Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 25 Number 1, January 2020 ISSN 1531-7714

# Detection of Learning in Longitudinal Assessment via Score-based Tests

Ting Wang, *The American Board of Anesthesiology*
Huaping Sun, *The American Board of Anesthesiology*
Yan Zhou, *The American Board of Anesthesiology*
Ann E. Harman, *The American Board of Anesthesiology*

Longitudinal assessment is a type of assessment involving repeated measures over a period to evaluate whether and when an attribute (e.g., ability, skill) changes. Thus, change detection is of central interest in longitudinal assessment. In the assessment setting, change in the desired direction (typically upward) is often referred to as "learning". While many methods have been proposed to detect change, they all require a prior definition of a changing point. However, this information is often unknown in practice. As an alternative, we focus on a family of tests based on stochastic processes of case-wise derivatives of the likelihood function (i.e., scores). These score-based tests could detect "learning" without prior information of a changing point and signal the changing point to the users. In this article, we will illustrate what the score-based tests are and the novel application by using the data of two physicians participating in longitudinal assessment for a medical specialty certifying board's continuous certification program.

Longitudinal assessment is a type of assessment involving repeated measures over a period of time to evaluate whether and when an attribute (e.g., ability, skill) underlying a set of observations changes. Thus, change detection is of central interest in longitudinal assessment. In the assessment setting, change in the desired direction (typically upward) is often referred to as "learning". If a new education intervention is implemented, it would be valuable to assess whether the abilities of the participants have improved, and if so, at what time point the improvement occurs. From a statistical modelling standpoint, change is manifested by parameter instability. There are extensive studies about parameter instability detection in econometrics (Andrews, 1993; Brown, Durbin, & Evans, 1975; Hansen, 1992; Hjort & Koning, 2002; Horn & McArdle, 1992; Nyblom, 1989), policy analysis (Zeileis & Hornik, 2007) and drug intervention (Hothorn & Zeileis, 2008). In these studies, the computational tool is a family of statistical tests based on stochastic processes of case-wise derivatives of the likelihood function (referred to as scores). These score-based tests require estimation of the null model only (i.e., when parameter stability is assumed to hold), and have been applied to linear models, generalized linear models (GLMs), Rasch models and factor analysis models (Merkle, Fan, & Zeileis, 2014; Merkle & Zeileis, 2013; Strobl, Kopf, & Zeileis, 2015; Wang, Merkle, & Zeileis, 2014; Zeileis & Hornik, 2007).

The goal of this article is to demonstrate the use of score-based tests to detect learning in an individualized longitudinal assessment program comprising single best answer multiple-choice questions exclusively. First, we introduce the theoretical framework of the score-based tests. Second, we apply the score-based tests to the assessment data of two physicians with the same end-of-year percent correct scores and demonstrate how the proposed test differentiates between the learner and the non-learner with tutorial code. Finally, we discuss the tests' applications and limitations.

## Overview of Longitudinal Assessment Platform

In 2016, a medical specialty certifying board launched a web-based longitudinal assessment platform for its continuous certification program. Physicians who register for the program are required to answer 30 single best answer multiple-choice questions in each calendar quarter – a total of 120 questions in a calendar year. Each response is scored dichotomously (correct is coded as 1 and incorrect is coded as 0), yielding a series of binary data for each physician. In the following, we illustrate how to model these data in -GLM, a flexible generalization of ordinary linear regression that allows for the binary response data to be related to a linear model via a link function.

## Model: Generalized Linear Model (GLM)

For the binary data collected for each physician, GLM in binomial family can be utilized with logit link in the following form:

$$y_i \sim \text{Bernoulli}(p), \quad (1)$$

$$\log\left(\frac{p}{1-p}\right) = b_0, \quad (2)$$

where $i \in 1, 2, 3, \ldots n$ represents the number of observations for each physician, with $n = 120$ for the current data set; $y_i \in 0, 1$ representing physicians' incorrect (0) or correct (1) response to each question. Parameter $b_0$ is considered as the mean of the logit function of the percent correct score $p$. Therefore, the instability of percent correct score is reflected in parameter $b_0$ change in the GLM, which could be detected in the score-based tests. In the following section, we describe how score-based test can be utilized as a tool to detect parameter instability.

