PSYCHOMETRIKA—VOL. 82, NO. 4, 1149–1161
DECEMBER 2017
https://doi.org/10.1007/s11336-016-9531-z

SOME REMARKS ON APPLICATIONS OF TESTS FOR DETECTING A CHANGE POINT TO PSYCHOMETRIC PROBLEMS

SANDIP SINHARAY

PACIFIC METRICS CORPORATION

Tests for a change point (e.g., Chen and Gupta, Parametric statistical change point analysis (2nd ed.). Birkhuser, Boston, 2012; Hawkins et al., J Qual Technol 35:355–366, 2003) have recently been brought into the spotlight for their potential uses in psychometrics. They have been successfully applied to detect an unusual change in the mean score of a sequence of administrations of an international language assessment (Lee and von Davier, Psychometrika 78:557–575, 2013) and to detect speededness of examinees (Shao et al., Psychometrika, 2015). The differences in the type of data used, the test statistics, and the manner in which the critical values were obtained in these papers lead to questions such as "what type of psychometric problems can be solved by tests for a change point?" and "what test statistics should be used with tests for a change point in psychometric problems?" This note attempts to answer some of these questions by providing a general overview of tests for a change point with a focus on application to psychometric problems. A discussion is provided on the choice of an appropriate test statistic and on the computation of a corresponding critical value for tests for a change point. Then, three real data examples are provided to demonstrate how tests for a change point can be used to make important inferences in psychometric problems. The examples include some clarifications and remarks on the critical values used in Lee and von Davier (Psychometrika, 78:557–575, 2013) and Shao et al. (Psychometrika, 2015). The overview and the examples provide insight on tests for a change point above and beyond Lee and von Davier (Psychometrika, 78:557–575, 2013) and Shao et al. (Psychometrika, 2015). Thus, this note extends the research of Lee and von Davier (Psychometrika, 78:557–575, 2013) and Shao et al. (Psychometrika, 2015) on tests for a change point.

Key words: likelihood ratio test, statistical process control, statistical quality control.

1. Introduction

Tests for a change point (e.g., Chen & Gupta, 2012; Hawkins, Qiu, & Kang, 2003) are methods from statistical quality control (SQC; e.g., Allalouf, 2007; Montgomery, 2013; von Davier, 2012) and are intended to detect whether there has been any change in the parameter(s) underlying a sequence of random variables. Lee and von Davier (2013) successfully used a *test for a change point* (TFCP) to detect an unusual change in the mean score of a sequence of administrations of an international language assessment. Shao et al. (2015) successfully used a TFCP to detect speededness in non-adaptive tests. The two *tests for a change point* (TFCPs) in Lee and von Davier (2013) and Shao et al. (2015) are quite different, in that the underlying random variable is continuous in the former but binary in the latter, the associated test statistic is like a T statistic in the former but a likelihood ratio test (LRT) statistic in the latter, and the critical value was obtained from a table in Hawkins et al. (2003) in the former but using simulations in the latter. These differences lead to at least the following questions

- What type of psychometric problems can be solved by TFCPs?
- What test statistics should be used with TFCPs in psychometric problems?
- What critical value would be appropriate for the test statistic?

The goal of this paper is to answer some of these questions by providing a general overview of TFCPs and then illustrating using three real data examples how these tests can be used to make

Correspondence should be made to Sandip Sinharay, Pacific Metrics Corporation, Princeton, NJ USA.
Email: ssinharay@ets.org

important inferences in psychometric problems, especially focusing on the choice of test statistics and corresponding critical values. The general overview encompasses the TFCPs of both Lee and von Davier (2013) and Shao et al. (2015). It is noted that two applications of TFCPs may involve the same type of data and the same test statistic but two different critical values depending on how long quality-control procedures have been used for the assessment that produced the data (e.g., Hawkins et al., 2003).

The general overview of TFCPs is then brought to bear on three real data examples that involve one data set each from the SAT Critical Reading, an adaptive licensure test, and a non-adaptive licensure test; TFCPs are applied in these examples to detect, respectively, a change in the mean score, person misfit, and item pre-knowledge. The applications of TFCPs in Lee and von Davier (2013) and Shao et al. (2015) are briefly revisited in these data examples and some clarifications and remarks are made on them.

2. A General Overview of Tests for a Change Point

Suppose that the data set consists of several observations X_1, X_2, …, X_n obtained at a sequence of time points. Let us also assume that the X_i's were produced by an underlying statistical model. A TFCP is employed to determine if there is a time point τ so that the model parameter underlying X_1, X_2, …, $X_{\tau-1}$ is substantially different from that underlying X_τ, $X_{\tau+1}$, …, X_n. The time point τ is referred to as the *change point*. There are several formulations of TFCPs, which are also referred to as tests for structural break/change, but the formulation that is most relevant to psychometric problems is discussed in, for example, Andrews (1993), Chen and Gupta (2012, p. 2), and Csorgo and Horvath (1997, p. 1), and is discussed next.

Let $X_1, X_2, \ldots, X_n$ be independent and unidimensional random variables. Let the probability function (that is a mass function if X_i's are discrete and a distribution function if X_i's are continuous) of X_i be $f_i(X_i; \psi_1, \eta)$ for $i = 1, 2, \ldots, \tau - 1$, and $f_i(X_i; \psi_2, \eta)$ for $i = \tau, \tau + 1, \ldots, n$, where ψ_1 and ψ_2 are unidimensional model parameters of interest and η is a unidimensional parameter not of interest (or, a nuisance parameter; an example of such a parameter is the population variance of X_i's when one is interested only in testing a hypothesis regarding the population mean of X_i's). Note that the X_i's have not been assumed to be identically distributed because, for example, they could denote scores on test items of an examinee in which case their distributions would not be identical (due to differences in the parameters over the items).

