Detecting Earnings Management Using Discontinuity Evidence*

David Burgstahler  
Julius A. Roller Professor of Accounting  
University of Washington/Seattle

Elizabeth Chuk  
Assistant Professor  
University of Southern California

September 30, 2014

Abstract

Evidence of earnings management based on discontinuities in earnings distributions at prominent benchmarks is pervasive, yet there is little evidence of earnings management based on tests for abnormal accruals. Together with previous results on the characteristics of accruals-based tests, results in this paper on the statistical characteristics of discontinuity tests reconcile this apparent inconsistency. The distribution of the standardized difference statistic is derived, relaxing the Burgstahler and Dichev (1997) assumption that the two components in the numerator of the standardized difference are independent and simulation results confirm that the "correction" introduced into the literature by Beaver, McNichols, and Nelson (2007) is, in fact, incorrect. Simulation results also suggest that the independence assumption is not likely to induce important errors in typical applications. The power of the standardized difference test is evaluated for alternative assumptions about the pattern of earnings management to meet benchmarks. In contrast to results from previous research that suggest accruals tests have reasonable power only in samples where there are high rates and large amounts of accruals management, discontinuity tests have the power to detect far smaller amounts and much lower rates of total earnings management to meet a benchmark.

Preliminary and incomplete draft prepared for discussion at UTS. Comments invited but please do not quote, cite, or distribute without permission

JEL classification:  G14, M40

Key Words:  Earnings management, discontinuities

* We appreciate helpful comments and suggestions from Russell Lundholm, Norman Strong, and seminar participants at the University of British Columbia and the University of Illinois.
1. Introduction

There is pervasive evidence of discontinuities in distributions of reported earnings at prominent earnings benchmarks, where distributions comprise fewer observations immediately below the benchmark and more observations immediately above the benchmark than would be expected if the distribution were smooth.\(^1\) The body of evidence is consistent with the theory that managers take actions to ensure that earnings meet benchmarks, e.g., earnings are managed to avoid small losses, small earnings decreases, and small negative earnings surprises.\(^2\) This interpretation is further supported by survey evidence in Graham, Harvey, and Rajgopal (2005) indicating that managers are willing to incur real costs in order to meet benchmarks. However, there has been no formal evaluation of the statistical properties of discontinuity tests to detect earnings management.\(^3\)

Inferences from empirical evidence depend on an understanding of the determinants of the size and power of the statistical tests employed. If a test is sensitive to violations of assumptions, a significant test statistic might be attributable to violations of assumptions rather than to a false null hypothesis. On the other hand, if a test has low power, insignificant results might be attributable to low power rather than to a true null hypothesis. Further, significant results from a test with low power lead to little or no revision in beliefs.\(^4\)

---

\(^1\) The smoothness assumption is discussed in more detail in Section 2.

\(^2\) For example, Burgstahler and Dichev (1997, hereafter BD) show that distributions of earnings levels and distributions of earnings changes exhibit discontinuities at zero and there is widespread evidence of discontinuities in distributions of earnings surprises, e.g., DeGeorge, Patel, and Zeckhauser (1999), Brown (2001), Matsumoto (2002), Brown and Caylor (2005), Burgstahler and Eames (2006). There is also evidence of discontinuities at prominent non-zero earnings benchmarks, e.g., Carslaw (1988), Thomas (1989), Das and Zhang (2003), Grundfest and Malenko (2009), and Burgstahler and Chuk (2013).

\(^3\) In contrast, the power of methods of detecting earnings management based on measures of abnormal accruals is evaluated in Dechow, Sloan, and Sweeney (1995), Dechow et al (2010), and Ecker et al (2011).

\(^4\) See Burgstahler (1987) for further discussion of the roles of size and power in inference from empirical results.
The purpose of this paper is to (i) refine the derivation of the distribution of the Burgstahler and Dichev (1997, hereafter BD) standardized difference statistic under the null (no earnings management) hypothesis, relaxing the simplifying assumption that the two numerator components of the test statistic are independent, (ii) derive the distribution of the standardized difference statistic under the alternative hypothesis, and (iii) analyze the determinants of the power of standardized difference tests under various versions of the alternative (earnings management) hypothesis. The derivation, verified using simulation results, shows the independence assumption typically has only a minor effect on empirical assessments of significance and also confirms that the "correction" to the variance of the standardized difference statistic published in Beaver, McNichols, and Nelson (2007, hereafter BMN) is erroneous. Specifically, the BMN “correction” results in a substantial overstatement of the variance of the test statistic and a corresponding substantial understatement of significance.

The power analysis suggests the standardized difference test has power to detect management of relatively small amounts of earnings by even a small proportion (such as .25–.50% of MVE for .1-.2%) of sample firms. In contrast, previous research suggests that accruals-based tests have reasonable power only for much larger amounts and much higher rates of earnings management (such as 5% or more of total assets for 100% of firms identified as potential earnings managers). Together with evidence from previous papers, the analysis here suggests that discontinuity tests have far greater power to detect much less pervasive, and much smaller amounts of, earnings management than do tests based on abnormal accruals. Thus, these results reconcile what has sometimes been interpreted as an inconsistency in the literature,

\[\text{\textsuperscript{5}}\text{ As explained further below, .1%-2% is the proportion of firms managing earnings when, for example, 1% of the population has pre-managed earnings a small amount below a benchmark and 10%-20% of firms with earnings a small amount below a benchmark manage earnings upwards to meet the benchmark.}\]
namely common and widespread evidence of earnings management using discontinuity evidence yet little supporting evidence of earnings management from abnormal accruals models.

The paper is organized as follows. Section 2 provides background. Section 3 reviews and refines the theoretical derivation of the distribution of the standardized difference statistic. Section 4 provides simulation evidence to confirm the derivation. Section 5 analyzes the determinants of power of the standardized difference test and discusses implications for research design. Section 6 concludes.