## Method: Score-based Tests

In this section, we review the score-based tests' theoretical background and describe a test statistic that particularly fits the purpose of change detection. Related descriptions can be found in the literature (Merkle & Zeileis, 2013).

### Score

The above GLM's log-likelihood function can be written as the sum of observations' log-likelihoods

$$\ell(b_0; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i|p), \quad (3)$$

where $f()$ is the GLM's parametric distribution.

Maximizing the model's log-likelihood function is equivalent to solving the first-order conditions

$$\sum_{i=1}^{n} s(\widehat{b_0}; y_i) = 0, \quad (4)$$

where

$$\widehat{b_0} = argmax_{b_0} \ell(b_0; y_1, y_2, \ldots, y_n). \quad (5)$$

and

$$s(\widehat{b_0}; y_i) = \frac{\partial \ell(y_i, b_0)}{b_0}|_{b_0=\widehat{b_0}} \quad (6)$$

### Distribution Theory

The functions of the scores obtained above follow a stochastic process along an auxiliary variable $V$ (i.e., time). We can build the following intuition for the tests. We examine observations' scores as V is moved from its least value to the greatest. If the parameter is stable, the scores should fluctuate around zero. Conversely, the scores will significantly shift from zero when the parameter demonstrates instability.

To obtain test statistics, we define a cumulative score as

$$\mathbf{B}(t; \widehat{b_0}) = \hat{\mathbf{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{[nt]} \mathbf{s}(\widehat{b_0}; \mathbf{y}_{(i)}) \quad (0 \le t \le 1), \quad (7)$$

where $\mathbf{y}_{(i)}$ represents the observed data vector for *i*th-largest observation, with the order determined by the auxiliary variable $V$. $\hat{\mathbf{I}}$ denotes the estimate of the covariance matrix of the scores, which decorrelates the fluctuation processes associated with observation model parameters. To account for the possible autocorrelation, we use the heteroskedasticity and autocorrelation consistent (HAC) estimation to adjust the covariance matrix. $[nt]$ is the integer part of $nt$ (i.e., a floor operator), and $0 \le t \le 1$. With a sample size of $n$, $\mathbf{B}(t; \widehat{b_0})$ changes at $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$. For $t = 1$the cumulative score vector always equals 0, as defined in Equation (4).

Under the null hypothesis of parameter stability, $\mathbf{B}(t;\widehat{b_0})$ converges in distribution to an independent Brownian bridge (Hjort & Koning, 2002):

$$\mathbf{B}(t;\widehat{b_0}) \xrightarrow{d} \mathbf{B}^0(\cdot), \quad (8)$$

where $\mathbf{B}^0(\cdot)$ is a unidimensional Brownian bridge associated with the parameter $b_0$.

To define a unidimensional Brownian bridge, let $W(t)$ represent the value of the stochastic process at the point $t$, then the Brownian bridge satisfies the following conditions:

$$\mu(t) = 0 \; \forall \, t, \quad (9)$$

$$\mathrm{Cov}(t_1, t_2) = \min(t_1, t_2) \quad (10)$$

$$W(0) = 0, \quad (11)$$

$$W(1) = 0, \quad (12)$$

where $\mu(t)$ and $\mathrm{Cov}(t_1, t_2)$ respectively represent the mean *t* and covariance between points $t_1$ and $t_2$. The beginning and ending are restricted to be 0 ("tied down"). A graph of a simulated Brownian bridge is shown in Figure 1.

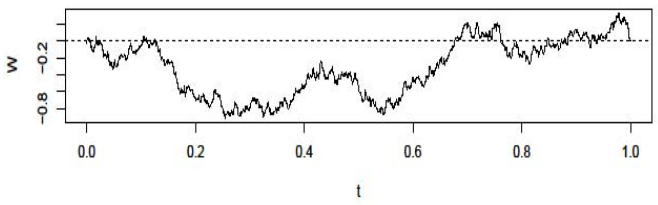


**Figure 1**. Example of a simulated unidimensional Brownian bridge. The dashed line represents 0, which is the beginning and ending of the stochastic process.

