit method, Tukey slope, and the most published LR slope. Theil–Sen and Tukey were compared for precision and robustness by Johnstone and Velleman (1985); however, that was with Monte Carlo data not reflective of typical PM data, which uses a variety of scale types and series lengths. School PM data are unique and are not well reflected in data from the stock market, glacier retreats, or postsurgery recovery. For that reason, a field trial of three slopes (Tukey, LR, and Theil–Sen) and two indices of trend consistency (Pearson's R^2 and Kendall's tau) was undertaken with typical PM data. Specific research questions include: (a) How do R^2 and tau compare in size, and at what level are they related? and (b) How well do Theil–Sen and Tukey slopes match the LR slope in size?

Method

The field trial was on a sample of 372 data series during interventions, collected from journals in special education, school psychology, and applied behavioral psychology. Most interventions were in school settings with students, pre-K through Grade 12. Interventions represented a roughly even mix of academic and nonacademic (social and functional) behaviors. All types of disabilities were represented, along with low achievers and students at risk. The great majority of the intervention data series in this study were preceded by a prior baseline monitoring period, which was not relevant for these analyses.

This convenience sample was collected in stages over multiple years, beginning with PsychLit and ERIC searches for any articles with graphed single-client monitoring data. The original sample was augmented to broaden the sample, include new publications, specifically increase the number of studies that used rating scales, and increase the number of *Journal of Applied Behavior Analysis* articles. Article graphs were scanned at high resolution, digitized using I-Extractor software (Linden, 1998), and replotted from reconstituted scores saved to a spreadsheet. The reliable data capturing process is described in detail in an earlier article (Parker, Cryer, & Byrns, 2006). The final sample was composed of 372 data series from 82 different articles found in 29 different journals.

For the 372 data series, LR slopes were plotted and R^2 was calculated within the Number Cruncher Statistical System linear regression module, which offers bootstrap inference tests. Kendall's tau trendedness indices and the Theil–Sen slope were obtained from StatsDirect (StatsDirect, Ltd., London, England). The Theil–Sen output was cross-checked with the U.S. government's KTRLine output, which matched perfectly. StatsDirect output includes exact estimates of standard errors, *p* values, and confidence intervals for Theil–Sen and tau. For the Tukey slope, the online Mood Median inference test (Mood, 1950) was used, which includes Fisher exact *p* values.

Results

This study applies the Theil–Sen slope to existing PM data to see how it performs in comparison to other more commonly used methods. Specifically we were interested in how tau compares in size to R^2 and how closely they are related, As well as in a comparison of Theil–Sen, Tukey, and LR slope sizes.

Anticipating typical sizes for a trend index is expected to increase its usability. Users have that sense about Pearson R^2, the ubiquitous standard. We know less about tau—only that it ranges from −1 to +1. Table 2 presents the quartile distributions of R^2 and tau for the sample of $N = 374$ intervention data series.

TABLE 2
Quartile Distributions of R^2 and tau Across

	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Actual Values					
R^2	−.618	−.262	−.000	.272	.702
tau	−.672	−.400	−.080	.349	.702
Absolute Values					
R^2 abs	.004	.042	.255	.583	.810
tau abs	.000	.087	.320	.576	.779

Notes. R^2 cannot be negative. Negative signs were added to R^2 after calculations to permit direct comparison with tau.

The top half of Table 1 contains actual R^2 and tau values, but maintains the negative signs from R to R^2. The bottom half of the table contains only absolute values, discarding negative signs. Both halves show tau and R^2 to be similar in size, differing by less than .10 over most of the score distribution (median difference .07). Recall the somewhat similar interpretations of R^2 as percentage of variance in scores accounted for by time'' and tau as percentage of scores which improve over time. For series with the most trendedness (90th percentile), R^2 and tau are equal, but tau values are relatively larger below the 50th percentile. Finally, R^2 and tau correlated at a strong $R = .97$ level, despite their very different assumptions, theoretical foundations, and computational procedures.