2. Background

Earnings management occurs when the perceived benefits of managing earnings (including both benefits to the firm and potentially disparate benefits to the manager) exceed the costs. Healy and Wahlen (1999) note: "Despite the popular wisdom that earnings management exists, it has been remarkably difficult for researchers to convincingly document it." For example, there are relatively few examples of evidence of significant abnormal accruals in situations where earnings management behavior is hypothesized. Further, Ball (2013) outlines a number of reasons to be skeptical about significant abnormal accruals results that imply that earnings management of “enormous amounts” based on accruals is “rife.” In contrast, there is widespread evidence of discontinuities in earnings distributions at prominent benchmarks, consistent with earnings management.6

2.1 Evidence of discontinuities

Earnings management to meet a benchmark transforms some pre-managed earnings observations below the benchmark into reported (i.e., post-managed) observations above the

6 Two recent papers by Durtschi and Easton (2005 and 2009) purport to show that the discontinuities in earnings distributions are explained by “deflation, sample selection, and a difference between the characteristics of profit and loss observations.” Burgstahler and Chuk (2013) show that the Durtschi and Easton "explanations" for discontinuities are erroneous.
benchmark. The result is a decrease in the frequency of earnings observations below the benchmark and an increase in the frequency above the benchmark, creating a discontinuity at the benchmark in the distribution of reported (post-managed) earnings.\(^8\)

The set of alternative methods of earnings management is broad. To meet a benchmark, managers might enter into real operating, investing, or financing transactions that generate additional current profits, where these transactions might be purely incremental with no effects on future profitability or they might have adverse effects on future profitability. In addition, managers might make accounting choices to increase current reported profits. For example, managers might accelerate recognition of revenues or defer recognition of expenses as permitted within GAAP. Managers might record transactions in a way that violates GAAP or record fictitious or fraudulent transactions. Any combination of these actions can be used to manage earnings to meet the benchmark, and managers are expected to choose the lowest-cost combination. Because reported earnings reflect all of these actions, discontinuities provide evidence that some combination of these actions has been taken to meet the benchmark, but do not reveal which actions were taken.\(^10\)

Multiple studies document discontinuities around alternative earnings benchmarks, including the profit/loss benchmark (Hayn, 1995; BD, 1997; Degeorge et al., 1999); prior-year

---

\(^7\) Researchers have examined incentives to manage earnings upward to create the appearance of better current performance, or downward to create the appearance of worse current performance or to transfer earnings to future periods from the current period. This paper focuses on upward management to increase current performance, the phenomenon that creates discontinuities at current earnings benchmarks.

\(^8\) Kerstein and Rai (2007) provide direct evidence that earnings management converts observations below the benchmark into observations above the benchmark.

\(^9\) See Jensen (2001, 2003) for multiple examples of transactions and decisions that increase current earnings but decrease expected future earnings and firm value.

\(^10\) Schipper (1989) defines "disclosure management" as "purposeful intervention in the external financial reporting process, with the intent of obtaining some private gain…" (p. 92) but also notes that there are additional forms of "real" earnings management. i.e. real operating, investing, or financing decisions that alter reported earnings. Healy and Whalen (1999) use a broad definition that encompasses a broad range of actions: "Earnings management occurs when managers use judgment in financial reporting and in structuring transactions to alter financial reports to either mislead some stakeholders about the underlying economic performance of the company or to influence contractual outcomes that depend on reported accounting numbers." (p. 368)
earnings (BD, 1997; Degeorge et al., 1999; Beatty et al., 2002; Donelson et al., 2013); and analyst forecasts (Degeorge et al., 1999; Burgstahler and Eames, 2006; Donelson et al., 2013). Leuz, Nanda, and Wysocki (2003) report evidence of loss avoidance for an international sample of firms from 31 countries. Suda and Shuto (2005) provide strong evidence that Japanese firms engage in earnings management to avoid earnings decreases and losses. Haw et al. (2005) document discontinuities for the return on equity (ROE), where there is an unusually high proportion of firms reporting ROEs slightly greater than the 10% benchmark. Daske, Gebhardt, and McLeay (2006) report even more pronounced discontinuities in the European Union than in the US for the profit/loss, prior year earnings, and analyst forecast benchmarks.

2.2 Evidence of abnormal accruals

Accruals tests identify as abnormal the effects of any transactions or events where accruals deviate significantly from a model of "normal" accruals. Thus, abnormal accrual models are not designed to detect earnings management via methods other than accruals. Further, abnormal accrual models will not detect earnings management structured to fit the model of normal accruals, such as fraudulent transactions structured so that the accruals fit within the "normal" range. Also, as discussed further below, accrual models may incorrectly identify other unusual circumstances, such as unusual growth or other forms of extreme financial performance, as "abnormal accruals."

Previous research suggests that in order for accruals-based tests to have reasonable power, earnings management amounts must be large and earnings management must be pervasive among sample firms. Dechow, Sloan, and Sweeney (1995) show that the power of accrual-based models for detecting earnings management is low even for relatively large amounts (1 to 5 percent of total assets) of earnings management by 100% of sample firms. More recently, Dechow et al (2010, p. 24-25) develop a test with greater, but still relatively low,
power. Similarly, Ecker et al (2011) examine the effect of peer firm selection on the power of
discretionary accruals models to detect earnings management and show that peer selection based
on lagged assets provides higher power than selection based on industry. However, the power to
detect moderate to large amounts of earnings management remains relatively low, even for high
rates and relatively large amounts of earnings management. For example, the Ecker et al (2011)
Figure 1 summary shows that for accruals management equal to 4% (10%) of total assets for
100% of sample firms, the power of tests conducted at the .05 level of significance is generally
on the order of .10 (.25).\textsuperscript{11}

While the magnitude of accruals management required to achieve even modest power is
clearly large relative to total assets, the required magnitude is even more striking relative to
expected earnings. For example, assume long-run expected earnings is on average on the order
of 10% of book value of equity. For a firm with no debt in the capital structure where equity is
equal to total assets, the Ecker et al estimates suggest the power of tests based on discretionary
accruals is only .10 (.25) when abnormal accruals average 40% (100%) of long-run expected
earnings. For a firm with 50% debt in the capital structure where equity is equal to one-half of
total assets, the power is only .10 (.25) when abnormal accruals average 80% (200%) of expected
earnings.\textsuperscript{12}

In addition to the limitations of low power and narrow focus on one specific method of
earnings management, accruals-based approaches are subject to substantial risk of spurious
significance due to factors other than earnings management. Accruals tests compare estimates of
pre-managed accruals with reported accruals for each individual observation. The estimates of

\textsuperscript{11} Stubben (2010) develops a test for abnormal revenue that appears to have somewhat greater power than accruals-
based tests, though the test by design detects only revenue-based earnings management.
\textsuperscript{12} Consistent with Ball (2013), this numerical example illustrates why evidence of significant abnormal accruals
implies earnings management of “enormous amounts.”
pre-managed earnings are prone to both estimation error and bias, so abnormal accruals are subject to estimation error and bias. For example, Dechow, Sloan, and Sweeney (1995) show that accruals-based models “reject the null hypothesis of no earnings management at rates exceeding the specified test-levels when applied to samples of firms with extreme financial performance.” Similarly, Collins and Hribar (2002) show that the error in the balance-sheet approach to estimating accruals is correlated with economic characteristics of the firm, which can lead to a spurious relation between estimates of accruals management and the correlated economic characteristics.