To obtain scalar test statistics, we summarize the empirical behavior of Equation (6) and compare it to the analogous scalar summary of the Brownian bridge.

### Test Statistic

After summarizing the empirical cumulative score process via a scalar, the asymptotic distribution of the scalar can be obtained by applying the same summary to the asymptotic Brownian bridge. This yields critical values and *p* values. Various statistics have been proposed, and selection of a statistic could be based on the plausibility of potential instability patterns.

In this application, the following method is used to detect $b_0$ instability. Parameter stability is rejected if the largest component of the empirical cumulative score vector is greater than the critical value. The value of *V*, at which the violation occurs, indicates the location of the detected component. This statistic is called the "double maximum" ($DM$):

$$DM = \max_{i=1,\ldots,n} |\mathbf{B}(\widehat{b_0})_i|. \quad (13)$$

Specifically, time is the auxiliary variable of interest and cumulative sum scores fluctuate as more observations are added. If parameter stability holds, then observation scores will fluctuate randomly around zero. If parameter changes, the score associated with the parameter will be greater than the critical value.

## Empirical Study

In this empirical illustration, we apply the score-based test described above to 120-question response data from two physicians participating in the longitudinal assessment program. These two physicians have the same percent correct scores for the entire calendar year, but their quarterly response patterns are different. The aim is to detect whether there are parameter changes with respect to time for these two physicians (i.e., whether these two physicians "learn" during the four quarters).

### Descriptive Statistics

The two physicians' quarterly and end-of-year percent correct scores are displayed in Table 1. The former (Q1-Q4 columns of Table 1) is calculated based on 30 responses received quarterly; the latter is based on 120 responses collected by the end of the year (last column of Table 1). Their overall percent correct scores for the entire year are the same (59.17%). However, Physician 1 has a lower percent correct score in the second quarter (Q2) in comparison to Q4; whereas Physician 2 has relatively "stable" percent correct scores across the four quarters. The descriptive statistics presented in Table 1 do not tell us whether the higher percent correct score in Q4 for Physician 1 indicates "learning" or just random fluctuation of performance. In addition, even if "learning" occurs, we do not know where the exact changing point is (e.g., the end of Q3,

the beginning of Q4, or the end of Q4). In the following section, we will use the score-based test to answer these questions. The test is conducted in R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria).

**Table 1.** Quarterly and end-of-year percent correct scores for two physicians. Q1- Q4 indicates quarter 1 to quarter 4, respectively. Sum indicates the end-of-year percent correct score.

| | Q1 | Q2 | Q3 | Q4 | Sum |
|---|---|---|---|---|---|
| Physician 1 | 53.33% | 46.67% | 53.33% | 83.33% | 59.17% |
| Physician 2 | 60.00% | 56.67% | 66.67% | 53.33% | 59.17% |

## Score-based Test Results

When the score-based test is applied to the two physicians' responses, each response has a score ($b_0$, corresponding to the logit function of the percent correct score) to describe how well the model describes the observation. The responses are ordered according to the response time, and we search changing points in the scores on that sequential order. In this case, if the $b_0$ parameter changes, the statistic $DM$ will be greater than its corresponding critical value. With the test statistic's fluctuation displayed across the entire year, peaks higher than the critical value indicate changing points.

To demonstrate how to conduct the score-based test analysis, we use two physicians' data for simplicity. The tests can be scaled up to all physicians, which will be described later. We first load the data.

```
load("physician1.rda")
load("physician2.rda")
```

The data are a sequence of 120 binary data for each physician.

