Practical Assessment, Research, and Evaluation

Volume 28 Article 2

February 2023

Using Cumulative Sum Control Chart to Detect Aberrant Responses in Educational Assessments

Siyu Wan
University of Massachusetts Amherst

Lisa A. Keller
University of Massachusetts - Amherst

Follow this and additional works at: https://scholarworks.umass.edu/pare

Recommended Citation

Wan, Siyu and Keller, Lisa A. (2023) "Using Cumulative Sum Control Chart to Detect Aberrant Responses in Educational Assessments," *Practical Assessment, Research, and Evaluation*: Vol. 28, Article 2.
DOI: https://doi.org/10.7275/pare.1257
Available at: https://scholarworks.umass.edu/pare/vol28/iss1/2

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 28 Number 2, February 2023 ISSN 1531-7714

Using Cumulative Sum Control Chart to Detect Aberrant Responses in Educational Assessments

Siyu Wan, *University of Massachusetts Amherst*
Lisa A. Keller, *University of Massachusetts Amherst*

Statistical process control (SPC) charts have been widely used in the field of educational measurement. The cumulative sum (CUSUM) is an established SPC method to detect aberrant responses for educational assessments. There are many studies that investigated the performance of CUSUM in different test settings. This paper describes the CUSUM procedure and shows how it can be used to monitor the test-taking process and detect aberrant responses. It aims to provide an accessible guide for the CUSUM control chart.

Keywords: CUSUM, Aberrant response, PFS

Introduction

When using any assessment scores, the underlying assumption is that examinees' responses should not be affected by unusual behaviors. However, various aberrant responses are commonly seen in practical testing situations. For example, Ward et al. (2016) found that around 10% of examinees had careless responses during a test. Besides carelessness, test behaviors such as speededness, lack of motivation, cheating, pre-knowledge of items, warm-up effect, and fatigue all cause aberrant responses (Sinharay, 2017b; Zhang et al., 2020). Having aberrant responses will also undermine the accuracy of the ability estimation of examinees, leading to invalid conclusions and inferences based on assessment scores (Shao, 2016). It is crucial to detect aberrant responses to reduce their negative influences on test score validity.

There are different approaches to detect aberrant responses. For instance, one direction is to directly model aberrant behaviors based on different assumptions. In addition to using traditional unidimensional item response models (IRT), models have been developed, such as the Mixture Model (Bolt, Cohen, & Wollack, 2002), the Hybrid Model (Yamamoto & Everson, 1997), and the Graduate Change Model (Wollack & Cohen, 2004). This study adopted another approach, modeling the responses using standard models and flagging responses or response patterns that do not fit the typical response model. In the early development of this approach, researchers utilized different person-fit statistics (PFS) to identify test takers who showed abnormal item response patterns (which resulted in artificially high- or low-test scores). They then separate those test takers from those who exhibited normal item response patterns (Karabatsos, 2003). However, traditional PFS indices do not perform well in detection (De La Torre & Deng, 2008). Most person-fit research focuses on paper-and-pencil examinations in the literature (van Krimpen-Stoop & Meijer, 2000). Statistical process control (SPC) methods were recently introduced to detect aberrant responses, particularly the cumulative sum control chart (CUSUM).

In 1998, Bradlow, Weiss, and Cho firstly adopted the CUSUM method to detect four types of aberrant behaviors: warm-up effects, fatigue, sub-expertise, and lack-of-fit, in a computerized adaptive test (CAT). Since then, many articles have utilized the CUSUM

method to detect aberrant response behaviors in a test (Armstrong & Shi, 2009a, 2009b; Meijer, 2002; Tendeiro & Meijer, 2012; Tendeiro et al., 2013; van Krimpen-Stoop & Meijer, 2000, 2001). It provides a practical methodology to examine changes numerically and visually, using person- fit statistics (PFS) over time (Armstrong & Shi, 2009a).

Compared to other SPC methods, such as change-point analysis, the CUSUM owns simplicity in terms of computation. However, the procedure of generating a CUSUM plot is not straightforward. This study will demonstrate the standard CUSUM procedure using one widely used index, the weighted residual between expected and observed scores. This study aims to provide an accessible guide for researchers who want to apply CUSUM in their research. First, we introduce the CUSUM method under the traditional SPC framework. Then the procedure of CUSUM will be explained in the context of aberrant response detection in educational testing, including the steps of determining the boundary values. For illustration, we simulate two aberrant behaviors, the warm-up effect, and random responses. The definitions of these two behaviors are given in the later section. We can easily observe differences between normal and aberrant behaviors through its corresponding CUSUM chart. Finally, we discuss the advantages and disadvantages of the CUSUM method.

A Review of CUSUM Procedures

Yu and Cheng (2022) defined SPC as "a collection of methods for monitoring, controlling, and improving a random process through statistical analysis" (p. 2). It was initially developed and utilized to monitor product quality in production or manufacturing areas. Using SPC, the product quality can be actively measured and charted simultaneously while manufactured things are mass-produced. For example, if one company makes chocolate beans, each bag should have a certain amount of beans. If each bag has too many beans, that will increase the costs of materials for the company. Nevertheless, customers will not be satisfied if too few beans are in a bag. Therefore, the company needs to control the number of chocolate beans during production, which can be achieved through SPC.

The production or process that needs to be statistically controlled usually has a stable distribution. Schafer, Coverdale, Luxenberg, and Jin (2011) described a general procedure of SPC: Plotting the index of the product, such as means of groupings of products, on a chart within certain limits. Technicians visually analyze charts to see whether deviations from expectations are beyond specific boundaries. Suppose the graph contains any pattern or the points deviate too far from the expected values. In that case, the process is considered "out of control." Some variability is expected due to sampling variations and variances across sampled groups. Suppose the variations are within predefined limits, and the pattern of deviations appears random. In that case, the process is considered "under control." When this situation occurs, there is no need to do any following analysis.

Montgomery (2013) provides a detailed review of many different types of control charts. Recently, SPC was introduced to the educational measurement field to detect aberrant responses. For example, Omar (2010) used Shewhart's mean and standard deviation charts to measure and monitor the consistency of rating performance items in operational assessments. The CUSUM (Page, 1954) is another established SPC method widely used in educational measurement to detect aberrant responses. It is effective to detect small changes in the variable being measured.