A TFCP commonly involves the testing of the null hypothesis $H_0 : \psi_1 = \psi_2$ against the alternative two-sided hypothesis $H_a : \psi_1 \neq \psi_2$ or the alternative one-sided hypothesis $H_a : \psi_1 > \psi_2$ or $H_a : \psi_1 < \psi_2$. Only the case of unknown ψ_1, ψ_2, η and τ, which is the most relevant to psychometric problems, will be considered here.

2.1. *Appropriate Test Statistics*

The choice of a test statistic for testing the abovementioned null hypothesis depends on the nature of the X_i's and hence of the f_i's.

2.1.1. Distribution of the Observations is Normal If the f_i's are adequately approximated by a normal distribution, with ψ_1 and ψ_2 as the means and η as the common variance (that is most common setting of TFCPs), then the generalized LRT of $H_0 : \psi_1 = \psi_2$ versus $H_a : \psi_1 \neq \psi_2$ can be performed using the test statistic

$$T_{\max,n} = \max_{1\le j\le n-1} |t_{jn}|, \tag{1}$$

$$\text{where } t_{jn} = \sqrt{\frac{j(n-j)}{n}}\,\frac{\bar{X}_{jn} - \bar{X}^*_{jn}}{s_{jn}}, \tag{2}$$

$$\bar{X}_{jn} = \frac{1}{j}\sum_{i=1}^{j} X_i,\ \bar{X}^*_{jn} = \frac{1}{n-j}\sum_{i=j+1}^{n} X_i,\ \text{and } s^2_{jn} = \frac{\sum_{i=1}^{j}(X_i - \bar{X}_{jn})^2 + \sum_{i=j+1}^{n}(X_i - \bar{X}^*_{jn})^2}{n-2}.$$

(see, e.g., Hawkins et al., 2003; Lee & von Davier, 2013; Montgomery, 2013). One rejects the null hypothesis if $T_{\max,n}$ is larger than an appropriately chosen critical value h_n.

Note that the comparison of $T_{\max,n}$ with h_n actually involves $(n-1)$ comparisons, that of $|t_{1n}|$ with h_n, $|t_{2n}|$ with h_n, …, and $|t_{n-1,n}|$ with h_n. However, while the comparisons in typical multiple comparison problems such as those considered in Benjamini and Hochberg (1995) are usually independent (mostly because they are performed on different individuals), those in a TFCP are dependent; for example, if X_{1n} is much larger than the rest of the observations in the sample, then $|t_{j1}|$ will be much larger than h_n and $|t_{j2}|$ will most likely be much larger than h_n as well.

If the f_i's are adequately approximated by a normal distribution and the alternative hypothesis is one-sided and is, for example, $H_a^{(1)}: \psi_1 > \psi_2$, one can use an one-sided version of $T_{\max,n}$ that is given by

$$T^{(1)}_{\max,n} = \max_{1\le j\le n-1} t_{jn}, \tag{3}$$

and is similar to a statistic suggested by Sen and Srivastava (1975).

2.1.2. Distribution of the Observations is Not Normal If the f_i's cannot be approximated adequately by a normal distribution, which could happen if the X_i's denote scores on binary items as in Shao et al. (2015), then the test statistic given by Eq. 1 cannot be applied. However, researchers such as Andrews (1993) and Csorgo and Horvath (1997) showed that in such a case, one can use the LRT statistic

$$L_{\max,n} = \max_{n_1\le j\le n-n_1} L_{jn} \tag{4}$$

to test the null hypothesis of no change versus a two-sided alternative, where

$$\begin{aligned} L_{jn} = 2\{&L_{j1}(\hat{\psi}_{1j}, \hat{\eta}_a; X_i, i = 1,2,\ldots,j) + L_{j2}(\hat{\psi}_{2j}, \hat{\eta}_a; X_i, i = j+1, j+2,\ldots,n) \\ &-L(\hat{\psi}_0, \hat{\eta}_0; X_i, i = 1,2,\ldots,n)\}, \end{aligned} \tag{5}$$

$$L_{j1}(\psi_1, \eta; X_i, i = 1,2,\ldots,j) = \sum_{i=1}^{j} \log f_i(X_i; \psi_1, \eta) \tag{6}$$

denotes the log-likelihood of $X_1, X_2, \ldots, X_j$ at ψ_1 and η, $L_{j1}(\hat{\psi}_{1j}, \hat{\eta}_a; X_i, i = 1,2,\ldots,j)$ and $L_{2j}(\hat{\psi}_{2j}, \hat{\eta}_a; X_i, i = j+1, j+2,\ldots,n)$ jointly maximize $L_{j1}(\psi_1, \eta; X_i, i = 1,2,\ldots,j)$ and $L_{j2}(\psi_2, \eta; i = j+1, j+2,\ldots,n)$, and $L(\hat{\psi}_0, \hat{\eta}_0; X_i, i = 1,2,\ldots,n)$ is the maximum value of $L(\psi, \eta; X_i, i = 1,2,\ldots,n)$. One rejects the null hypothesis if $L_{\max,n}$ is larger than an appropriately chosen critical value h_n. Note that $T_{\max,n}$ is a special case of $L_{\max,n}$ for normal f_i's. To increase the stability of the test, Andrews (1993) recommended setting n_1 equal to the integer nearest to $0.15n$ that restricts the change point to roughly the middlemost 70 % of the observations.