Since the 372 slopes came from different scales with different ranges, means, and standard deviations, slope comparisons are problematic. Instead, Theil–Sen and Tukey slopes were each compared on the same dataset against the LR slope as the standard. Standardized deviation scores (from the LR slope) were created for Tukey and Theil–Sen slopes. The Tukey and Theil–Sen slopes were each subtracted from the LR slope, and the difference was divided by the LR slope, creating a standardized "percentage of LR slope difference" deviation score. Tukey slopes had median deviation scores (from LR slopes) of −.22, or 22% smaller than LR slopes. The IQR (middle half of scores) for Tukey slopes was −.91 to +1.05 deviation, or about 100% larger to 100% smaller than LR slopes. The Theil–Sen median deviation score was zero, and the IQR for the deviations ranged from −.43 to +.27. Thus, Tukey slopes deviated two to four times more from the LR slope than did Theil–Sen slopes. Only the most extreme deviation scores of Tukey and Theil–Sen (at the 10th and 90th percentiles) were similar. In summary, for the large majority (75%+) of datasets, Theil–Sen substantially surpassed Tukey in matching the LR slope .

In a second analysis, the three slopes were intercorrelated by rank (Spearman's rho), revealing strong similarities: LR with Theil–Sen: .94; LR slope with Tukey: .94; Theil–Sen with Tukey: .93. These results do not contradict the very large deviation score results immediately preceding, but do not. This analysis tested the rank-order similarity of the slopes, not their differences in size. Within a variable sample such as this, two slope methods can be rank ordered the same, yet differ greatly in size of slope coefficients.

In summary, the three slopes are highly intercorrelated and all at about the same level; nevertheless, there are large differences in the size of the trendline slope, especially between the Tukey and LR slopes. Theil–Sen differed much less from LR slope size than did Tukey, by a factor of one half to one fourth. Neither

TABLE 3
p-values for Three Trend Analyses at Quartile Markers from 372 Sample Datasets

	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
LR/*R*	.830	.575	.144	.004	.0001
Theil–Sen/tau	.860	.600	.172	.012	.0007
Tukey	1.000	.833	.569	.193	.0286

Tukey nor Theil–Sen showed a tendency toward extremely deviant slopes.

Discussion

This study takes place in the context of increased accountability for schools and the likelihood that PM data will play a central role in RtI for making medium-to-high stakes decisions. Useful instruments and procedures for PM have been developed over nearly four decades within applied behavior analysis, precision teaching, and curriculum-based measurement, but largely within special education, school psychology, and applied behavior analysis research, insulated from other measurement fields. Other fields occasionally produce and validate methods that show promise for school application. If attractive, they should be field tested with actual school data to ascertain their appropriateness for school purposes.

A new contender for summarizing PM data was compared to the two trendline slopes most commonly in schools' PM. Teachers are most likely to use the median-based hand-fit line, White's "split-middle," Koenig's "quarter-intersect," or the successor, the Tukey tri-split slope. Researchers are more likely to employ the ubiquitous LR slope "b," which is widely available in free software, even in teacher-friendly applications such as Chart-Dog (http://www.jimwrightonline.com). Between these two current choices was introduced a third option, the Theil–Sen slope. The Theil–Sen is "distribution free," so within high-stakes decision making it cannot be challenged for failing to meet parametric assumptions, and its successful track record outside of education is in high-stakes decision arenas. Another asset of Theil–Sen is its companion, Kendall's tau, conveniently interpreted as "the percentage of data that show improvement over time." Because Theil–Sen has apparently not been published on student performance data, PM practitioners have not been given a sense of how it performs with typical PM data compared with LR and Tukey summaries.

There would be less reason to consider a new trend technique were there not issues with the available options. The increasingly popular daily behavior rating scales and school rules monitoring may not be interval-level, which LR assumes. For low-stakes decisions only, these issues should cause little concern, but for PM data used to determine disability or service provision (high-stakes decisions), this is a large concern. Our trend analysis performed on 372 published datasets showed that nearly three-fourths failed one or more of the parametric assumptions of linearity, constant variance, and normality, according to the best tests available: Shapiro-Wilk (Shapiro & Wilk, 1965) and Modified Levene.