Traditional CUSUM Procedure

van Krimpen-Stoop and Meijer (2000) described a traditional CUSUM procedure: Let Z_t be the value of a standard normally distributed statistic Z collected at a time point t (t = 1, 2, …) from a sample of size N (e.g., the standardized average number of chocolate beans). Let d be a reference value (the choice of d value is given below); statistics sums are accumulated in two directions only if they surpass the "goal value" by more than d units: when $Z_t > d$ (e.g., there are more chocolate beans in a bag than expected), positive values are accumulated in C^+; when $Z_t < -d$ (e.g., there are fewer chocolate beans in a bag than expected), negative values are accumulated in C^-. The starting values are $C^+ = C^- = 0$. The two-sided CUSUM process for each time point t is shown below:

$$C_1^+ = \max\{0, Z_1 - d\}, \quad (1)$$

To understand equation 1 better, Z_1 represents the value of *some* statistic that is computed to monitor the statistical process, and d is the tolerated deviation from that statistic. Since Z is a value from a standard normal distribution, it is centered around 0. Assume the value

of *Z* is 1.3. Then the difference between *Z* and *d* would be 0.8, and the value of equation (1) would be 0.8. If the value of the difference is negative, the value of equation (1) goes to zero.

$$C_2^+ = \max\{0, (Z_1 - d) + (Z_2 - d)\}$$
$$= \max\{(Z_2 - d) + C_1^+\}, \quad (2)$$

The process from equation (1) continues, and we accumulate the values of the differences between *Z* and *d*. Assume this time that the value of *Z* is 0.4. In this case, the difference between *Z* and *d* would be -0.1, and the value of $C^+$is 0.8 + (-0.1) = 0.7.

$$C_3^+ = \max\{0, (Z_3 - d) + C_2^+\}, \quad (3)$$

$$C_t^+ = \max\{0, (Z_t - d) + C_{t-1}^+\}, \quad (4)$$

At the same time,

$$C_t^- = \min\{0, (Z_t + d) + C_{t-1}^-\}, \quad (5)$$

C_t^+reflects the sum of consecutive positive values of $Z_t - d$, and C_t^- reflects the sum of consecutive negative values of $Z_t + d$,. Sums are accumulated on both sides simultaneously. It is worth noting that both C_t^+ and C_t^- can go back to 0, as they are not constantly accumulating. Only $|Z_t| > d$, Z_t will be taken into account. Let *h* be a pre-determined threshold value. When $C_t^+ > h$, or $C_t^- < -h$, the process is "out-of-control." Otherwise, the process is "in control." A complete example with computations is provided in the illustration.

The underlying assumption of the traditional CUSUM procedure (van Krimpen-Stoop & Meijer, 2000) is that the Z_t values are asymptotically standard normally distributed. The *d* and *h* values are determined based on this assumption. The value of *d* is often chosen to be one-half the magnitude of the mean shift (in $Z_t$units) to be detected; for instance, *d* = 0.5 is a good option for detecting a shift of one standard deviation of Z_t. Many analytical studies investigated how to determine h values after choosing the *d* value. Briefly speaking, the *d* and *h* parameters should be selected to provide good average run length performance (ARL). ARL is separated into two types. ARL_0 is similar to type one error and represents the expected number of samples until a control chart signals, given that the process is in control; in other words, it sends a false alarm. At the same time, ARL_Δ is similar to a true positive, the expected number of samples until a control chart signals, given that the process is out of control. The process is like balancing type I error and power in the same way as hypothesis testing. Montgomery (2013) provided some general recommendations for selecting *d* and *h* and the underlying rationale for this choice (p.422). Usually, *h* = 4 *or* 5 when *d* = 0.5.

CUSUM Procedure Based on the Weighted Residual

Like many other SPC methods, CUSUM requires identifying a variable that represents the quality of the process needing statistical control (Omar, 2010). Researchers have proposed various CUSUM indices to represent the test-taking quality in an educational test. van Krimpen- Stoop and Meijer (2000) offered 8 CUSUM indices, all of which were some kinds of residuals (weighted or unweighted) between the expected and observed scores of an item. Other CUSUM indices are based on the likelihood ratio test (Sinharay, 2016; Armstrong & Shi, 2009a). Yu and Cheng (2022) provided a comprehensive review of these statistics. This section introduces the CUSUM procedure using the most widely used index (T_i, see equation 6), the weighted residual, for illustration purposes. Researchers can use other indices to suit their needs, but the CUSUM process remains the same.

Let *i* denote the i^{th} item in a test. T_i is the residual between the observed score (X_i,) and the expected score of the i^{th} administered item, given the length of the test (N):

$$T_i = \frac{1}{N}\left[X_i - E(X_i|\hat{\theta})\right]. \quad (6)$$

The expected score, $E(X_i|\hat{\theta})$ is usually calculated based on the unidimensional (dichotomous or polytomous) IRT models. In this study, we used the 1-parameter logistic model (1 PL) for its simplicity:

$$E(X_i|\hat{\theta}) = \frac{1}{1 + e^{-D(\hat{\theta} - b_i)}}, \quad (7)$$

where b_i is the item difficulty parameter. Student estimated ability is represented by $\hat{\theta}$. D is a scaling constant (D = 1.7 to scale the logistic to the normal ogive metric; D = 1 to use the logistic metric). If the examinee correctly answers an item, the T_i will be positive; otherwise, the T_i will be negative.

Similar to the traditional CUSUM procedure, the start points C_t^+ and C_t^-r are 0. The residuals given in equation (6) are summed across consecutive items. For each examinee, after each administered item *i*, the CUSUM can be shown as:

$$C_t^+ = \max\{0, T_i + C_{i-1}^+\}, \quad (8)$$

$$C_t^- = \min\{0, T_i + C_{i-1}^-\}. \quad (9)$$

C_t^+ and C_t^- are the cumulative sum of the consecutive positive and negative residuals (T_i). A series of consecutive positive values of T_i will make C_t^+ larger, while consecutive negative values of T_i will make C_t^- smaller. Let UB and LB represent the pre-specified upper and lower bound, respectively. If $C_t^+ >$ UB, or $C_t^- <$ LB at some points, we can identify this response pattern as aberrant. Otherwise, we will classify this item score pattern as fitting the unidimensional IRT model. The CUSUM control chart is the scatter plots of C_t^+ and C_t^- at every item point *i*.