If the f_i's cannot be approximated by a normal distribution and the alternative hypothesis of interest is one-sided, for example $H_a^{(1)}: \psi_1 > \psi_2$, then one can use the one-sided Wald-type test statistic suggested by Estrella and Rodrigues (2005); in our context, the statistic is given by

$$W_{\max,n} = \max_{n_1\le j\le n-n_1} W_{jn}, \tag{7}$$

$$\text{where } W_{jn} = \frac{\hat{\psi}_{1j} - \hat{\psi}_{2j}}{\left[I^{-1}_{1j}(\hat{\psi}_0) + I^{-1}_{2j}(\hat{\psi}_0)\right]^{1/2}}, \tag{8}$$

where $I_{1j}(\hat{\psi}_0)$ and $I_{2j}(\hat{\psi}_0)$, respectively, are the estimated Fisher information on ψ contained in items 1 to j and $j+1$ to n, respectively, and computed at $\psi = \hat{\psi}_0$. Note that $T^{(1)}_{\max,n}$ is a special case, for normally distributed observations, of $W_{\max,n}$.

2.2. *Determining Critical Values for the Test Statistics*

Hawkins et al. (2003, p. 358), Montgomery (2013, p. 206), and Woodall and Montgomery (1999) noted that an application of SQC processes could be either a Phase I application or a Phase II application. Woodall and Montgomery (1999, pp. 378–379) noted that it is very important to distinguish between the two phases.

In Phase I, or, *retrospective analysis phase*, a set of data is gathered and analyzed. This data set is referred to as a static set of data and statistical methods are applied to this data set all at once in a *retrospective* analysis. Any hypothesis test performed at this phase involves a critical value at a desired significance level (such as 0.01 or 0.05) α, as in traditional hypothesis testing. A change point in the mean is a distinct possibility in a Phase I application (Hawkins et al., 2003, p. 358). Any unusual patterns in this data set indicate a lack of statistical control and lead to adjustments and fine tuning until a clean data set is achieved under stable (or in-control) operating conditions. This clean data set is often used to estimate the in-control distribution of the variable including its mean and standard deviation.

Phase II, or, *process monitoring phase*, starts after stability of the process has been achieved in Phase I and forms a never-ending or dynamic process. As each new data observation accrues, the SQC checks (that could include TFCPs) are reapplied; thus, the sample keeps growing. At Phase II, fixed significance critical values are not appropriate (Hawkins et al., 2003, p. 358); instead, one is concerned about the run length. Critical values h_n are determined based on the desired 'hazard' or 'alarm rate' α that is the conditional probability of a false alarm at any n given that there was no previous false alarms; the in-control average run length (ARL) is then $1/\alpha$.

Thus, the investigator should pick an appropriate critical value depending on whether the application is a Phase I application or a Phase II application. In addition, the distribution of either of the test statistics $T_{\max,n}$, $T^{(1)}_{\max,n}$, $L_{\max,n}$, and $W_{\max,n}$ under the null hypothesis is quite cumbersome and not of a well-known form in either of Phase I and Phase II applications even for large samples.

2.2.1. *Critical Values for $T_{\max,n}$ and $T^{(1)}_{\max,n}$*

Hawkins et al. (2003, p. 358) and Sullivan and Woodall (1996) recommended for $T_{\max,n}$ and two-sided alternatives the following critical values depending on whether the application is a Phase I or a Phase II application:

- Those provided in a table in Worsley (1979) in Phase I applications; Worsley (1979) obtained the critical values using theoretical calculations for $3 \leq n \leq 10$ and simulations for $15 \leq n \leq 50$.
- Those provided in tables in Hawkins et al. (2003) and on the website www.stat.umn.edu/hawkins in Phase II applications; these critical values were obtained using simulations.

For example, the critical value for $T_{\max,n}$ for $n = 50$ is 3.16 at significance level of 0.05 in a Phase I application (Worsley, 1979, p. 367), but is 2.354 for a hazard rate of 0.05 in a Phase II application when one starts after at least three observations have been collected (Hawkins et al., 2003, p. 360).

Note, regarding the critical value of 3.16 in Phase I applications, that if one were performing a 2-sample t test with 50 observations in the two samples together, the critical value at level 0.05 would have been 2.01; note further that if a Bonferroni correction were applied to adjust for the 49 comparisons (it was mentioned earlier that a TFCP using $T_{\max,n}$ involves 49 comparisons that are not independent), the critical value at level 0.05 would have been 3.50. The critical value 3.16 falls between these two values.

Critical values for $T^{(1)}_{\max,n}$ can be found using simulations. For example, the critical value for $T^{(1)}_{\max,n}$ for $n = 50$ is 2.88 (obtained using simulations) at significance level of 0.05 in a Phase I application.

2.2.2. Critical Values for $L_{max,n}$ and $W_{max,n}$ Researchers such as Andrews (1993) and Csorgo and Horvath (1997) provided asymptotic results on the distribution of $L_{\max,n}$ under the null hypothesis and tables with asymptotic critical values of the statistic for several values of n_1/n and significance levels in Phase I applications. Estrella and Rodrigues (2005) provided similar information for $W_{\max,n}$ in Phase I applications. These asymptotic critical values are often slightly inaccurate when the sample is not large (e.g., Gombay & Horvath, 1996). One can use simulations to obtain critical values of $L_{\max,n}$ and $W_{\max,n}$ for small samples. No critical values known to the authors are available for $L_{\max,n}$ or $W_{\max,n}$ in Phase II applications, but such values can be obtained using simulations.