Applying Tukey, Theil–Sen, and LR slopes and R^2 and tau trendedness indices to a sizeable sample of published intervention data addresses other questions regarding their comparative sizes. Most R^2 and tau indices differed by around .07 points, and up at the 90th percentile (most consistent trends), they both equaled +.70 for positive trends, and at the 10th percentile, −.62 and −.67 for negative trends. This score similarity with R^2 should be reassuring. In addition, R^2 and tau rank-correlated at .97. In short, Theil–Sen appears similar to R^2 and Tukey by size and relationship.

The three slopes also rank-correlated in a tight cluster, from rho = .93 to .94. However, slope values differed widely, especially by Tukey. Most Tukey slopes (IQR) differed from the LR slopes in the range of ±100% of the LR slope. In contrast, most Theil–Sen slopes (IQR) differed from LR slopes by ±35% of the LR slopes (which is still substantial). So Theil–Sen showed onethird the deviation from the LR slope that Tukey did. This analysis also emphasizes the difficulty in comparing slopes and indicates that slope as a progress summary has limitations.

The large typical differences between Tukey and LR slopes may be of concern at present, as the field uses both slope estimation techniques concurrently. To our knowledge this is the first large application comparing the two indices. Teachers are trained to hand-calculate Tukey slopes, yet sometimes they are also trained with regression-fit trend lines, and the two results tend to differ. Also, teachers are encouraged to use LR slopes by user-friendly applications such as Chart-Dog and Microsoft Excel. This study highlights the danger of conflicting results from the two methods. Use of Theil–Sen should reduce the magnitude of differences with LR slope by two thirds.

Statistical power for Theil–Sen was 89% that of LR slope, using a rough approximation method, which was the number of datasets reaching $p \leq .05$. Though only a rough approximation, that result is comparable to Theil–Sen's published Pittman efficiency of 91%. Tukey power was substantially less, though it was cautioned that statistical power is of little importance when a convenient, teacher-friendly PM summary is used for routine classroom decisions.

Supplemental analyses were conducted with 124 shorter datasets (N = 6 to 10) to see whether findings differed from those obtained for the full dataset. Findings did not differ. With this short dataset, the prevalence of significant results at $p \leq .10$ was counted, finding 30% with LR, 27% with Theil–Sen, and 12% with Tukey. The results show that using the Theil–Sen or LR slope, inference testing can be fruitful for quite short datasets, as one third or more of them met the standard $p \leq .05$ test often used for moderate and higher stakes decisions.

A strength of this study is the relatively large sample of individual time series datasets. In addition, they were diverse, covering academic skills, functional skills and communication, and social behavior with a variety of at-risk and disability types. The breadth of coverage is a strength in that it helps ensure the generalizability of findings across scale types, and the diverse sample should have improved most correlation analyses.

Sampling together from very different scales is a study limitation. We cannot say from these results specifically how Theil–Sen, Tukey, and LR slopes perform with only a specific subset of measures. Rather, the datasets analyzed in this study were undifferentiated by social versus academic behaviors. To answer questions about Theil–Sen performance particularly for subsets of measures, further studies would be needed on those particular scales. One might assume greater power and sensitivity with all three slopes if scales were limited to academic growth (rather than social behavior performance). But that should be assumed only if probes are well spaced out in administration and over a longer time periods. Credible, reliable measured gains through PM require that probes be spaced out over a longer time period than is often practiced. In an ORF study by Ardoin and Christ (2009), high variability and nonsignificant progress resulted over eight weeks, even with carefully selected probes.