However, the null distributions of proposed CUSUM indices are usually far from normal. For example, residual T_i given in equation (6) follows a binomial distribution. Therefore, it is not appropriate to set $d = 0.5$ and UB = 4, LB = −4. There are several methods to determine UB and LB. In general, it is necessary to define a level of statistical significance (usually at a 5% level) first. Then the most extreme value is found, and the UB and LB are values for which 2.5% of the most extreme values lie above (UB) or below (LB). This process can be achieved through Monte Carlo simulation or based on the empirical dataset at hand. We will explain this procedure later.

Illustration through the Simulation

To demonstrate how to utilize CUSUM charts to identify abnormal response behaviors, we simulate normal responses and two aberrant behaviors: warm-up effects and random responses for a 40-item test. In this simulation study, we used the 1PL model. Item difficulty parameters are generated from a standard normal distribution (mean=0, SD=1). Table 1 provides the item parameter values. An examinee has a warm-up effect when he or she performs poorly at the beginning of the exam for some reason, such as nervousness or unfamiliarity, and then gradually returns to normal as the exam progresses. Random responses mean the examinee makes a random choice for items. Random responses might be caused by a lack of time (speededness) or motivation. They are more likely to occur near the end of the exam. CUSUM could, of course, be used to detect other aberrant behaviors such as item pre-knowledge (van Krimpen-Stoop & Meijer, 2002), fatigue (Armstrong & Kung, 2011), and sub-expertise (Bradlow et al., 1998), and speededness (Yu & Cheng, 2022).

As previously stated, one critical step of the CUSUM process is finding upper and lower boundaries for non-normally distributed data. We can complete this step in an operational setting through Monte Carlo simulation or empirical data analysis. In this study, we used Monte Carlo simulation to determine the boundary values (UB and LB):

1) The latent abilities of 10,000 examinees (θ) were generated from the standard normal distribution. Item scores were generated based on examinees' θs and item parameters (Table 1).

2) The probability of endorsing items, T_i based on estimated $\hat{\theta}$, and cumulative weighted residuals (C^+ and C^-) were calculated.

3) The maximal C^+ and minimal C^- values of each examinee were collected. Then the LB and UB were identified as 2.5% and 97.5% percentile of 10,000 extreme values, respectively.

4) We repeated the previous steps 100 times, and a final LB and UB was the average value across 100 replications. UB= 0.114, LB= -0.114.

We can also use the empirical dataset at hand to determine the UB and LB. Instead of generating responses of examinees, we can sample a group of examinees with approximately similar $\hat{\theta}$ of the population. Then, as illustrated in steps 2 and 3, we calculate the UB and LB using extreme C^+ and C^- values of the subsample of examinees. There is a drawback of using empirical item responses instead of simulation: the existing misfitting item scores might affect boundary values. However, we expect most item score patterns will fit the underlying IRT model in reality. Meijer (2002) used both simulation and empirical datasets to investigate the influence of misfitting item responses and found similar bounds. So, we recommend using an empirical dataset if your sample size is big enough.

Table 1. Item Parameter Values for the Simulated Test

Item	Difficulty (b)	Item	Difficulty (b)
1	-0.560	21	-1.068
2	-0.230	22	-0.218
3	1.559	23	-1.026
4	0.071	24	-0.729
5	0.129	25	-0.625
6	1.715	26	-1.687
7	0.461	27	0.838
8	-1.265	28	0.153
9	-0.687	29	-1.138
10	-0.446	30	1.254
11	1.224	31	0.426
12	0.360	32	-0.295
13	0.401	33	0.895
14	0.111	34	0.878
15	-0.556	35	0.822
16	1.787	36	0.689
17	0.498	37	0.554
18	-1.967	38	-0.062
19	0.701	39	-0.306
20	-0.473	40	-0.380

We can also use the empirical dataset at hand to determine the UB and LB. Instead of generating responses of examinees, we can sample a group of examinees with approximately similar $\hat{\theta}$ of the population. Then, as illustrated in steps 2 and 3, we calculate the UB and LB using extreme C^+ and C^- values of the subsample of examinees. There is a drawback of using empirical item responses instead of simulation: the existing misfitting item scores might affect boundary values. However, we expect most item score patterns will fit the underlying IRT model in reality. Meijer (2002) used both simulation and empirical datasets to investigate the influence of misfitting item responses and found similar bounds. So, we recommend using an empirical dataset if your sample size is big enough.

Data

We generated the latent abilities of 2,000 examinees (θ) from the standard normal distribution (mean=0, SD=1). 90% of examinees performed as expected without aberrant responses: item scores were generated based on the 1PL model. 5% of examinees (100) performances were affected by warm-up effects: the first 10 items' responses were generated based on their true θ - 2, and the remaining responses were based on true θ. The non-invariant ability setting was from van Krimpen-Stoop and Meijer (2000). The remaining 5% of examinees (100) had random responses: items scores of the first 30 items still followed the 1PL model. However, the probability of answering items correctly for the last 10 items was fixed to 0.2 (guess rate for 5-option multiple choice questions; Yamamoto & Everson, 1997).

Analyses

We used items' true parameters and examinees' estimated $\hat{\theta}$s to calculate the expected scores (equation 7) to mimic practical testing situations. The CUSUM method is based on the weighted residual, T_i (Equation 6) was applied to all 2,000 examinees in the sample. It should be noted that CUSUM's index is a statistic based on the cumulative sum. The CUSUM index is calculated continuously. When one item's C^+ or C^- exceeds the critical value, it does not mean the aberrant behavior happened near or at that specific item. Lai (2001) suggested that the location of an item with a CUSUM index closest to 0 before the critical value

should be taken as the change point. We used R (R Core Team, 2021) to analyze and draw CUSUM plots.