2.3 Reconciling discontinuity and abnormal accruals evidence

In several settings where discontinuity tests have identified significant discontinuities, abnormal accruals tests have not identified significant abnormal accruals. For example, Dechow, Richardson, and Tuna (2003) are unable to find any evidence that boosting of discretionary accruals is the key driver of the discontinuity at zero earnings. Similarly, Ayers et al (2006) find little or no evidence of abnormal accruals among firms with earnings that fall just above the benchmark. One possible explanation for these seemingly inconsistent results is the broader scope of discontinuity tests (which reflect all methods of earnings management) versus accruals tests (which reflect only earnings management via accruals manipulation). A second possible explanation is the low power of accruals-based tests shown in previous research coupled with higher power of discontinuity tests. However, to date there has been no formal evaluation of the power of discontinuity tests. Section 5 below provides this evaluation, showing that discontinuity tests have the ability to detect much lower rates and much smaller amounts of earnings management than do abnormal accruals tests.

13 See McNichols (2000).
14 See also Kothari, Leone, and Wasley (2005) and Dechow et al (2010).
3. Distribution of the standardized difference test statistic

The standardized difference statistic defined in BD is designed to detect effects of earnings management to meet a benchmark on the distribution of earnings. The name is derived from the fact that the statistic for interval i is the difference between the number of observations in interval i and the average of the numbers in the two adjacent intervals standardized by the approximate standard deviation of the numerator difference. When there is no management in the vicinity of interval i, the expected difference between the number of observations in interval i and the average in the two adjacent intervals is approximately zero, and the expectation of the standardized difference is approximately zero. On the other hand, management in the vicinity of interval i will create a non-zero expectation. Management that transforms observations into interval i from other intervals creates an excess of observations in interval i relative to the average in the adjacent intervals, leading to a positive standardized difference. Management that transforms observations from interval i into other intervals creates a dearth of observations in interval i, leading to a negative standardized difference.

Denote the number of observations in the entire distribution by N, the number of observations in interval i by n_i and the number of observations in the intervals immediately above and below interval i by n_{i-1} and n_{i+1}. Similarly, denote the probability of an observation in interval i by p_i and the probabilities of an observation in the intervals immediately below and above by p_{i-1} and p_{i+1}, respectively.

The numerator of the standardized difference is a function of two multinomial random variables, the observed frequency in interval i, n_i, and the sum of the frequencies in the intervals immediately above and below interval i, n_{i-1} and n_{i+1}.

15 The standardized difference statistic has also been used to test for effects of upward management of variables other than earnings (see, for example, Dichev and Skinner 2002 or Dyreng, Mayew, and Schipper 2012) and to test for downward management of size variables (see, Bernard, Burgstahler, and Kaya 2013). However, to simplify the exposition, the examples and discussion in this paper focus specifically on upward management of earnings.
immediately below and above interval i, \( n_{i-1} + n_{i+1} \). Thus, the numerator can be interpreted as the sum of the first multinomial variable and the second variable after is has been multiplied by \(-1/2\) (to subtract the average of \( n_{i-1} + n_{i+1} \)). The marginal distributions for the two terms of the numerator sum are binomial, with respective expectations \( Np_i \) and \(-1/2N(p_{i-1}+p_{i+1})\), variances \( Np_i(1-p_i) \) and \( 1/4N(p_{i-1}+p_{i+1})(1-p_{i-1}-p_{i+1}) \) and covariance \((-1/2Np_i (p_{i-1}+p_{i+1}))\).

Define \( \delta \) as the difference between the probability for interval i and the average of the probabilities for intervals i–1 and i+1,

\[
\delta = p_i - \frac{1}{2} (p_{i-1}+p_{i+1}).
\]

Thus, \( \delta \) is positive when interval i is a local peak (i.e., when the probability in interval i is greater than the average of the probabilities in the two adjacent intervals) and \( \delta \) is negative when interval i is a local trough.

Using the definition of \( \delta \), the expectation of the numerator of the standardized difference is

\[
E[n_i - \frac{1}{2} (n_{i-1} + n_{i+1})] = N[p_i - \frac{1}{2} (p_{i-1}+p_{i+1})] = N \delta.
\]

The variance of the numerator is the sum of the variances of the numerator components plus twice the covariance:

\[
V(\text{numerator}) = Np_i(1-p_i) + 1/4N(p_{i-1}+p_{i+1})(1-p_{i-1}-p_{i+1}) + 2 (-1/2) \{ -Np_i (p_{i-1}+p_{i+1}) \} = Np_i(1-p_i) + 1/4N(p_{i-1}+p_{i+1})(1-p_{i-1}-p_{i+1}) + Np_i (p_{i-1}+p_{i+1}).
\]

Because \( (p_{i-1}+p_{i+1}) \equiv 2 (p_i - \delta) \), the variance in (3) can also be written as

\[
V(\text{numerator}) = Np_i(1-p_i) + 1/2N(p_i-\delta)(1-2p_i+2\delta) + 2Np_i (p_i-\delta)
\]

---

16 See, for example, Johnson and Kotz (1969, p. 281 and 284).

17 Burgstahler and Dichev (1997, footnote 6) assume the two numerator components are approximately independent so that the covariance of the numerator components is approximately zero. The effect of this simplifying assumption is examined in Section 3.1.1 below.
\[= Np_i(1-p_i) + 1/2N[p_i - 2p_i^2 - \delta + 4p_i\delta - 2\delta^2] + 2Np_i^2 - 2Np_i\delta \]
\[= Np_i - Np_i^2 + 1/2Np_i - Np_i^2 - 1/2N\delta + 2Np_i\delta - N\delta^2 + 2Np_i^2 - 2Np_i\delta \]
\[= 3/2 N \ p_i - N\delta(1/2+\delta). \]

For the case of a smooth earnings distribution where \(\delta = 0\), the variance expression in (4) simplifies to
\[V(\text{numerator}) = 3/2 N \ p_i. \quad (5)\]

The standardized difference statistic is defined as the numerator, \([n_i - \frac{1}{2}(n_{i-1} + n_{i+1})]\), standardized by the square root of the variance in equation (3) or (4), where each unobservable \(p_j\) is replace by the empirical estimate \(n_j/N\). The multinomial random variables in the numerator of the standardized difference are distributed approximately normal for large \(N\), so the standardized difference is distributed approximately normal \((0,1)\) and the significance of the standardized difference statistic is commonly evaluated by reference to a standard normal distribution.\(^{18}\)

Commonly-used rules of thumb from the statistics literature suggest that the normal approximation to the binomial is reasonably accurate for \(Np(1-p) \geq 25\).\(^{19}\) In typical applications where \(p\) is small so that \(Np(1-p) \equiv Np\), the statistical rule of thumb is approximately equivalent

\(^{18}\) Note that the normal approximation to the binomial does not rely on normality of the distribution of earnings. In contrast, some alternative significance tests for discontinuities rely on more stringent assumptions about the form of the distribution of earnings. For example, the test in Hayn (1995) requires the additional assumption that the distribution of pre-managed earnings is normal. Similarly, Chen et al (2010) assume the pre-managed distribution has a specific distributional form. In both cases, a significant test statistic could be due to a discontinuity but could instead be due to departures from the assumed pre-managed earnings distribution. Other statistical tests assume the pre-managed distribution is known or can be estimated without error. For example, Bollen and Pool (2009) assume that the pre-managed distribution of hedge fund returns is perfectly described by a distribution fitted to the histogram of reported hedge fund returns, as they assume there is no variance in their test statistic attributable to the use of the fitted distribution as an estimate of the unmanaged hedge fund return distribution. Still other alternative tests rely on the assumption that variability of the test statistic for the test interval can be estimated based on variability of a similar statistic in non-test intervals. For example, Degeorge, Patel, and Zeckhauser (1999) test whether the increment in observations at the earnings benchmark is significant relative to the variance of increments in a symmetric set of 10 intervals surrounding the benchmark.