2.3. What Critical Values Should be Used in Psychometric Problems?

The above discussion implies that in an application of a TFCP to a psychometric problem, one can use the statistic $T_{\max,n}$, $T^{(1)}_{\max,n}$, $L_{\max,n}$, or $W_{\max,n}$ depending on the nature of the random variable and the nature of the alternative hypothesis, but the choice of the critical value depends on the answers to the following two questions:

- Is the application a Phase I application or a Phase II application? Researchers such as Allalouf (2007), Lee and Haberman (2013), Lee and von Davier (2013), and von Davier (2012) noted that there have been few applications of SQC methods to psychometric problems and to existing educational and psychological assessments. Thus, the stability desired in Phase II applications has been achieved for very few assessments, if any. Further, the assessments, especially the educational ones, keep changing with change in policies (for example, the *No Child Left Behind Act* of 2001 was replaced in 2015 by the *Every Student Succeeds Act*). Therefore, most of the applications of TFCPs to psychometric problems are likely to be Phase I applications and critical values from Worsley (1979) and Andrews (1993) would be applicable to them. Alternatively, one can obtain critical values using simulations, especially in applications of $L_{\max,n}$ or $W_{\max,n}$ to samples that are not large. Critical values designed for Phase II applications can be used for an assessment only after SQC methods appropriate for Phase I applications have been used for it for a while, in which case the assessment can be considered "in-control" or "stable."
- Is one TFCP being performed or are multiple TFCPs being performed? If one performs multiple TFCPs as in Shao et al. (2015), then one may choose a critical value that adjusts for multiple comparisons by controlling the family-wise error rate (using, for example, a Bonferroni correction) or the false discovery rate or FDR (using the procedure of Benjamini & Hochberg, 1995). Shao et al. (2015) recommended controlling the FDR.

Thus, it seems that the choice of the test statistic and the critical value in an application of a TFCP to a psychometric problem is not straightforward. The next three sections bring to bear the above overview of TFCPs, especially the issues of the choice of test statistics and critical values, to three real data examples.

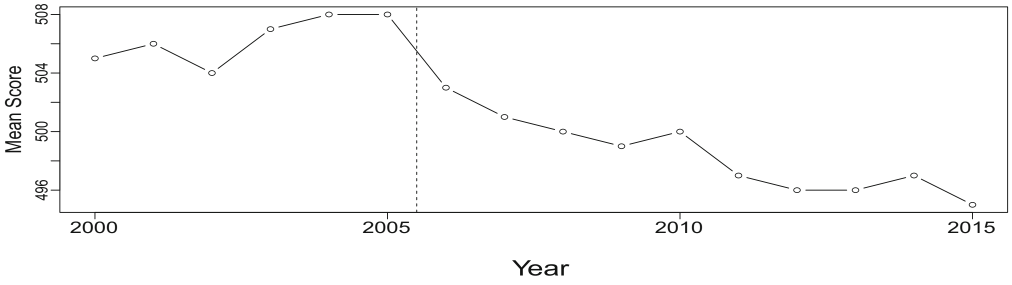

FIGURE 1.
Mean scores on SAT critical reading for the total group between 2000 and 2015.

3. Detecting a Change in the Mean Scores of SAT Critical Reading

3.1. Data

Let us consider the 2015 total-group profile report for College-bound seniors published by the College Board that is available at the website https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2015. Let us consider the mean scores on SAT Critical reading for the total group between the years 2000 and 2015, which are shown in Fig. 1.

Even slight changes in the mean of the SAT lead to big newspaper headlines; the goal here will be to examine if a TFCP indicates a change (either upward or downward) in the above mean scores.

3.2. The Appropriate TFCP for the SAT Data

Lee and von Davier (2013, pp. 561–563) considered X_i, $i = 1, 2, \ldots, n$, which are residuals from the application of a harmonic regression model (Lee & Haberman, 2013) to mean scores of a sequence of administrations of an international language assessment. They assumed the X_i's to be independent and to follow a normal distribution with unknown means and variance. They were interested in testing the null hypothesis of no change in the means of the X_i's against a two-sided alternative hypothesis—so they used the test statistic $T_{\max,n}$.

Our set up is exactly the same as the set up of the application of the TFCP in Lee and von Davier (2013). [1] Therefore, the test statistic $T_{\max,n}$ will be appropriate here as well.

3.3. Appropriate Critical Values for $T_{max,n}$

Lee and von Davier (2013, p. 564) used tables from Hawkins et al. (2003), which are appropriate in Phase II applications, to obtain a critical value for $T_{\max,n}$ in their TFCP. However, the application to the SAT data is actually a Phase I application because:

- The TFCP will be performed using the data all at once in a *retrospective* analysis here. In other words, $T_{\max,n}$ will be computed and compared to h_n only once, for $n = 16$. A Phase II application would have involved the comparisons of $T_{\max,n}$'s to h_n's for $n = 1, 2, \ldots, 16$.
- SAT has no history of established standards for SQC.

In our Phase I application, the critical values in Worsley (1979) would be more appropriate for $T_{\max,n}$, as discussed above. [2] The critical values at significance level = 0.05 are 3.36 for $n = 15$

[1] Of course, in our case, X_i's are the mean scores on SAT Critical Reading and n = 16.

[2] The author is of the opinion that critical values from Worsley (1979) would have been more appropriate in von Davier (2013) as well because theirs was also a Phase I application for the same reasons as in this case.

and 3.28 for $n = 20$; by linear interpolation (e.g., Thisted, 1988, p. 300), the critical value at significance level = 0.05 for $n = 16$ is $3.36 - \frac{16-15}{20-15}(3.36 - 3.28) \approx 3.34$. The critical value at significance level = 0.01 for $n = 16$ can be obtained similarly as 4.28.