CUSUM Outputs

Normal Responses

Table 2 provides information on one score pattern without any aberrant flagging response. The true ability of this examinee is 0.637, and the estimated ability is 0.685. The first item is correct, and $P_1(\hat{\theta}) = .78$, which results in $T_1 = \frac{1}{N}\left[X_1 - P_1(\hat{\theta})\right] = \frac{1-.78}{40} = .0056$ (Equation 6). Substituting this value in Equation 4 results in C_1^+=0.0056 and Equation 5 results in C_1^-=0. Answering the second item correctly results in T_2 =.0071. $C_2^+ = \max\{0, C_1^+ + T_2\} = \max(0, 0.0056 + .0071) = 0.0127$, and $C_2^- = \min\{0,\ C_1^- + T_2\} = \min\{0, 0 + .0071\} = 0.$ This examinee makes an incorrect choice for the third item, $T_3 = -.0074$, $C_3^+ = \max\{0, C_2^+ + T_3\} = \max\{0.0.0127 + (-.0074)\} = 0.0054$, and $C_3^- = \min\{0, C_2^- + T_3\} = \min\{0, 0 + (-.0074)\} = -0.0074$. The procedure runs on both sides till the last item of a test.

The whole response answering process is in Figure 1. The item number and cumulative residual are represented by the horizontal and vertical axes, respectively. Two horizontal red dash lines show the UB (0.114) and the LB (-0.114), respectively. The largest value of C_1^+ is 0.0706 (item 23), and the smallest value of C_1^- is -0.054 (item 14), and neither is across boundary lines. So, this pattern does not contain an aberrant response.

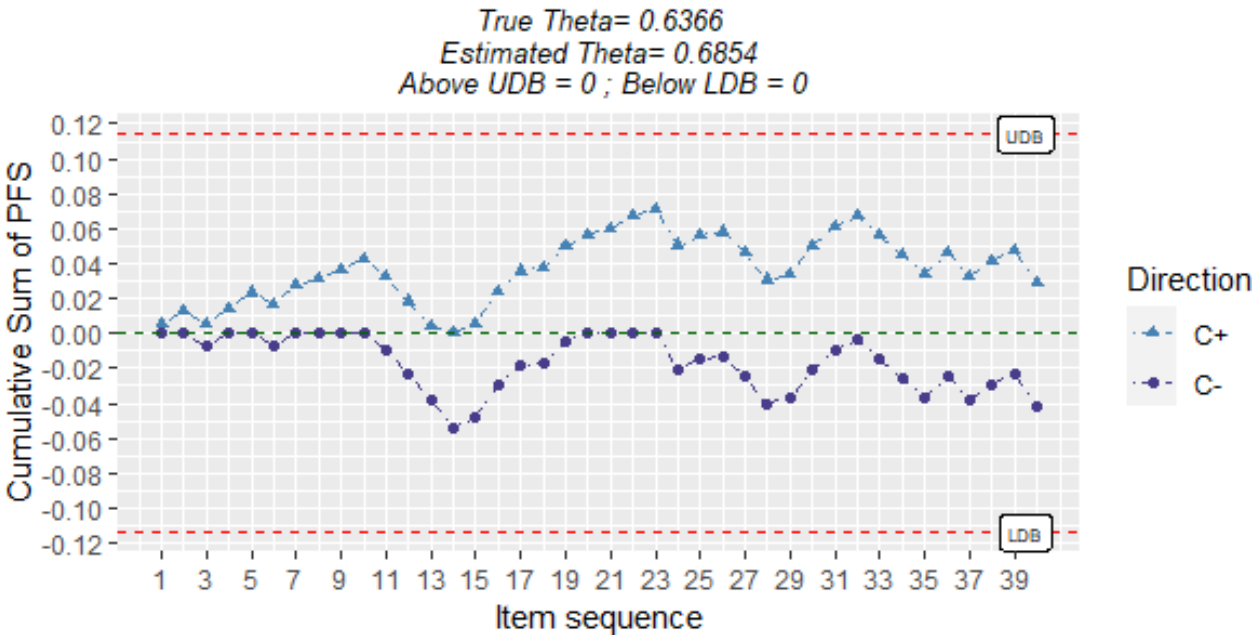
Table 2. CUSUM Procedure for a Regular Response Pattern

Item	Item score	P	T_i	C^+	C^-
1	1	0.7766	0.0056	0.0056	0
2	1	0.7141	0.0071	0.0127	0
3	0	0.2946	-0.0074	0.0054	-0.0074
4	1	0.6491	0.0088	0.0141	0
5	1	0.6356	0.0091	0.0233	0
6	0	0.2631	-0.0066	0.0167	-0.0066
7	1	0.5559	0.0111	0.0278	0
8	1	0.8755	0.0031	0.0309	0
9	1	0.7978	0.0051	0.0359	0
10	1	0.756	0.0061	0.042	0
11	0	0.3685	-0.0092	0.0328	-0.0092
12	0	0.5807	-0.0145	0.0183	-0.0237
13	0	0.5707	-0.0143	0.004	-0.038
14	0	0.6399	-0.016	0	-0.054
15	1	0.7758	0.0056	0.0056	-0.0484
16	1	0.2495	0.0188	0.0244	-0.0296
17	1	0.5467	0.0113	0.0357	-0.0183
18	1	0.9341	0.0016	0.0373	-0.0166
19	1	0.496	0.0126	0.0499	-0.004
20	1	0.761	0.006	0.0559	0
21	1	0.8524	0.0037	0.0596	0
22	1	0.7117	0.0072	0.0668	0
23	1	0.847	0.0038	0.0706	0
24	0	0.8044	-0.0201	0.0505	-0.0201
25	1	0.7876	0.0053	0.0558	-0.0148
26	1	0.9147	0.0021	0.058	-0.0127

27	0	0.462	-0.0115	0.0464	-0.0242
28	0	0.63	-0.0157	0.0307	-0.04
29	1	0.861	0.0035	0.0342	-0.0365
30	1	0.3616	0.016	0.0501	-0.0205
31	1	0.5644	0.0109	0.061	-0.0096
32	1	0.7272	0.0068	0.0678	-0.0028
33	0	0.4478	-0.0112	0.0566	-0.014
34	0	0.452	-0.0113	0.0453	-0.0253
35	0	0.466	-0.0117	0.0337	-0.037
36	1	0.4992	0.0125	0.0462	-0.0244
37	0	0.5328	-0.0133	0.0329	-0.0378
38	1	0.6786	0.008	0.0409	-0.0297
39	1	0.7294	0.0068	0.0477	-0.023
40	0	0.7438	-0.0186	0.0291	-0.0416

Note: P=probability of endorsing an item given $\hat{\theta}$.