\(^{19}\) Although they are not considered here, corrections for small sample sizes, including a simple continuity correction (see Johnson and Kotz equations 33 and 36, p. 62-64) are available and could be used to improve evaluations of significance when \(Np_i < 25\).
to the condition that the expected interval sample size, \( N_{pi} \), is 25 or larger.\(^{20}\) For sufficiently large expected values of \( N_{pi} = E[n_i] \) and \( N(p_{i-1} + p_{i+1}) = E[n_{i-1} + n_{i+1}] \), the standardized difference statistic is distributed approximately normal, but we use simulation evidence to assess the behavior of the standardized difference statistic for smaller interval sample sizes where the continuous normal approximation to the discrete multinomial distribution may be more problematic.

### 3.1 Relation to previous derivations in the literature

#### 3.1.1 BD assumption of independence of the numerator components

BD invoke a simplifying assumption that the number of observations in interval \( i \) and the sum of the numbers of observations in intervals \( i-1 \) and \( i+1 \) are independent. Therefore, the BD variance expression omits the covariance term in (4), \( 2N_{pi} (p_i - \delta) \). Omission of the covariance term results in a misstatement relative to the correct variance in (4) of:

\[
\text{Relative Misstatement} = -\frac{2N_{pi} (p_i - \delta)}{3/2N_{pi} - N\delta(1/2 + \delta)}.
\]

Because the expression \( -[2N_{pi} (p_i - \delta)] \) is virtually always negative, the BD assumption generally results in an understatement of the variance.\(^{21}\) Further, the percentage understatement is on the order of the probability in interval \( i \). For example, for the specific case of a smooth distribution where \( \delta = 0 \), the understatement simplifies to

\[
= -\frac{2N_{pi}^2}{3/2N_{pi}}
= -(4/3) p_i
\]

---

\(^{20}\) When the interval sample size is 25 or larger, the error induced by using estimated probabilities in calculating the variance is likely to be small. However, when the standardized difference test is applied to much smaller interval sample sizes, the use of estimated probabilities can lead to more serious errors. For example, in the extreme case where a very low interval sample size gives rise to an empirical frequency of zero, the estimated variance based on the zero frequency is zero.

\(^{21}\) This expression is never positive but it can be equal to zero if \( p_i = 0 \) or if both \( p_{i-1} \) and \( p_{i+1} \) are equal to zero.
The understatement of the variance results in a corresponding (though slightly smaller) percentage overstatement of the test statistics.22

3.1.2 BMN "correction"

BMN footnote 12 claims that the variance of the standardized difference test statistic derived in BD and widely used in the literature is incorrect: "The correct variance, however, is \( Np_i(1-p_i) + 1/4N(p_{i-1}+p_{i+1})(2-p_{i-1}-p_{i+1}) \). Because of the difference in the first term in the last parentheses, the estimated standard deviation used in BD and related papers is understated, resulting in an overstatement of the standardized difference test statistic." However, BMN provide neither a derivation for their "correction", nor an explanation of the purported error in the BD derivation.

The BMN expression for variance is substantially larger than the variance in equation (4). To see this, the BMN expression can be rewritten as the variance in equation (4) less the covariance term (the term that is omitted in the BD simplification), \( 2Np_i (p_i-\delta) \), plus a final unexplained term, \( 1/4N(p_{i-1}+p_{i+1}) \):

\[
V_{BMN} = [Np_i(1-p_i) + 1/4N(p_{i-1}+p_{i+1})(2-p_{i-1}-p_{i+1})] \\
= [Np_i(1-p_i) + 1/4N(p_{i-1}+p_{i+1})(1-p_{i-1}-p_{i+1})] + [1/4N(p_{i-1}+p_{i+1})] \\
= [Np_i(1-p_i)+1/4N(p_{i-1}+p_{i+1})(1-p_{i-1}-p_{i+1})] + [2Np_i(p_i-\delta)] - [2Np_i(p_i-\delta)] + [1/4N(p_{i-1}+p_{i+1})] \\
= 3/2Np_i - N\delta(1/2+\delta) - [2Np_i(p_i-\delta)] + [1/4N(p_{i-1}+p_{i+1})]
\]

The resulting misstatement relative to the correct variance in (4) is:

\[
\text{Relative Misstatement} = \{ -[2Np_i(p_i-\delta)] + [1/4N2(p_i-\delta)] / [3/2Np_i - N\delta(1/2+\delta)] \} .
\]

---

22 Since the standardized difference is standardized by the square root of the variance, a \((4/3)p_i\) relative understatement of the variance translates into understatement of the standard deviation by the square root of \([1 - (4/3)p_i]\) and a corresponding overstatement of the standardized difference statistic by the reciprocal of \([1 - (4/3)p_i]^{1/2}\). For example, when \( p_i = .01, .03, \) or \(.05 \) and the distribution is smooth, the BD approximation overstates the test statistic by about .7%, 2.1%, or 3.5%, respectively.
The BMN expression results in an overstatement of the variance, where the percentage overstatement is on the order of 33%. For example, for the specific case of a smooth distribution where $\delta=0$, the misstatement term reduces to

$$\text{Relative Misstatement} = \frac{\{-2N\pi_r^2 + \frac{1}{4N}2^2\}}{\frac{3}{2}N\pi_r} = -\frac{4}{3}\pi_r + \frac{1}{3}.$$ 

Because the variance is overstated by slightly less than 1/3, standardized differences constructed using the overstated variance have a standard deviation slightly less than the reciprocal of the square root of 4/3 which is slightly greater than .866.