3.4. Results and Validation of the Estimated Change Point

The value of $T_{\max,n}$ for these data was found to be equal to 7.65, which is much larger than the critical value 4.28 at level 0.01. The function 'WorsleyLikelihoodRatio' in the package 'climtrends' in the R software (R Core Team, 2016) can be used to compute $T_{\max,n}$. Thus, the value of $T_{\max,n}$ is statistically significant at level = 0.01—so it can be concluded that a statistically significant change occurred in the mean. The estimated change point was 2006. Figure 1 shows that this estimate makes sense—the trend in the mean changed from an upward trend to a downward trend in 2006. A question that naturally arises here is "Can the estimated change point of 2006 be validated with evidence that there was a major change in SAT between 2005 and 2006?" A study of the history of SAT (e.g., https://en.wikipedia.org/wiki/SAT) shows that SAT indeed underwent a major change in 2005—the questions on analogies were eliminated from the critical reading (or verbal) section and a new writing section, with an essay, was added; also, several sources such as CBS News (e.g., http://www.cbsnews.com/news/sat-scores-take-sharp-drop/) reported a sharp drop in the SAT scores in 2006. The College Board, the owner of SAT, explained that the drop was partly due to some students taking the newly lengthened test only once instead of multiple times. Thus, even though this example involves real data, it can be claimed that the TFCP provided an estimated change point that corresponds to a true change in the assessment concerned.

3.5. What if the Alternative Hypothesis Were One-Sided

One may wonder what the conclusion from a TFCP would have been if one were interested in testing the null hypothesis of no change against a one-sided alternative, for example, one corresponding to a drop in the mean score. To perform a TFCP against this one-sided alternative, $T^{(1)}_{\max,16}$ was also computed for the data. The value of $T^{(1)}_{\max,16}$ in this case is equal to $T_{\max,16}$ and is 7.65. [3] The 95th and 99th percentiles of the distribution of 10,000 values of $T^{(1)}_{\max,16}$ simulated under no change are 2.95 and 3.84, respectively. Thus, for example, the critical value for a one-sided alternative at level 0.05 would have been 2.95.[4] So, one would have rejected the null hypothesis at level 0.01 if one were interested in testing against the one-sided alternative of a drop in the mean. However, if one were interested in testing the null hypothesis of no change against the one-sided alternative of an increase in the mean, then one would have required to use the test statistic $\max_{1\le j\le n-1}(-t_{jn})$ and examine if it is larger than, for example, the critical value 2.95 at level 0.05; the value of $\max_{1\le j\le n-1}(-t_{jn})$ is negative for the SAT data—so one would not have rejected the null hypothesis (due to insufficient evidence in favor of the alternative of an increase in the mean) for the SAT data.

4. Assessment of Person Fit in Computerized Adaptive Testing

In a report for the Council of Chief State School Officers, Olson and Fremer (2013) recommended the use of person-fit statistics (PFSs), in addition to other methods, to detect irregularities in answering behavior. Researchers such as Bradlow, Weiss, and Cho (1998), Meijer (2002), and van Krimpen-Stoop and Meijer (2000) suggested, for use with computerized adaptive tests (CATs),

[3] This is because $t_{jn} > 0$ for the j for which $T_{\max,16} = |t_{jn}|$.

[4] Note that this value of 2.95 is close to what one would obtain as a critical value at level 0.10 (for a two-sided alternative hypothesis) for $T_{\max,16}$ from the table of Worsley (1979).

several PFSs that are based on the cumulative sum (CUSUM) procedure (e.g., Montgomery, 2013). Each CUSUM-based PFS involves a cumulative sum of positive and negative residuals after each item—a cumulative sum that is too large in absolute value indicates a person misfit. For example, van Krimpen-Stoop and Meijer (2000) and Meijer (2002) defined the iterative "upper" and "lower" cumulative statistics based on residual item scores on an n-item CAT as

$$C_i^+ = \max\{0, T_i + C_{i-1}^+\}; \; C_i^- = \min\{0, T_i + C_{i-1}^-\}, i = 1, 2, \ldots, n, \tag{9}$$

$$\text{where } T_i = \frac{1}{n}\left[X_i - P_i(\theta)\right], X_i = \text{Score on item } i, \text{ and } P_i(\theta) = P(X_i = 1)\cdot \tag{10}$$

Person misfit is concluded when, for an appropriate critical value h, C_i^+ is larger than h or C_i^- is smaller than $-h$ for some i. The CUSUM-based PFSs have been successful in detecting a string of consecutive correct or incorrect answers (e.g., Meijer, 2002, p. 223), which is mostly associated with an abrupt change in the test performance (in the form of tiredness, speededness, loss of concentration, item preknowledge, etc.) of an examinee.

Researchers such as Hawkins et al. (2003, p. 357) noted that to detect an abrupt change in the context of statistical process control, the CUSUM procedures are the most powerful when the parameters of the underlying statistical model before and after the change are known; however, if one or more of the parameters are unknown, the application of TFCPs may be more appropriate than that of the CUSUM procedures. Given that the examinee ability parameter is unknown [5] in CATs, TFCPs may be successful in detecting an abrupt change in test performance. However, examples of person-fit assessment (PFA) using TFCP in the context of CATs are severely lacking with the exception of Shao, Kim, Cheng, and Luo (2015).

4.1. Data

The available data set includes information on about 70,000 examinees who took a large-scale high-stakes health care licensure examination over a few months in 2015. The examination has been computer-adaptive in the last several years and currently is a variable-length CAT. Each examinee is administered a minimum of 60 operational items and a maximum of 250 operational items. The unidimensional Rasch model is used for item calibration and scoring. The current cut score used for passing is 0 in the logit scale. The item selection mechanism is based on the constrained CAT procedure of Kingsbury and Zara (1989); first, the content area of the item is chosen; then, to control item exposure, an optimal item is randomly selected from 15 items that provide the most information at the current ability estimate. On average, the examinees in the data set took 126 operational items. The goal in this example is to perform PFA using a TFCP.