Figure 1. The CUSUM Chart for a Response Pattern without Aberrant Response



Warm-up Effect

Table 3 provides information on one score pattern with a warm-up effect. This examinee's true ability (θ) is 1.222, while the estimated ability ($\hat{\theta}$) is -0.063. The lower estimated $\hat{\theta}$ value is probably because the first 10 items' responses were generated based on the $\theta_{warm-up} = \theta - 2 = -0.778$. The non-invariant ability setting for the warm-up effect severely underestimates this examinee's ability level. This example illustrates the negative influence of aberrant responses to ability estimation. And the column P is the probability of correctly answering items based on the examinee's estimated $\hat{\theta}$ and item parameters.

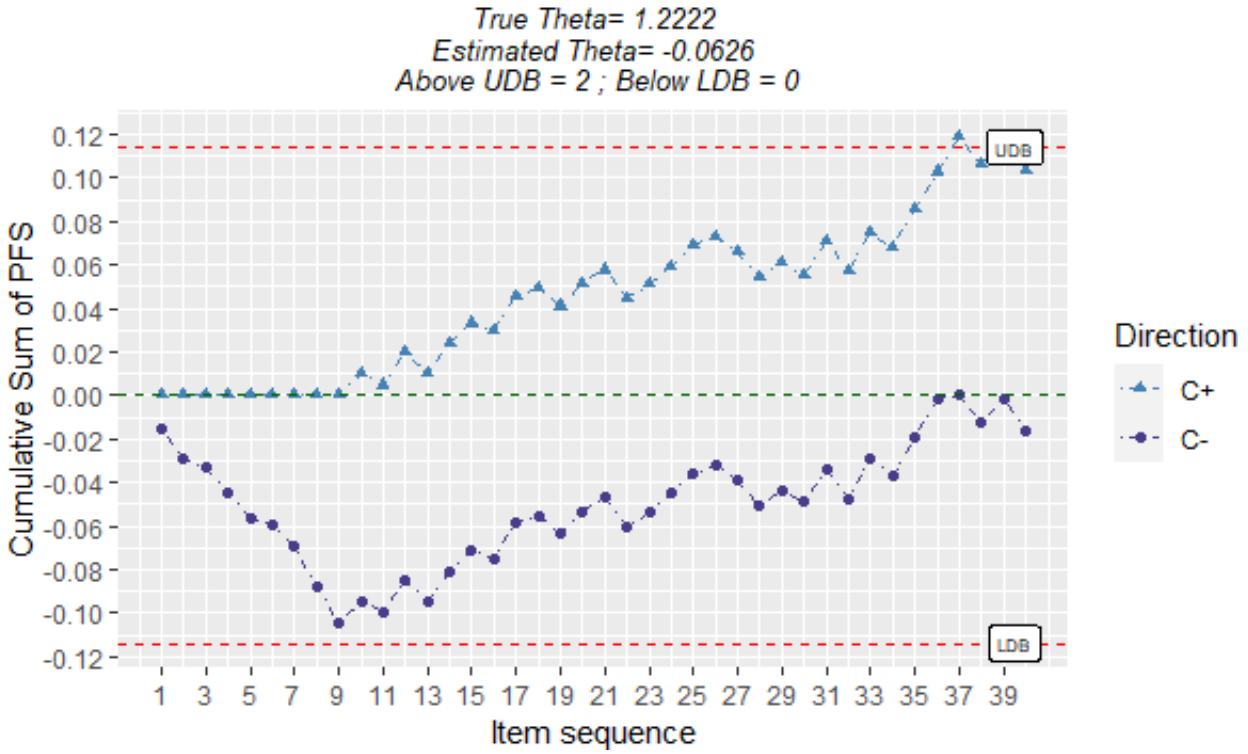
Figure 2 gives the CUSUM chart of this response pattern. This examinee answers all of the first 9 items incorrectly. These consecutive incorrect responses do not lead the C_i^- line across the LB. Then this examinee performs better in the middle and late stages of the test. We can see an upward trend from the chart after the beginning stage. It even leads to the last item above the UB. It is because the T_i, C^+, and C^- were calculated based on the underestimated $\hat{\theta}$. So, the first 10

consecutive items with abnormal errors are not cumulative enough for the C^- line to across the LB. However, the remaining expected responses accumulate a lot for the C^+ line. It is challenging to locate the change point solely based on the chart.

Table 3. CUSUM Procedure for a Warm-up Response Pattern

Item	Item score	P	T_i	C^+	C^-
1	0	0.622	-0.0155	0	-0.0155
2	0	0.5418	-0.0135	0	-0.0291
3	0	0.165	-0.0041	0	-0.0332
4	0	0.4668	-0.0117	0	-0.0449
5	0	0.4522	-0.0113	0	-0.0562
6	0	0.1446	-0.0036	0	-0.0598
7	0	0.372	-0.0093	0	-0.0691
8	0	0.769	-0.0192	0	-0.0883
9	0	0.6512	-0.0163	0	-0.1046
10	1	0.5946	0.0101	0.0101	-0.0945
11	0	0.2164	-0.0054	0.0047	-0.0999
12	1	0.3959	0.0151	0.0198	-0.0848
13	0	0.3862	-0.0097	0.0102	-0.0944
14	1	0.4568	0.0136	0.0238	-0.0809
15	1	0.6209	0.0095	0.0332	-0.0714
16	0	0.1359	-0.0034	0.0298	-0.0748
17	1	0.3634	0.0159	0.0457	-0.0589
18	1	0.8703	0.0032	0.049	-0.0556
19	0	0.3178	-0.0079	0.041	-0.0636
20	1	0.6011	0.01	0.051	-0.0536
21	1	0.7321	0.0067	0.0577	-0.0469
22	0	0.5388	-0.0135	0.0442	-0.0604
23	1	0.7238	0.0069	0.0511	-0.0535
24	1	0.6607	0.0085	0.0596	-0.045
25	1	0.637	0.0091	0.0687	-0.0359
26	1	0.8354	0.0041	0.0728	-0.0318
27	0	0.289	-0.0072	0.0656	-0.039
28	0	0.4462	-0.0112	0.0544	-0.0502
29	1	0.7456	0.0064	0.0608	-0.0438
30	0	0.2114	-0.0053	0.0555	-0.0491
31	1	0.3801	0.0155	0.071	-0.0336
32	0	0.5579	-0.0139	0.0571	-0.0475
33	1	0.2773	0.0181	0.0751	-0.0295
34	0	0.2808	-0.007	0.0681	-0.0365
35	1	0.2923	0.0177	0.0858	-0.0188
36	1	0.3206	0.017	0.1028	-0.0018
37	1	0.3506	0.0162	0.119	0
38	0	0.4998	-0.0125	0.1065	-0.0125
39	1	0.5606	0.011	0.1175	-0.0015
40	0	0.5788	-0.0145	0.103	-0.016