In medicine and the physical sciences, the Theil–Sen slope with its related tau index of trend consistency is a method of choice for making high-stakes decisions with time series data, proving itself in dozens of studies. This study is the first to our knowledge to apply Theil–Sen to school PM data. In this first field test, the new measure proved competitive with or superior to the two indices now commonly used. Theil–Sen measures similar attributes similar to those measured by LR and Tukey, and tau and Theil–Sen tie with LR summaries in matching visual analysis judgments of progress. Interpretation of Kendall's tau as "percentage of data showing improvement over time" is more direct and intuitive than Pearson's *R*. Finally, tau and Theil–Sen can be performed on freely downloadable, menu-driven software. These positive results and software availability argue for further examination of tau and Theil–Sen within narrower performance domains.

REFERENCES

Ardoin, S. P., & Christ, T. J. (2009). Curriculum based measurement of oral reading: Estimates of standard error when monitoring progress using alternate passage sets. *School Psychology Review, 38,* 266–283.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91–97.

Berkeley, S., Bender, W. N., Peaster, G. L., & Saunders, L. (2009). Implementation of Response to Intervention: A snapshot of progress. *Journal of Learning Disabilities, 42,* 85–95.

Bijou, S. W., & Baer, D. M. (1961). *Child development: A systematic and empirical theory.* New York: Appleton-Century-Crofts.

Bolger, H. (1965). The case study method. In B. B. Wolman (Ed.), *Handbook of clinical psychology.* New York: McGraw-Hill.

Bolt, S. (2005). Reflections on practice within the heartland problem-solving model: The perceived value of direct assessment of student needs. *The California School Psychologist, 10,* 65–80.

Brown-Chidsey, R., & Steege, M. W. (2005). *Response to intervention: Principles and strategies for effective practice.* New York: Guilford.

Christ, T. J. (2006). Short term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128–133.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47,* 55–75.

Christ, T. J., & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum, 1,* 75–86.

Deno, S., Fuchs, L. S., Marston, D. B., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30,* 507–524.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140.

Fantuzzo, J. W., Polite, K., & Grayson, N. (1990). An evaluation of reciprocal peer tutoring across elementary school settings. *Journal of School Psychology, 28*, 309–323.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315–342.

Fuchs, D., Deshler, D., & Reschly, D. (2004). National Research Center on Learning Disabilities: Multi-method approaches to the study of LD identification and classification. *Learning Disability Quarterly, 27*, 189–195.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In H. L. Swanson, K. R. Harris, & S. E. Graham (Eds.), *Handbook on learning disabilities* (pp. 431–449). New York: Guilford.

Gerber, M. (2005). Teachers are still the test: Limitations of response to intervention strategies for identifying children with learning disabilities. *Journal of Learning Disabilities, 38*(6), 516–524.

Good, R. H., & Jefferson, G. (1988). Contemporary perspectives on Curriculum Based measurement Validity. In M. R. Shinn (Ed.), Advanced applications of Curriculum-based Measurement (pp. 61–88). New York: Guilford.

Gresham, F. M. (2002). Teaching social skills to high-risk children and youth: Preventive and remedial strategies. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 403–432). Bethesda, MD: National Association of School Psychologists.

Hale, J. B. (2006). Differential medication response in ADHD subtypes. *US Special Populations Pediatrics Review*, 62–69.

Hale, J. B., Naglieri, J. A., Kaufman, A. S., & Kavale, K. A. (2004). Specific learning disability classification in the new Individuals with Disabilities Education Act: The danger of good ideas. *The School Psychologist, 58*(1), 6–14,

Hintze, J., & Stecker, P. (2006). *Data-based instructional decision making.* Retrieved from http://www.studentprogress.org/weblibrary.asp#data

Holdnack, J. A., & Weiss, L. G. (2006). IDEA 2004: Anticipated implications for clinical practice–Integrating assessment and intervention. *Psychology in the Schools, 43*, 871–892.

Jenkins, J. R., Zumeta, R., Dupree, O., & Johnson, K. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research and Practice, 20*(4), 245–253.

Johnstone, I., & Velleman, P. F. (1985). Efficient scores, variance decompositions, and Monte Carlo swindles. *Journal of the American Statistical Association, 80*, 851–862.