4.2. The Test Statistic

An appropriate test statistic in this case is the $L_{\max,n}$ statistic defined in Eq. 4 for the Rasch model, where the ability estimated based on the first several items plays the role of $\hat{\psi}_{1j}$, the ability estimated based on the last several items play the role of $\hat{\psi}_{2j}$ and $\hat{\eta}_a$ is the null set in Eq. 5. Note also that Shao et al. (2015) and Shao et al. (2015) used the $L_{\max,n}$ statistic[6] for the 2PL model to detect speededness and warm-up effects, respectively, for non-adaptive and adaptive tests, respectively. Therefore, the test statistic used here is a special case of the statistic used in Shao et al. (2015) and Shao et al. (2015) with the slope parameters being equal over all items. The operational item parameters were used for the computation of $L_{\max,n}$. The maximum likelihood

[5] The item parameters are usually assumed known during CATs (e.g., Bradlow et al., 1998).

[6] Though they referred to the statistic as Δl_i.

estimate (MLE) of ability, restricted between −4.5 and 4.5, was used in the computations. It is expected that $L_{\max,n}$ will be large (and indicate person misfit) when an examinee produces a string of consecutive correct or incorrect answers, which is mostly associated with person misfit caused by tiredness, speededness, loss of concentration, item preknowledge, etc.

4.3. *Appropriate Critical Values for $L_{max,n}$*

To determine the null distribution and critical value of $L_{\max,n}$, Shao et al. (2015) and Shao et al. (2015) used simulations so that the FDR (Benjamini & Hochberg, 1995) is equal to 0.20. The same procedure, whose steps are given below, was used:

1. For each examinee, randomly permute the item responses and then compute the value of $L_{\max,n}$ 100 times; these (simulated) values can be considered to be those under the null hypothesis of no change (or no person misfit)
2. Set the critical value as the smallest value of T for which the FDR (Benjamini & Hochberg, 1995) given by

$$FDR = \frac{\text{The proportion of values of } L_{\max,n} \text{ simulated in Step 1 that are larger than } T}{\text{The proportion of values of } L_{\max,n} \text{ in the real data set that are larger than } T}$$

is equal to 0.20.

Because it is also of interest to examine what proportion of examinees are flagged by a TFCP at a fixed significance level, theoretical critical values for $L_{\max,n}$ provided in Andrews (1993) and Csorgo and Horvath (1997) were also used; for $n_1/n = 0.15$, the critical values are 8.85 at level = 0.05 and 12.35 at level = 0.01.

4.4. *Results*

The critical value corresponding to FDR = 0.20 was found to be 16.1 for these data and it led to the flagging of 0.8 % of the examinees in the data set.

The use of the theoretical critical values led to the flagging of 5.6 and 1.2 % examinees at significance levels of 0.05 and 0.01, respectively, for the data set.

Figure 2, which is like Fig. 4 of Shao et al. (2015), shows the score patterns of four examinees whose $L_{\max,n}$ was found significant at 1 % level and was also larger than 16.1. The item number is shown along the X-axis and the score on the item (0 for "Incorrect" and 1 for "Correct") is shown as a hollow circle along the Y-axis. Thus, for example, a hollow circle near the top of a panel represents a correct response. The estimate of the change point for each examinee is shown as a vertical dashed line. For example, the estimated change point on the top left panel is 27. The title of each panel shows two ability estimates (MLE up to the first decimal place), one based on the items up to the estimated change point and the other based on the items after the estimated change point. All the examinees represented in the figure received between 60 and 70 items.

Figure 2 shows that the estimated change points seem to represent the change in performance of the examinees quite accurately. The top two panels represent examinees whose performance dropped substantially during the test and the MLE after the estimated change point is much smaller compared to that up to the estimated change point. These examinees most likely suffered from fatigue or speededness. The bottom two panels represent examinees whose performance improved substantially during the test. The examinee represented in the bottom right panel most likely had trouble settling in or warming up (a phenomenon mentioned by, for example, Meijer, 2002, p. 227). Thus, $L_{\max,n}$ seems to perform reasonably well in detecting person misfit for the data set. Sinharay (2016) provided more details on $L_{\max,n}$ and showed in a simulation study that $L_{\max,n}$