Figure 2. The CUSUM Chart for a Response Pattern with a Warm-up Effect



Random Response

Table 4 provides information on one score pattern with random responses. This examinee's true ability (θ) is 0.537, while the estimated ability ($\hat{\theta}$) is 0.043. The lower estimated $\hat{\theta}$ value is likely because the last 10 items' responses were generated based on the random guessing probability of 0.2. It is a mimic of rapid guessing behavior. The column P is the probability of correctly answering items based on the examinee's estimated $\hat{\theta}$ and true item parameters.

Figure 3 gives the CUSUM chart of this response pattern. This examinee performs well at the first 29 items. The line of C_1^+ even crossed over the UB at the 24th and 29th items. However, from the 30th item, all choices made by this examinee are wrong. The consecutive incorrect responses result in 1 item below the LB. We can observe a clear downward trend of C^- line from the 30th item. The change point of this examinee is around the 30th item based on Figure 3.

Discussion

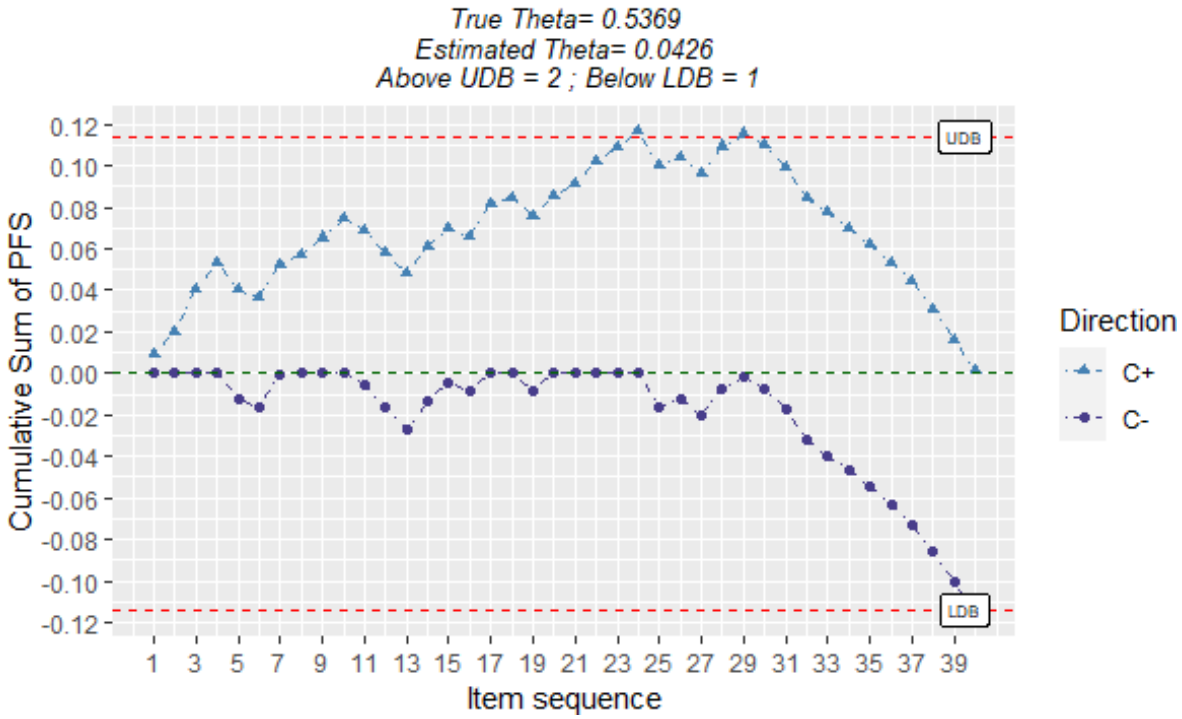
One of the most critical quality control tasks in testing is developing a mechanism to monitor examinees' behavior for potentially abnormal responses. Failure to address this problem may lead to inaccurate item and ability parameter estimations, biased equated scores, and misunderstanding of examinee performance. Researchers have developed many PFS in educational assessment fields to detect aberrant responses. Karabatsos (2003) compared the performances of 36 different PFS in detecting aberrant responses. However, these PFS indices were overall statistics based on all item responses. The positive (negative) residual in one place in the sequence might cancel a negative (positive) residual in another place.