Kavale, K. A., & Forness, S. R. (2000). What definitions of learning disability say or don't say: A critical analysis. *Journal of Learning Disabilities, 33*(3), 239–256.

Kavale, K. A., Holdnack, J. A., & Mostert, M. P. (2006). Responsiveness to intervention and the identification of specific learning disability: A critique and alternative proposal. *Learning and Disability Research & Practice, 29*, 113–127.

Kern, L., & Bambara, L. (2002). Class-wide curricular modification to improve the behavior of students with emotional or behavioral disorders. *Behavior Disorders, 27*(4), 317–326.

Koenig, C. (1972). *Charting the future course of behavior.* Kansas City, KS: Precision Media.

Marston, D. B. (1989). A Curriculum-based Measurement approach to assessing academic performance: Wate it is and why to do it. In M. Shinn (Ed). Curriculum-based Measurement Assessing Special Children. New York: Guildford.

Mood, A. M. (1950). *Introduction to the theory of statistics.* New York: McGraw-Hill.

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *American Statistician, 4*, 19–21,

Naglieri, J. A., & Crockett, D. P. (2005). Response to intervention (RTI): Is it a scientifically proven method? *Communique, 34*(2), 38–39.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single case research. *School Psychology Quarterly, 21*(4), 418–444,

Parker, R., Stein, M., & Tindal, G. (1992). Estimating trend in progress monitoring data: A comparison of simple line-fitting methods. *School Psychology Review, 2*, 300–313.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum based measurement. *Journal of Psychoeducational Assessment, 23*, 226–238.

Restori, A. F., Gresham, F. M., & Cook, C. R. (2008). Old habits die hard: Past and current issues pertaining to response-to-intervention. *The California School Psychologist, 13*, 67–79.

Reynolds, C. R., & Shaywitz, S. E. (2009). Response to intervention: Ready or not? Or, from wait-to-fail to watch-them-fail. *School Psychology Quarterly, 24*(2), 130–145.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *American Statistical Association Journal, 63*, 1379–1389.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611.

Sidman, M. (1960). *Tactics for scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

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

Skinner, B. F. (1966). Science and human behavior. New York: McMillian.

Stecker, P. M., Lembke, E. S., & Sáenz, L. (2007). *Advanced application of CBM in reading: Instructional decision-making strategies.* Presentation at 2007 Summer Institute on Student Progress Monitoring, Nashville, TN. Retrieved from http://www.studentprogress.org/weblibrary.asp#data

Swanson, H. L., Kozleski, E., & Stegink, P. (1987). Disabled readers' processing of prose: Do any processes change because of intervention? *Psychology in the Schools, 24,* 378–384.

Theil, H. (1950). *A rank-invariant method of linear and polynomial regression analysis, III.* Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen A, *53,* 1397–1412.

Tukey, J. W. (1977). Straightening out plots (using three points). In *Exploratory data analysis* (pp. 169–181). Menlo Park, CA: Addison-Wesley.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics, 11,* 284–300. U.S. Office of Special Education Programs. National Center on Student Progress Monitoring http://www.studentprogress.org/weblibrary.asp dowloaded August 14, 2012.

White, O. R. (1972). *The prediction of human performances in the simple case: An example of four techniques.* (Working Paper No. 15), University of Oregon Regional Resource Center for Handicapped Children, Eugene, OR.

White, O. R. (1974). *The split middle: A quickie method of trend estimation* (3rd revision). Unpublished manuscript, University of Washington, Experimental Education Unit, Child Development and Mental Retardation Center, Seattle, WA.

Wodrich, D. L., & Schmitt, A. J. (2006). *Patterns of learning disabilities.* New York: Guilford Press.

AUTHORS' NOTE

Acknowledgments to Randy Morton, M.D. for assistance in the preparation of this manuscript.

Address correspondence to Dr. Kimberly Vannest, MS 4225, College Station, TX 77843-4225; E-mail: kvannest@tamu.edu.

MANUSCRIPT

Initial Acceptance: 3/19/12
Final Acceptance: 6/12/12