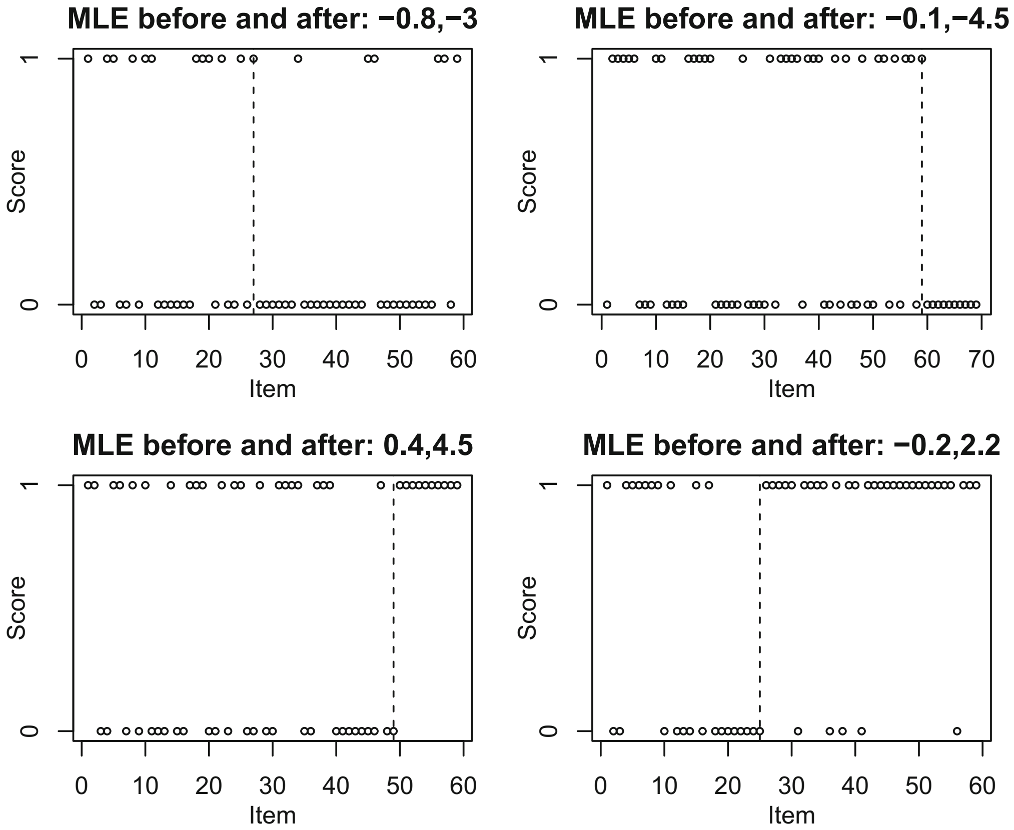

FIGURE 2.
Score patterns of four examinees who were flagged.

is more powerful than several CUSUM-based PFSs in detecting abrupt change in performance during an assessment.

5. Detection of Preknowledge in a Non-Adaptive Licensure Test

5.1. Data

Let us consider a data set, which was analyzed in several chapters of Cizek and Wollack (2017) with a focus on detecting test fraud by the examinees, from one form of a licensure examination. The test form includes 170 operational items that are dichotomously scored. Item scores on the form were available for 1644 examinees. The licensure organization that administers the examination identified as compromised 61 items on the form. The organization also flagged 48 individuals for the form as possible cheaters from a variety of statistical analysis and an investigative process that brought in other information. The goal of the analysis here will be to examine if a TFCP can flag some examinees, especially those who were flagged by the licensure organization.

5.2. Reformulation of the Problem as a TFCP

Researchers such as Sinharay (in press) suggested detecting item preknowledge by examining if the performance of an examinee is significantly better on the compromised items than on the non-compromised items. Further, Bradlow et al. (1998) recommended a CUSUM-based approach to examine if the performance of an examinee is much better on one subarea compared to the rest of the test (they referred to such examinees as subexperts) after reordering the test so that the s items from the subarea appear in the beginning of the reordered test (the order within the subarea is unimportant); the subexperts would undergo an abrupt performance change roughly at item number $s + 1$. It follows, from Sinharay (in press) and Bradlow et al. (1998) and from the observation in Hawkins et al. (2003) that TFCPs may be more appropriate than the CUSUM procedures in detecting abrupt changes, that the problem of detection of item preknowledge can be reformulated as a TFCP after reordering the test so that the 61 compromised items appear in the beginning of the reordered test. It is expected that the examinees with item preknowledge would undergo an abrupt performance drop roughly at item number 62.

Note that a TFCP is not claimed to be the most powerful test to detect item preknowledge; most likely, the tests described in Sinharay (in press) would be more powerful than a TFCP; however, it is demonstrated in this example that a TFCP has satisfactory power to detect item preknowledge.

5.3. *Test Statistic and Critical Value*

Because the alternative hypothesis of interest here is one-sided (as only a performance drop is of interest), the test statistic $W_{\max,n}$ would be appropriate for performing a TFCP.[7] So, the values of $W_{\max,n}$ were computed for all examinees in the data set after reordering the items.

The Rasch model is operationally used in the assessment—the difficulty-parameters under the Rasch model were estimated from the data set and were used in the computations. The MLE, restricted between -4.5 and 4.5, was used as the estimate of the examinee ability.

As in the previous data example, a simulation-based procedure (Shao et al., 2015) was used to determine a critical value for which FDR is equal to 0.20. The critical value was 3.40.

The theoretical critical values of $W_{\max,n}$, from Table 1 in Estrella and Rodrigues (2005), are 2.70 and 3.30 at significance levels 0.05 and 0.01, respectively, for $n_1/n = 0.15$. A simulation under no change point produced critical values that were very close—so these theoretical critical values were used.

5.4. *Results*

Using the theoretical critical value (at level of 0.05) of 3.30, the value of $W_{\max,n}$ was significant for 12.7 % examinees in the whole data set and for 33.3 % examinees among the 48 individuals flagged by the licensure organization as possible cheaters. Note that the signed LRT statistic (SLRTS), which was found as the most powerful among the statistics for preknowledge detection by Sinharay (in press), was significant at level 0.05 for 17 % examinees in the whole data set and for 38 % examinees among the 48 individuals flagged by the licensure organization; further, the correlation coefficient between $W_{\max,n}$ and the SLRTS was 0.91. Thus, $W_{\max,n}$ performs quite similar to the SLRTS. For the 16 examinees among the 48 flagged by the licensure organization for whom $W_{\max,n}$ was significant at level 0.05, the estimated change point associated with $W_{\max,n}$ was between items 58 and 66, which is very close to the expected change point of item 62.

The critical value of 3.4 (corresponding to an FDR of 0.20) for $W_{\max,n}$ led to the flagging of 4.7 % examinees in the whole data set and 13.5 % examinees among the 48 examinees flagged by the licensure organization as possible cheaters.

[7] We think that if the goal of the investigator is to detect only speededness, then the underlying alternative hypothesis is one-sided because speededness leads to a performance drop later in the test, and $W_{\max,n}$ would be more appropriate than $L_{\max,n}$. A brief simulation study shows that $W_{\max,n}$ leads to a slightly larger power than $L_{\max,n}$ in such cases. Of course, if the goal is to detect both speededness and warm-up, then $L_{\max,n}$ is more appropriate than $W_{\max,n}$.

TABLE 1.
Some details for five examinees for the real data.