SPC charts, widely used to monitor product quality in the manufacturing area, are also very promising in addressing the limitation of the overall PFS index. Many studies applied the CUSUM chart, one of the SPC charts, to detect aberrant responses (van Krimpen-Stoop & Meijer, 2000, 2001. 2002; Meijer, 2002; Armstrong & Shi, 2009a, 2009b; Tendeiro & Meijer, 2012). The CUSUM method calculates PFS for every item in the test. It identifies aberrant responses based on the cumulative values of the PFS. The visualization of the test-taking process can be provided immediately. It allows us to do a local inspection of the whole test-taking procedure. However, the CUSUM

Table 4. CUSUM Procedure for a Random Response Behavior

Item	Item score	P	T_i	C^+	C^-
1	1	0.6464	0.0088	0.0088	0
2	1	0.5678	0.0108	0.0196	0
3	1	0.18	0.0205	0.0401	0
4	1	0.493	0.0127	0.0528	0
5	0	0.4783	-0.012	0.0409	-0.012
6	0	0.1581	-0.004	0.0369	-0.0159
7	1	0.3969	0.0151	0.052	-0.0008
8	1	0.7871	0.0053	0.0573	0
9	1	0.6747	0.0081	0.0654	0
10	1	0.6197	0.0095	0.0749	0
11	0	0.2348	-0.0059	0.0691	-0.0059
12	0	0.4214	-0.0105	0.0585	-0.0164
13	0	0.4114	-0.0103	0.0483	-0.0267
14	1	0.483	0.0129	0.0612	-0.0138
15	1	0.6453	0.0089	0.0701	-0.0049
16	0	0.1488	-0.0037	0.0663	-0.0086
17	1	0.3881	0.0153	0.0816	0
18	1	0.8818	0.003	0.0846	0
19	0	0.341	-0.0085	0.0761	-0.0085
20	1	0.6261	0.0093	0.0854	0
21	1	0.7522	0.0062	0.0916	0
22	1	0.5648	0.0109	0.1025	0
23	1	0.7443	0.0064	0.1089	0
24	1	0.6839	0.0079	0.1168	0
25	0	0.661	-0.0165	0.1003	-0.0165
26	1	0.8493	0.0038	0.104	-0.0128
27	0	0.3111	-0.0078	0.0962	-0.0205
28	1	0.4723	0.0132	0.1094	-0.0073
29	1	0.7651	0.0059	0.1153	-0.0015
30	0	0.2295	-0.0057	0.1096	-0.0072
31	0	0.4052	-0.0101	0.0994	-0.0173
32	0	0.5836	-0.0146	0.0848	-0.0319
33	0	0.2989	-0.0075	0.0774	-0.0394
34	0	0.3025	-0.0076	0.0698	-0.047
35	0	0.3145	-0.0079	0.062	-0.0548
36	0	0.3439	-0.0086	0.0534	-0.0634
37	0	0.3749	-0.0094	0.044	-0.0728
38	0	0.5261	-0.0132	0.0308	-0.0859
39	0	0.5863	-0.0147	0.0162	-0.1006
40	0	0.6042	-0.0151	0.0011	-0.1157

Figure 3. The CUSUM Chart for a Response Pattern with a Random Responses



procedure based on PFS differs from the CUSUM procedure in production areas. The traditional CUSUM procedure has an underlying assumption: the variable to be investigated is approximately normally distributed. The distributions of PFS indices in educational assessments are usually far from normal. Therefore, the selection of boundary values for CUSUM charts based on PFS is not as straightforward as traditional procedures.

This article used a simulation study to demonstrate the procedure of CUSUM to detect aberrant response patterns. Two aberrant behaviors, warm-up effect, and random responses were used as examples. People who are not professionals in psychometrics can use the sequence of presenting information in the chart to identify various distinct kinds of unexpected response behaviors. Besides detecting aberrant responses, the CUSUM method can also be used as a diagnostic tool to identify compromised items. Lee and Lewis (2021) adopted the CUSUM method to detect items that might be exposed during continuous testing.

Despite all advantages previously discussed, the CUSUM method has some limitations. Firstly, to fully understand an examinee's test-taking process, we need to manually check the CUSUM control chart to locate the change point and identify potential aberrant behaviors. It can be time-consuming and laborious for a large sample. Secondly, if one response pattern contains too many aberrant responses, it might result in inaccurate ability estimates, as well as T_i, C^+, and C^- values. The performance of the CUSUM method might be negatively affected. Our warm-up effect example is in line with this situation. Hong and Cheng (2019) discussed this "masking effect" and how to use a robust estimation method to address the problem.

Identifying an aberrant response pattern is always contentious, especially in high-stakes testing. The CUSUM is a statistical method. Its classification of subjects, “with or without aberrant responses,” is a statistical inference that can only be used as a supplement to identify abnormal responding examinees. In addition to item scores, other sources of evidence, such as response time, examinee's self-reported surveys, and teacher evaluations, are needed to verify an abnormal respondent.

References

Allalouf, A., Gutentag, T., & Baumer, M. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional

module. *Educational Measurement: Issues and Practice, 36*(1), 58-68.

Armstrong, R. D., & Shi, M. (2009a). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*(4), 408–428.

Armstrong, R. D., & Shi, M. (2009b). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*(5), 391–410.

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language. *ETS Research Report Series, 1985*(1), i-57.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under Conditions of test speededness: Application of a Mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*(443), 910–919.

Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods, 51*(2), 573- 588.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica, 11*(2), 303–408.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29-37.

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement, 39*(3), 219-233.

Montgomery, D. C. (2013). *Introduction to statistical quality control* (7th ed.). Wiley.

Omar, M. H. (2010). Statistical process control charts for measuring and monitoring temporal consistency of ratings. *Journal of Educational Measurement, 47*(1), 18-35.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*(1–2), 100–115.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R- project.org/

Schafer, W. D., Coverdale, B. J., Luxenberg, H., & Ying, J. (2011). Quality control charts in large-scale assessment programs. *Practical Assessment, Research, and Evaluation, 16*(1), 15.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81*(4), 1118–1141.

Sinharay, S. (2017). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika, 82*(4), 1149-1161.

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36*(5), 420–442.

Tendeiro, J. N., Meijer, R. R., Schakel, L.,& Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement, 73*(1), 143–161.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In *Computerized adaptive testing: Theory and practice* (pp. 201-219). Springer, Dordrecht.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*(2), 199–217.

Wollack, J. A., & Cohen, A. S. (2004). *A model for simulating speeded test data* [Paper presentation]. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series, 1995(1)*, i-39.

Yu, X., & Cheng, Y. (2022). A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect test speededness. *Multivariate Behavioral Research*, *57(1),* 112-133.

Zhang, L., Wang, X., Cai, Y., & Tu, D. (2020). Change point analysis: A new method to detect aberrant responses in psychological and educational testing. *Advances in Psychological Science, 28*(9), 1462.