Physician 1:

```
1 0 1 1 0 0 0 1 1 0 1 1 0 0 1 1 0 1 0 0 0
0 0 1 1 0 1 1 1 1 0 1 0 0 0 1 0 0 1 0 0 1
1 0 0 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1 0 1
0 1 1 0 1 1 0 1 0 1 1 0 1 1 0 0 0 0 0 0 0
1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 1
1 0 1 1 1 1 1 1 1 0 1 1 1 1 1
```

Physician 2:

```
0 0 1 1 1 1 0 0 1 1 0 1 0 1 0 0 0 0 1 1 1
0 1 0 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 0 1
1 0 0 0 1 1 0 1 0 1 0 1 0 0 0 1 1 1 0 0 1
1 0 1 1 1 0 1 1 1 1 1 0 0 1 1 1 1 0 1 1 1
1 1 0 1 0 0 0 1 0 1 1 0 1 0 0 1 1 1 1 0 1
0 1 1 0 0 0 0 0 1 1 1 0 1 1 0
```

Next, we extract the cumulative score fluctuation using the function `gefp()` from `strucchange` package with the code below:

```
Cumscore_physican1 <- gefp(Response ~ 1,
  family = binomial,  vcov = kernHAC,
  data =  physician1)
```

where Response ~ 1 represents the intercept-only GLM, and the linking function belongs to the binomial family because of binary response data. These two arguments together implement Equation (1) and (2). The vcov argument is specified as "kernHAC" to account for possible autocorrelation among the responses. "data" argument specifies the data set. The same function is applied to Physician 2's data set.

```
Cumscore_physican2 <- gefp(Response ~ 1,
  family = binomial,  vcov = kernHAC,
  data =  physician2)
```

Finally, we compare the maximum of these cumulative scores against the critical value obtained from Brownian bridge, and retrieve $p$-value from `sctest()` function (in `strucchange` package):

```
sctest(Cumscore_physician1, "max")
sctest(Cumscore_physician2, "max")
```

where the "max" argument requests the $DM$ statistics described in Equation (13). The above code returns the statistics and the corresponding $p$ values for Physicians 1 and 2, respectively:

```
f(efp) = 1.6011, p-value = 0.01187
f(efp) = 0.60765, p-value = 0.854
```

These results indicate Physician 1 has experienced significant ($p$ value less than 0.05) percent correct score change whereas Physician 2 has not.

In addition to the test statistics, adding `plot = TRUE` argument in `sctest()` generates an instability plot for each physician to facilitate visual representation. Figure 2 displays the resulting plots. The left panel and right panel represent Physician 1 and Physician 2, respectively. The x-axis represents the auxiliary variable (i.e. time) as shown in Equation (7), with every 30 responses indicating a quarter in this example. The curved line depicts the test statistic fluctuation process for each observation (greater values reflect more instability), with the dashed horizontal line representing the critical value. As stated before, the hypothesis of parameter stability is rejected if the $DM$ test statistic (the maximum in the

fluctuation process or the peak in the plots) crosses the critical value. It can be observed that Physician 1's parameter changes in September, while Physician 2's parameter is stable across the year.

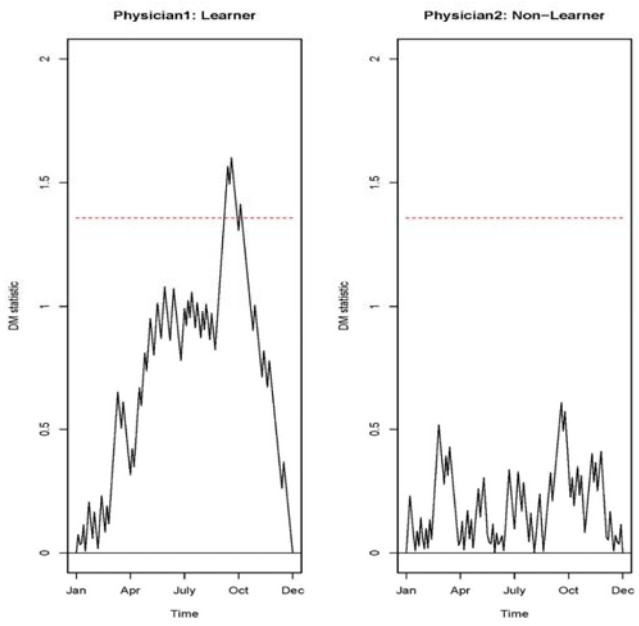


**Figure 2.** *DM* statistic with respect to time. Dashed red horizontal line represents the critical value at $\alpha$ =0.05. Left panel represents Physician 1 and right panel represents Physician 2, both testing parameter $b_0$, the logit function of percent correct scores.