**4. Simulation Evidence on the Distribution of the Standardized Difference Statistic**

In this section, we report the results of a simulation analysis that serves four purposes. First, the known parameters of the simulation allow us to verify the derivation of the distribution of the standardized difference test statistic, and specifically the expression for the variance in equation (4) or (5). Second, the simulation illustrates the behavior of the test statistic under conditions like those encountered in applications where the parameters that determine the variance are estimated rather than known. Results are provided for both large sample sizes, where the effect of using estimated variance should be minor, and for smaller sample sizes where the effect may be more substantial. Third, the simulation provides examples of the effect of estimated variances and departures from the null hypothesis assumption of smoothness on tests of significance for the distribution used to generate observations in the simulation, and these examples illustrate how this issue can be further explored for any other distribution posited as a description of the earnings distribution. Finally, the results provide evidence on the behavior of the test statistic using the simplified variance approximation from BD or using the incorrect expression from BMN.
Table 1 presents the results of a simulation of 1,000,000 simulation trials with observations generated from a normal distribution.\textsuperscript{23} The distribution of generated observations is segmented into 100 intervals between $-2$ and $+2$ standard deviations, where each interval has a width of .04 standard deviations. On each of the 1,000,000 simulation trials, standardized differences are evaluated for each of the 100 intervals. As illustrated in Figure 1, the 100 intervals are numbered relative to the mean of the simulated normal distribution, with interval numbers ranging from -50, …, -1, +1, …, +50. Column 1 of Table 1 contains the interval numbers and the expected interval sample size, $N_p$, is shown in column 2.

The normal distribution used to generate the observations is not smooth. As illustrated in Figure 1, for intervals more than 1 standard deviation below or above the mean (intervals $-50$ to $-26$ and $+26$ to $+50$), the normal distribution is convex, and in this range the expectations of the standardized differences are slightly negative. For intervals less than 1 standard deviation from the mean (intervals $-25$ to $+25$), the distribution is concave, the expectations are slightly positive. Column 3 shows the expected value of the standardized difference based on the known interval probabilities in the simulation.

Table 1 Panel A shows results for samples corresponding to 64,000 earnings observations, corresponding to typical sample sizes encountered in earnings management contexts.\textsuperscript{24} Panel B shows results for samples of 4,000 observations, illustrating the effects of a far smaller sample size, where there will be more substantial divergence between the continuous normal approximation to the discrete multinomial distribution.

\textsuperscript{23} The normal distribution is used to generate the simulation observations because the overall shape of the normal is broadly consistent with the overall shape of typical earnings distributions. However, the approximate normality of the standardized difference statistic does not depend on normality of the distribution generating the earnings observations but rather on the approximate normality of multinomial random variables for large interval sizes.

\textsuperscript{24} For example, BD examined distributions of earnings levels with a sample size of about 75,000 and distributions of earnings changes with a sample size of approximately 65,000.
4.1 Verification of distribution

Columns 4 and 5 show the means and standard deviations when the numerator of the generated standardized difference statistics is divided by the square root of the variance computed using the known interval probabilities. These standardized difference test statistics should have a mean equal to the theoretical expectation shown in column 3, and a standard deviation equal to one. In both panels, the difference between the average empirical mean in column 4 and the average theoretical mean in column 3 is less than .00001 and the standard deviation of the differences between the empirical mean in column 4 and the theoretical mean in column 3 is less than .001, the expected standard deviation for estimates of a random variable with unit variance when the estimates are based on 1,000,000 simulation trials. Also, the average of the standard deviations of the empirical standardized differences in column 5 is 1.00003 in Panel A and 1.0001 in Panel B, where the minimum of all the estimated standard deviations is .99244 and the maximum is 1.00825. Thus, the results in columns 4 and 5 are consistent with the distribution of the standardized difference derived in Section 3.

4.2 Distribution using estimated variances

In any application of the standardized difference statistic, the computed variance of the numerator must rely on estimated, rather than known, interval probabilities. Variances using estimated interval probabilities are subject to greater estimation error for smaller expected interval sample sizes (small N p), so there can be more substantial differences between the estimated and true variances.

Column 6 shows the average standardized differences when the variance is based on estimated probabilities. The average difference between the standardized differences using known versus estimated probabilities is less than .00001 and the standard deviation of the differences using known versus estimated probabilities is less than .001 in both Panels A and B.
Column 7 shows the empirical standard deviation of the standardized difference statistics based on estimated interval probabilities. In Panel A, where interval sample sizes are 16 times larger, the mean standard deviation in Column 7 is 1.0036 and the range is .9947 to 1.0125. In Panel B, the mean standard deviation in Column 7 is 1.0092 and the range is 1.0055 to 1.0182. Thus, the use of estimated probabilities in applications appears to result in a standardized difference statistics with standard deviations that average about .4% larger than 1 for the large interval sample sizes in Panel A and about .9% larger than 1 for the smaller interval sizes in Panel B.

4.3 Effects of departures from smoothness on significance tests

The null hypothesis of smoothness, which can also be described as a null hypothesis that the distribution that describes earnings is linear in the vicinity of the benchmark, is unlikely to hold precisely for distributions of unmanaged earnings. To illustrate the effect of departures from smoothness on the standardized difference statistic, observations in the simulation are generated from the normal distribution, which is concave within 1 standard deviation of the mean and convex beyond 1 standard deviation from the mean. Thus, the results of the simulation illustrate the effect of departures from smoothness for one example, the normal distribution. Perhaps more importantly, the analysis illustrates how to assess the effect of departures from smoothness for any other specified distribution.

The expected standardized difference in column 3 capture the degree of departure from smoothness. As the expectations become more positive, the proportion of significant test statistics should begin to exceed 5% and as expectations become more negative, the proportion should be less than 5%. The proportions of significant test statistic in column 10 are highly


25 For example, when the expected standardized difference is .021, the expected rejection rate is .0522, the probability of observing a value greater than 1.645 for a normal variate with mean .021 and standard deviation of 1. Thus, the expected rejection rate is about 5.2% in intervals -3 to +3 in Panel A. Similarly, when the expected
consistent with proportions predicted based on the expectation for each interval.\textsuperscript{26} For the normal distribution and interval definitions in the simulation, departures from the nominal 5% level are minimal – the maximum predicted proportion significant is 5.22% (for simulation intervals –1 and +1) and the maximum proportion significant realized in the simulation is 5.30% (for interval +2).

\subsection*{4.4 Effects of BD and BMN variances}

Column 8 shows the empirical standard deviation of the standardized difference statistics generated for each interval when the denominator variance is calculated using the BD independence assumption. Consistent with the analysis in section 3, the BD independence assumption yields a test statistic with a standard deviation roughly 1% greater than the theoretical standard deviation of 1. The average standard deviation in column 8 is 1.007 in Panel A and 1.012 in Panel B. This small inflation of the standard deviation of the standardized difference statistic leads to a small corresponding inflation of the proportion of significant test statistics. In column 11, the additional proportion significant using the BD independence assumption ranges from .01% to almost .1%, with an average of .055%. Thus, under the simulation conditions, the effect of using the BD independence assumption in tests of the significance of the standardized differences is very small.