Citation:
Wan, S., & Keller, L.A. (2023). Using cumulative sum control chart to detect aberrant responses in educational assessments. *Practical Assessment, Research, & Evaluation*, *28*(2). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/2/

Corresponding Author:
Siyu Wan
University of Massachusetts Amherst

Email: siyuwan93 [at] outlook.com

Appendix A. Monte Carlo Simulation Codes

```
#This code use Monte Carlo simulation to identify the UB and LB for CUSUM#####

#IMPORTANT NOTES:

#packages library(mirt)

library(mirtCAT)

library(dplyr)

#INPUT (files that need to be in working directory)

# null

#OUTPUT (files will be generated)

# 1. "MonteCarlo.csv" - it provides UB and LB for each replication

# 2. LB and UB value

TL=40

N=10000

set.seed(123)

b.par <- rnorm(TL)

slopeint <- -b.par #mirt used slope IRT parameters

 #####CUSUM process##############

plus.CUSUM <- function(PFS){

 test.length=length(PFS)

 Cplus=rep(NA,test.length)

 Cplus[1]=max(0,PFS[1])

 for (k in 2:test.length){


  Cplus[k]=max(0,Cplus[k-1]+PFS[k],na.rm = TRUE)
```

```
 }
 Cplus
}
 mins.CUSUM <- function(PFS){
  test.length=length(PFS)
  Cmins=rep(NA,test.length)
  Cmins[1]=min(0,PFS[1])
  for (k in 2:test.length){
   Cmins[k]=min(0,Cmins[k-1]+PFS[k],na.rm = TRUE)
 }
 Cmins
 }
 #create a matrix to save UB LB values
 LB=UB=rep(NA,100)
 CriPoint <- cbind(LB,UB)
 CriPoint <- as.data.frame(CriPoint)
 #response matrix
 Res <- matrix(NA,nrow = TL,ncol = 8)
 colnames(Res) <-
 c("id","theta","b","score","P","PFS","Cplus","Cmins")
  Res[,"b"]=b.par
 Res <- as.data.frame(Res)
 #matrix saving extremest CUSUM values for 1 person of 1 replication
 ExtreValue <- matrix(NA,nrow = N,ncol =4 )
```

```
colnames(ExtreValue) <- c("ID","Theta","cmax","cmin")
ExtreValue <- as.data.frame(ExtreValue)
# replicate 100 times
for (t in 1:100) {
 true.theta =rnorm(N) #each replication, randomly generate 10000
 thetas
dataset <-
      simdata(a=rep(1,TL),d=slopeint,itemtype="dich",guess=rep(0,TL),
      upper=rep(1,TL),
     Theta=true.theta)
#fixed ability estimation
pars <- data.frame(a1 =rep(1,TL),
          d=slopeint)
mod <- generate.mirt_object(pars, itemtype = '2PL')
# trait scores for pattern
theta.est <-fscores(mod, response.pattern = dataset,method="MAP")[,1]
for (i in 1:N) {
 #response matrix for each person
 Res <- Res%>% mutate(id=i,
     theta=theta.est[i],
     P=1/(1+exp(b-theta)),
     score=dataset[i,],
     PFS=(score-P)/TL,
     Cplus=plus.CUSUM(PFS),
```

```
      Cmins=mins.CUSUM(PFS))

# extreme values for each replication

ExtreValue[i,] <- Res%>%summarise(ID=first(id),

              Theta=first(theta),

              cmax=max(Cplus),

              cmin=min(Cmins))

}

CriPoint[t,] <- ExtreValue%>%summarise(LB=quantile(cmin,0.025),

                        UB=quantile(cmax,0.975))

print(t)

}

write.csv(CriPoint,file="MonteCarlo.csv")

#it tells us the value of LB and UB

apply(CriPoint,2,FUN=mean)%>%round(digits = 3)
```

Appendix B. CUSUM Plot Codes

```
UB=0.114
LB=0.114
library(ggplot2)
library(scales)
library(reshape2)
########################plot result#######################################
Plot.PFS <- function(n.order){ #need to specify the working directory
   of response files firstly
 filename=paste("Res",n.order,".csv",sep = "")

 Per.Res <- read.csv(filename)
 target= 0 true.theta=Per.Res$true.theta[1]
 est.theta=Per.Res$est.theta[1]
 widedate <- Per.Res[,c("X","Cplus","Cmins")]
 names(widedate) <- c("ITEM_SEQUENCE","C+","C-")
 N.up = sum(widedate$`C+`>UB)
 N.bot = sum(widedate$`C-`<LB)
 maintitle <- ("Cusum Chart")
 sub.t = paste("\n True Theta=",true.theta,"\n Estimated
       Theta=",est.theta, "\n Above UDB =",N.up,"; Below LDB
       =",N.bot,sep=" ")
 longdate <- melt(widedate,id.vars = 'ITEM_SEQUENCE', variable.name =
        'Direction', value.name = 'CUSUM')
 #######using ggplot to draw the plot
 ggplot(longdate,aes(x = ITEM_SEQUENCE, y = CUSUM, shape =
         Direction, color = Direction))+
  #draw the point and connected using lines
  geom_point()+
```

```
  scale_color_manual(values = c("C+" = 'steelblue','C-' =
  'darkslateblue')) + scale_shape_manual(values = c('C+' = 17,
  'C-' = 16))+ geom_line(linetype="dotdash")+
  #put UB and LB
  geom_hline(yintercept=c(UB,LB), linetype="dashed", color = "red")+
  geom_hline(yintercept=target, linetype="dashed", color =
  "darkgreen")+ geom_label(label="UDB", x=39.5,y=UB,size =2,color =
  "black")+ geom_label(label="LDB", x=39.5,y=LB,size =2,color =
  "black")+


  #define the title and x y lables
  ggtitle(maintitle,subtitle = sub.t)+
  xlab("Item sequence")+ ylab("Cumulative Sum of PFS")+
  theme(plot.title = element_text(face="bold",hjust =
  0.5,size = 11),
    plot.subtitle = element_text(face="italic",hjust = 0.5,size = 10))+

  #change the scale of x and y axis
  scale_x_continuous(breaks = seq(1, 40, by = 2))+
  scale_y_continuous(breaks =
  scales::pretty_breaks(n = 10))
}
```