In general, Physician 1's plot can be seen as a prototype of a "learner", whose parameter changes (and the percent correct score increases in this case); whereas Physician 2 can be seen as the typical response pattern of a "non-learner", with no statistics crossing the line of the critical value.

# Discussion

## Summary

In this article, we introduced the theoretical background of score-based tests and analyzed two physicians' response data with the same overall percent correct scores in a one-year longitudinal assessment program. The score-based test shows that Physician 1's percent correct score changes significantly and the changing point is in September, whereas Physician 2's percent correct score does not fluctuate significantly across the year. This analysis can be easily applied to a great number of physicians' response data by using `for` loop and parallel computation using `parallel` or `foreach` package for speed-up. Among 18,297 physicians who completed 120 questions in the 2016 longitudinal assessment program, 1,637 (9%) were detected to have changed in either direction. Among these, 92.3% (1,524 out of 1,637) of the physicians' percent correct scores increased, signaling "learning" in this setting.

## Applications

The score-based tests illustrated in this paper provide a convenient statistical tool to monitor changes in longitudinal assessment platforms, informing with the changing point(s). Multiple statistics could cover a wide range of applications. In particular, the auxiliary variable does not need to be continuous time points. For example, the auxiliary variable could be students' grade, age group or cognitive ability. In such cases, researchers can simply use the ordinal statistics by changing the argument in the `sctest()` with `WDMo` or `maxLMo`. The tests are easy to use for this purpose since no new model estimates are required.

## Extensions

In this article we focused on testing GLM estimated by maximum likelihood function. The score-based tests described here generally apply to estimation methods that maximize/minimize an objective function. For example, the tests have been applied using pairwise maximum likelihood estimation (Wang, Strobl, Zeileis, & Merkle, 2018). In addition, the estimated model could be multivariate models, such as structural equation modeling (Merkle & Zeileis, 2013) and item response theory (Wang et al., 2018). These previous studies focused on the measurement invariance issue, and the tests can be applied to detecting parameter instability in general.

## Limitations

Score-based tests are subjected to several limitations. First, score-based tests only identify "change". "Non-learners" can be those with constantly high, mediocre, or low performance. Second, in order to differentiate whether the parameter change represents an increase or a decrease, parameters before and after the changing point must be compared. Lastly, score-based tests do not tell us "why" the change occurs. The percent correct change might be attributable to factors other than change in individual's ability such as item

difficulty drift, familiarity with the platform or fatigue/burnout. Further examination on each individual could facilitate better understanding of the reason for change.

## References

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica:* Journal of the Econometric Society, 821-856.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society*. Series B (Methodological), 149-192.

Hansen, B. E. (1992). Testing for parameter instability in linear models. *Journal of Policy Modeling*, 14(4), 517-533.

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14(1-2), 113-132.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, 18(3), 117-144.

Hothorn, T., & Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics*, 64(4), 1263-1269.

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, 79(4), 569-584.

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: a generalization of classical methods. *Psychometrika*, 78(1), 59-82.

Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, 84(405), 223-230.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, 80(2), 289-316.

Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Frontiers in psychology*, 5, 438.

Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, 83(1), 132-155.

Zeileis, A., & Hornik, K. (2007). Generalized M - fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4), 488-508

**Citation**:

Wang, Ting, Sun, Huaping, Zhou, Yan, & Harman, Ann E. (2019). Detection of Learning in Longitudinal Assessment via Score-based Tests. *Practical Assessment, Research & Evaluation*, 25(1). Available online: https://scholarworks.umass.edu/pare/vol25/iss1/1/

**Corresponding Author**

Ting Wang, Psychometrician
The American Board of Anesthesiology
4208 Six Forks Road, Suite 1500
Raleigh, North Carolina 27609-5765

email: ting.wang [at] theABA.org