Column 9 shows the empirical standard deviation of the standardized difference statistics when the standard deviation of the statistic is calculated using the BMN expression for variance. Consistent with the analysis in section 3, the average standard deviations in column 9 are slightly

\begin{tabular}{l}
standardized difference is \(-.022\), the expected rejection rate is .0478, the probability of a value greater than 1.645 for a normal variate with mean \(-.022\) and standard deviation of 1. Thus, the expected rejection rate is about 4.8% in intervals \(-49\) to \(-47\) or \(+47\) to \(+49\) in Panel A. \\
\textsuperscript{26} The average difference (not reported in the table) between the realized proportion significant and the predicted rate is \(-.0014\) in Panel A and \(-.0007\) in Panel B and the standard deviation of the difference between the realized proportion significant and the predicted rate is .0011 in Panel A and .0025 in Panel B.
\end{tabular}
greater than .866, specifically .8702 in Panel A and .8709 in Panel B. Column 11 shows that deflation of the standard deviation of the test statistic leads to a relatively large corresponding deflation of the proportion of significant test statistics, ranging from 1.11% to 2.15%, with an average of 1.98% in Panel A and 1.60% in Panel B. Thus, under the simulation conditions, using the BMN expression yields an effective level of about 3% for a test with a nominal level of 5%. This lower effective level will result in substantially lower power of tests using the BMN expression.

4.5 Summary

In summary, the simulation results in Table 1 are consistent with the analysis in Section 3. When the variance of the numerator of the standardized difference statistic is computed using the interval probabilities that are known in the simulation, the empirical distribution of the standardized difference statistic is consistent with the theoretical distribution derived in Section 3. When the variance is computed using estimated, rather than known, interval probabilities the distribution continues to be consistent with the standard normal distribution for intervals with an expected sample size greater than 25. Even for the interval sample sizes in Panel B less than 25 where the normal approximation is more problematic, departures from the assumed standard normal remain relatively minor. Finally, standardized differences relying on the simplifying BD independence assumption to compute the denominator variance have standard deviations about 1% larger than standardized differences using the correct variance expression, resulting in rejection rates just slightly higher than rates using the correct variance expression.

5. Power of the Standardized Difference Test for Discontinuities

Earnings management to meet a benchmark transforms pre-managed observations in intervals below the benchmark into reported (post-managed) earnings above the benchmark.
When the unobservable distribution of pre-managed observations is smooth, earnings management creates a trough below the benchmark and a peak above the benchmark in the post-managed distribution. Standardized difference statistics are designed to be sensitive to departures from smoothness due to earnings management.

The power of standardized difference tests to detect earnings management is determined by 1) the pre-managed earnings distribution in the vicinity of the benchmark, and 2) the specific way that earnings are managed.

5.1 Pre-managed earnings distribution

The power of tests for a discontinuity at the benchmark depends on the probabilities of pre-managed earnings in the two intervals immediately below the benchmark, referred to here as intervals –2 and –1, and in the two intervals immediately above the benchmark, referred to here as intervals +1 and +2. These probabilities depend on where the benchmark falls relative to the distribution, i.e., how far the benchmark is in the tail of the distribution, and which side of the distribution the benchmark is on.

Tests for management with respect to benchmarks further in the tail will have lower power because there are fewer pre-managed observations below the benchmark to be managed. Power is determined by a combination of the total number of observations in the earnings distribution, N, and the probability of pre-managed observations in the interval(s) below the benchmark that could be managed to meet the benchmark. In the following analysis, we consider the specific case where only observations in the interval immediately below the benchmark are managed to meet the benchmark, so power is determined by N and the probability

27 Note that in section 4 where we were considering the null hypothesis, intervals were numbered relative to the mean of the distribution, where –1 and +1 were the intervals immediately below and above the mean. In this section, intervals are numbered relative to the benchmark, where –1 and +1 are the intervals immediately below and above the benchmark.
of pre-managed earnings observations in the interval immediately below the benchmark, referred to as the interval sample size for interval \(-1\). Denoting the cumulative probability that a pre-managed earnings observation falls in interval \(-1\) as \(p_{-1}\), the interval sample size for interval \(-1\) is \(n_{-1} = Np_{-1}\). In the example power calculations below, \(p_{-1}\) is set at .01 and power is evaluated for \(N\) ranging from 1,000 to 128,000, so the interval sample sizes range from 10 to 1,280. Results in the \(N=1,000\) and \(N=2,000\) columns where interval sample sizes are 10 and 20, respectively, should be interpreted cautiously in light of evidence that the normal provides a reasonable approximation to the multinomial for expected sample sizes of 25 or greater.

Power also depends on which side of the distribution the benchmark falls. For example, when the benchmark falls below the mode of a typical symmetric earnings distribution, \(p_{-2} < p_{-1} < p_{+1} < p_{+2}\). On the other hand, when the benchmark falls above the mode, the inequalities are all reversed. Power will also depend on the relative values for probabilities in adjacent intervals (e.g., how much larger is \(p_{+2}\) than \(p_{+1}\), or how much larger is \(p_{+1}\) than \(p_{-1}\)) which in turn depends on the specific earnings distribution and on the location of the benchmark relative to the distribution. Because the inequalities relating the probabilities in the four intervals could go either direction and because there is no general relationship among the relative values of the probabilities, we adopt the simplifying assumption that the pre-managed probabilities are all equal to the probability for interval \(-1\), so that \(p_{-2} = p_{-1} = p_{+1} = p_{+2} = .01\).

\[28\] The location of the benchmark relative to the peak of the pre-managed earnings distribution can also affect the interpretation of the standardized difference test. This issue is relatively unimportant for well-behaved theoretical earnings distributions such as the normal distribution used in the simulation. (See the results in Table 1 for the intervals immediately adjacent to 0.) However, because empirical distributions often are far more leptokurtic than the normal distribution, sometimes modeled as mixtures of normal distributions or as symmetric stable distributions, empirical evidence of a significant standardized difference at the peak of the distribution should generally not be interpreted as evidence of earnings management.
### 5.2 Assumptions about how earnings are managed

The power of the standardized difference test also depends on how earnings are managed. In this section, we illustrate power calculations for one assumption. To the extent this assumption is a reasonable representation of how earnings are managed, these calculations are interesting in their own right. However, the calculations also serve to illustrate how power can be calculated for any assumption about how earnings are managed. We consider the specific assumption that earnings management effects are concentrated entirely in the interval immediately below and the interval immediately above the benchmark, i.e., earnings in interval –1 are managed upward so that reported (post-managed) earnings is in interval +1.

#### 5.2.1 Rate of earnings management (r)

We assume the probability of earnings management is r, representing the rate of earnings management.29 Because earnings are managed from interval – to interval +1, the expected number of reported earnings observations in interval –1 is less than the average of the numbers of observations in the adjacent intervals, i.e., intervals –2 and +1. Thus, the left standardized difference, the standardized difference for the first interval left of the benchmark has a negative expectation. Similarly, the expected number of observations in interval +1 is greater than the average of the adjacent intervals, i.e., intervals –1 and +2 and therefore the expectation of the right standardized difference, the standardized difference for the first interval above the benchmark, is positive.30 Other patterns of earnings management can be described in terms of more complex concentration effects, considered next.