P_{comp}	$P_{\text{non-comp}}$	$\hat{\theta}_{\text{comp}}$	$\hat{\theta}_{\text{non-comp}}$	$W_{\max,n}$	ECP
0.90	0.61	2.81	0.39	5.67	63
0.87	0.60	2.45	0.34	5.16	61
0.90	0.68	2.81	0.76	4.96	66
0.95	0.79	3.60	1.42	4.67	59
0.93	0.79	3.28	1.42	4.39	64
0.33	0.33	−0.59	−0.85	1.21	NA

The first five rows of numbers in Table 1 provide, for five examinees flagged by the licensure organization, the proportion correct score (the number correct divided by the total number of items) on the compromised items (P_{comp}) and non-compromised items ($P_{\text{non-comp}}$), the corresponding ability estimates $\hat{\theta}_{\text{comp}}$ and $\hat{\theta}_{\text{non-comp}}$, $W_{\max,n}$, and the estimated change point (ECP). The last row of the table provides the same numbers for one examinee who was not flagged by the licensure organization. The performance on the compromised items was much better than that on the other items (both with respect to proportion corrects and ability estimates) for the first five examinees; naturally, all of these examinees were flagged by both the statistics at level of 0.01. Also, the expected change point was close to 62 for all of them. The examinee corresponding to the sixth row performed about equally well on the compromised and non-compromised items and the examinee was not flagged by any of the statistic. Thus, the results show that the TFCP led to conclusions that can be validated by other evidence (in the form of the information on flagging by the licensure organization).

6. Conclusions

Given an increased emphasis on quality control in educational and psychological assessments (e.g., Allalouf, 2007; von Davier, 2012), the applications of TFCPs to important psychometric problems in Lee and von Davier (2013) and Shao et al. (2015) are timely. This paper provides further insight on the TFCPs that promise to be useful to psychometricians; for example, it is discussed that critical values of the test statistics for use with TFCPs should be chosen depending on, for example, whether the application is a Phase I application or a Phase II application and whether an adjustment for multiple comparison is needed. The three data examples have some differences from those in Lee and von Davier (2013) and Shao et al. (2015); for example, one of the three real data examples involves a CAT and another involves a one-sided alternative, whereas the examples of Lee and von Davier (2013) and Shao et al. (2015) involved non-adaptive tests and two-sided alternatives. The TFCPs are shown to provide useful information in the three real data examples; the estimated change points are validated by other information in two of these examples (first and third). Thus, it is hoped that this paper, along with Lee and von Davier (2013) and Shao et al. (2015), will help psychometricians have a thorough understanding of the TFCPs.

The TFCPs including those of Lee and von Davier (2013) and Shao et al. (2015) should be explored further for possible applications to psychometric problems where some quantity is monitored over time for possible changes. Some of the examples involve monitoring of growth in academic achievement, monitoring of students for potential dropouts, monitoring of achievement gap, and monitoring of item statistics for potential exposures. TFCPs should also be explored to psychometric problems involving changes in multiple parameters, for example, to detect a change in both the mean and variance of the scores on a test (as mentioned by Lee & von Davier, 2013,

p. 573) or to both item scores and response times (as mentioned by Shao et al., 2015, p. 22). Furthermore, tests for multiple change points, tests for change points with epidemic alternatives, and Bayesian TFCPs (e.g., Chen & Gupta, 2012) could be considered in future.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26*(1), 36–46.

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *6*(61), 821–856.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 289–300.

Bradlow, E., Weiss, R. E., & Cho, M. (1998). Bayesian detection of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, *93*, 910–919.

Chen, J., & Gupta, A. K. (2012). *Parametric statistical change point analysis* (2nd ed.). Boston, MA: Birkhuser.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Washington, DC: Routledge.

Csorgo, M., & Horvath, L. (1997). *Limit theorems in change-point analysis*. New York, NY: Wiley.

Estrella, A., & Rodrigues, A. (2005). *One-sided test for an unknown breakpoint: Theory, computation, and application to monetary theory* (Staff Reports No. 232). Federal Reserve Bank of New York.

Gombay, E., & Horvath, L. (1996). On the rate of approximations for maximum likelihood tests in change-point models. *Journal of Multivariate Analysis*, *56*, 120–152.

Hawkins, D., Qiu, P., & Kang, C. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, *35*, 355–366.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 557–575.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*, 557–575.

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, *39*, 219–233.

Montgomery, D. C. (2013). *Introduction to statistical quality control*. New York, NY: Wiley.

Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test securities irregularities*. Washington, DC: Council of Chief State School Officers.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: Austria.

Sen, A. K., & Srivastava, M. S. (1975). On tests for detecting a change in mean. *Annals of Statistics*, *3*, 98–108.

Shao, C., Kim, D., Cheng, Y., & Luo, X. (2015, April). *Detection of warm-up effect in cat using change-point analysis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Shao, C., Li, J., & Cheng, Y. (2015). A change point based method for test speededness detection. *Psychometrika*. (Advance online publication. doi:10.1007/s11336-015-9476-7).

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics*, *41*, 521–549.

Sinharay, S. (in press). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*.

Sullivan, J. H., & Woodall, W. H. (1996). A control chart for preliminary analysis of individual observations. *Journal of Quality Technology*, *28*, 265–278.

Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation*. London: Chapman and Hall.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Netherlands: Springer.

von Davier, A. A. (2012). *The use of quality control and data mining techniques for monitoring scaled scores: An overview* (ETS Research Report No. RR-12-20). Princeton, NJ: ETS.

Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, *31*, 376–386.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, *74*, 365–367.

Manuscript Received: 10 FEB 2016
Final Version Received: 25 MAY 2016
Published Online Date: 21 OCT 2016