---

29 For example, BD estimate that the rate of earnings management among firms with small pre-managed losses is in the range of 30-44% and the rate among firms with small pre-managed earnings decreases is 8-12%.

30 For this simple form of earnings management, the left and right standardized difference are alternative, but highly-correlated, tests for the effect of earnings management. For more complex forms of earnings management, either the left or right standardized differences may provide a more appropriate test, as discussed in more detail later.
5.2.2 Concentration

Concentration is determined by both how earnings are managed and by the choice of interval width in the research design. The results below analyze the research design choice of interval width, holding constant the assumption about how earnings are managed.

Specifically, we consider the effect on power 1) when interval width is cut in half so that earnings are being managed from two intervals below the benchmark and managed to two intervals above the benchmark, and 2) when interval width is doubled so that the first interval below the benchmark includes both the original set of observations where earnings are being managed and a second set of observations where earnings are not being managed.

More complex models might describe which pre-managed earnings observations are subject to earnings management, and how the post-managed earnings observations are distributed. For example, observations might be managed from more than one interval below the benchmark, rather than only the single interval immediately below the benchmark, where the rate of management likely decline for intervals further below the benchmark. Further, earnings might be managed to more than one interval above the benchmark. Evaluation of these possibilities is left for future research.

5.3 Analysis of the Power of the standardized difference test

The mean and variance of the standardized difference statistic given in equations (2) and (4), respectively, can be used to compute power under any specified alternative hypothesis. Power, denoted by \( 1 - \beta \), is a function of the level of the test, denoted by \( \alpha \).\(^{31}\) An observed

---

\(^{31}\) The logic for rejection of the null when a significant statistic is observed is that a significant statistic is more likely to occur under the alternative (with probability \( 1 - \beta \)) than under the null (with probability \( \alpha \)). Thus, the strength of the evidence that a significant result provides in favor of the alternative is described by the ratio \( (1 - \beta) / \alpha \). As the ratio approaches one (i.e., as power falls close to the level of the test), the effect of a significant test statistic on belief in null versus the alternative disappears. For further discussion, see, for example, Burgstahler (1987).
standardized difference statistic is significant at level $\alpha$ when the probability under the null hypothesis of a statistic as extreme as the observed value is less than or equal to the size of the test, $\alpha$. The power of the standardized difference test (the probability of rejecting a false null hypothesis) can be assessed for various combinations of the interval sample size ($N_{p-1}$), the rate of earning management ($r$), and concentration. To explore the effect of concentration, we begin with the simple case where earnings are managed from only the interval immediately left of the benchmark and earnings are managed to only the interval immediately right of the benchmark, i.e., where earnings management is concentrated entirely in the two intervals adjacent to the benchmark (Section 5.4.1). Then we explore the effects of smaller (Section 5.4.2) or larger (Section 5.4.3) interval widths.

5.3.1 Concentration in the two intervals immediately adjacent to benchmark

Table 2 evaluates the power of the standardized difference test statistic for various rates of earnings management, shown in the rows of the table and for overall sample sizes ranging from $N=1,000$, a sample size much lower than is typical in applications, to $N=128,000$, which is 50% to 100% greater than is typical in applications. In combination with the assumption that $\alpha=.01$, these sample sizes translate into interval sample sizes ranging from $N_{p-1}=10$ up to $N_{p-1}=1,280$. Panel A shows results for left standardized differences and Panel B for right standardized differences.

To illustrate how the values in Table 2 are calculated, consider the example of $N=64,000$ and $r=.10$. The assumptions above imply that the probabilities of post-managed observations in intervals $-2$ and $+2$ are the same as the pre-managed probabilities, $p_{-2} = p_{+2} = .01$ and the effect of an earnings management rate of 10% is to reduce $p_{-1}$ to .009, and increase $p_{+1}$ to .011. Table 2 shows that for an $\alpha=.05$ test, the power of the left standardized difference under these conditions.
assumptions is .938 and the power of the right standardized difference is .916. The assumptions in this example correspond to an expectation of 640 pre-managed observations in the interval left of the benchmark, with just 640 x .1 = 64 observations expected to be managed to exceed the benchmark. Thus, the results show that standardized difference tests have power in excess of 90% to detect management for just 64 observations in a distribution of 64,000, or just .1% of the total number of observations.

As another example, for N=16,000, an earnings management rate of just .20 yields power in excess of 90%. Thus, .20 x 160 = 32 observations out of 16,000 observations managed to exceed the benchmark or just .2% are sufficient for the standardized difference tests to have power greater than 90%. As a final example, Panel A shows that the left standardized difference is expected to be significant more than 90% of the time when just 8,000 x .01 x .300 = 24 observations are managed to meet the benchmark in a sample of 8,000 observations.

The shading in Panels A and B demarcates different strata of power. Cells where power is 1.000 are shaded dark gray, cells with power greater than .900 are medium gray, and cells with power greater than .500 are light gray. Overall, the results show that when the effects of earnings management are concentrated entirely in the interval below and above the benchmark and when the intervals in the vicinity of the benchmark contain about 1% of the distribution, the standardized difference tests have power greater than .900 to detect small rates and small numbers of observations managed to meet the benchmark.

---

32 These values are computed by using equations (2) and (4) to compute the mean and standard deviation, respectively, of the left-standardized difference given the values p_{-2}, p_{-1}, and p_{+1} and of the right-standardized difference given the values p_{+1}, p_{+2}, and p_{+1}. Then, the mean and standard deviation of each standardized difference distribution are used to find the probability of a left-standardized difference less than the critical value of –1.64485 and the probability of a right-standardized difference greater than 1.64485.

33 More precisely, the power in these cells is greater than .9995, so the values round to 1.000 with three decimal places of accuracy.
Of course, in applications it is unlikely that interval widths can be identified such that the effects of earnings management are concentrated entirely in the interval below and above the benchmark. Instead, it is likely that the rate of earnings management declines for observations further below the benchmark and that probability that an observation will be managed to exceed the benchmark by a given amount will decline with the magnitude of the amount. Thus, interval width is a difficult and important choice in designing tests of earnings management. While it is not possible to delineate and evaluate all the possible effects of interval width choice for all possible types of earnings management behavior, we provide results below showing the effect of choosing an interval width that is too narrow or too wide relative to the case illustrated in Table 2, where earnings management behavior is such that all effects are concentrated in the interval immediately below and above the benchmark. That is, we maintain the assumptions that earnings are managed from an interval below the benchmark with pre-managed probability = .01 at some rate r that applies uniformly to the entire interval, and that the managed observations are distributed uniformly in an interval above the benchmark with pre-managed probability = .01. In Section 5.4.2, we evaluate the effect on power of using a research design with an interval width that is only half as wide as the intervals in which earnings are concentrated (so that the implemented intervals widths have pre-managed probabilities = .005). In Section 5.4.3, we evaluate the effect on power of using a research design with an interval width that is twice as wide as the intervals in which earnings are concentrated (so that the implemented intervals widths have pre-managed probabilities = .02).

5.3.2 Intervals half as wide as the .01 interval in which earnings management is concentrated

Table 3 shows the power of the standardized difference test statistic when the pre-managed interval width is .005 and the effects of earnings management are concentrated in two intervals below the benchmark and in two intervals above the benchmark. The rates in the rows
of the table represent the rate of earnings management among pre-managed earnings observations in the first two intervals below the benchmark, interval –2 and –1, where the cumulative probability of pre-managed observations in these two interval \( p_{-2} + p_{-1} \), is assumed to be .01. Managed observations are assumed to move to the first two intervals above the benchmark, intervals +1 and +2. Panel A shows results for left standardized differences and Panel B for right standardized differences.

Because the effects of earnings management are now spread across multiple intervals, the prominence of the trough and peak relative to the surrounding intervals are reduced, and the power of the standardized difference tests is also reduced. Focusing on the same example we considered in Section 5.4.1 where \( N=64,000 \) and \( r=.10 \), the assumptions imply that the probabilities of post-managed observations in the four intervals are \( p_{-2} = p_{-1} = .0045 \), and \( p_{+1} = p_{+2} = .0055 \). In words, earnings management is expected to transform 10% of the pre-managed observations in intervals –2 and –1 into post-managed observations in intervals +1 and +2, so that the probabilities of a post-managed observation in intervals –2 and –1 are .0045 (so the total probability is .009, the same as in interval –1 in Section 5.4.1) and the probabilities of a post-managed observation in interval +1 or +2 are .0055 (so the total probability in the two intervals is .011, the same as in interval +1 in Section 5.4.1). Table 3 shows that for an \( \alpha=.05 \) test, the power of the left standardized difference under these assumptions is .447 and the power of the right standardized difference is .409, computed by using equations (2) and (4) to compute the mean and standard deviation and finding the probability of a left-standardized difference less than \(-1.64485\) and the probability of a right-standardized difference greater than \(1.64485\). Thus, using interval width that is much narrower than the interval width in which earnings management is concentrated results in a substantial loss of power for this example.
Panels C and D facilitate the comparison of the power results in Table 3 with the corresponding amounts in Table 2, and shows that the use of a too narrow interval width can result in a substantial loss of power. Panels C and D show gray shading for the cells where there cannot be much loss of power when the interval width is too narrow, either because the power is very high even with the narrow interval width in Table 3 (the dark gray shading corresponds to cells where the power in Panels A and B exceeds .900) or because the power is very low even with the optimal interval width in Table 2 (the light gray shading corresponds to cells where the power in Table 2 is less than .100). The remaining, unshaded cells in Table 3 Panels C and D show that the power using the too narrow interval width assumed in Table 3 is often on the order of 40%-60% of the power using the optimal interval width in Table 2. Thus, too narrow intervals can substantially reduce the power of standardized difference tests.

5.3.3 Intervals twice as wide as the .01 interval in which earnings management is concentrated

Table 4 shows the power of the standardized difference test statistic when the pre-managed interval width is .02 so that observations affected by earnings management are combined with observations not affected by earnings management in the interval below the benchmark and the interval above the benchmark. The rates in the rows of the table represent twice the rate of earnings management among pre-managed earnings observations in the interval below the benchmark where the cumulative probability of pre-managed observations in the interval, \( p_{-1} \), is now assumed to be .02. Managed observations are assumed to move to the first interval above the benchmark, intervals +1. Panel A shows results for left standardized differences and Panel B for right standardized differences.

Because intervals affected by earnings management are now combined with intervals not affected by earnings management, the power of the standardized difference tests is again reduced. Focusing on the same example we have used before where \( N=64,000 \) and \( r=.10 \), the
assumptions imply that the probabilities of post-managed observations in the four intervals are \( p_{-2} = .02, p_{-1} = .019, p_{+1} = .021, \) and \( p_{+2} = .02 \). In words, earnings management is expected to transform 10% of half of the pre-managed observations in interval –1 into post-managed observations in interval +1, so that the probability of a post-managed observation in interval –1 is .019 (i.e., the total probability is .01 + .009, where .009 is the same as in interval –1 in Section 5.4.1) and the probability of a post-managed observation in interval +1 is .021 (i.e., the total probability is .011 + .01, where .011 is the same as in interval +1 in Section 5.4.1). Table 4 shows that for an \( \alpha = .05 \) test, the power of the left standardized difference under these assumptions is .717 and the power of the right standardized difference is .698. Thus, using interval width that is much wider than the interval width in which earnings management is concentrated results in a substantial loss of power.

Panels C and D facilitate the comparison of the power results in Table 4 with the corresponding amounts in Table 2, and shows that the use of an interval width that is too wide can result in a substantial loss of power. As in Table 3, Panels C and D show gray shading for the cells where there cannot be much loss of power either because the power is very high even with the wide interval width in Table 4 or because the power is very low even with the optimal interval width in Table 2. The remaining, unshaded cells in Table 4 Panels C and D show that the power using the too narrow interval width assumed in Table 3 is often on the order of 50%-70% of the power using the optimal interval width in Table 2. Thus, too wide intervals can substantially reduce the power of standardized difference tests.

6. Conclusion

A large body of literature documents discontinuities in earnings distributions at prominent earnings benchmarks. These discontinuities have been widely interpreted as consistent with the theory that managers take actions to ensure that earnings meet benchmarks.
This interpretation is supported by survey evidence in Graham, Harvey, and Rajgopal (2005) indicating that managers are willing to incur real costs in order to meet such benchmarks. However, in situations where both discontinuity evidence and accruals-based models have been applied, abnormal accruals models have failed to find evidence that earnings management using accruals explain discontinuities.

This paper reviews and refines the derivation of the distribution of the Burgstahler and Dichev (1997) standardized difference statistic and corrects an error introduced into the literature in Beaver, McNichols, and Nelson (2007). The paper also provides a formal evaluation of the statistical properties of discontinuity tests to detect earnings management. The analysis shows the importance of defining interval widths that result in concentration of the effects of earnings management in the interval immediately below and the interval immediately above the benchmark – much narrower or much wider interval widths can substantially reduce the power of standardized difference tests.

The results show that standardized difference tests have the power to detect management of relatively small amounts of earnings (say .25%–.50% of MVE) by a small proportion of firms (say .1%–.2%). In contrast, previous research suggests that accruals-based tests have reasonable power only for far larger amounts of earnings management by a much larger proportion of sample firms (say 5% or more of total assets for all sample firms). Thus, together with evidence from previous papers, the evidence suggests that discontinuity tests have far greater power to detect smaller amounts and lesser rates of earnings management than do tests based on abnormal accruals.
References


Dyreng, S., Mayew, W., Schipper, K., 2012. Evidence that managers intervene in financial reporting to avoid working capital deficits. Working paper.


Figure 1
Intervals in Simulation