Item Ordering and Computerized Classification Tests With Cluster-Based Scoring: An Investigation of the Countdown Method

Finkelman, M.D.; Lowe, S.R.; Kim, W.; Gruebner, O.; Smits, N.; Galea, S.

DOI
10.1037/pas0000470

Publication date
2018

Document Version
Final published version

Published in
Psychological Assessment

Psychological Assessment
2018, Vol. 30, No. 2, 204–219

1040-3590/18/$12.00 http://dx.doi.org/10.1037/pas0000470

Item Ordering and Computerized Classification Tests With Cluster-Based Scoring: An Investigation of the Countdown Method

Matthew D. Finkelman
Tufts University School of Dental Medicine

Sarah R. Lowe
Montclair State University

Wonsuk Kim
Measured Progress, Inc., Dover, New Hampshire

Oliver Gruebner
Robert Koch Institute, Berlin, Germany

Niels Smits
University of Amsterdam

Sandro Galea
Boston University School of Public Health

The countdown method is a well-known approach to reducing the average length of screening instruments that are presented by computer. In the countdown method, testing is terminated once the result of the screener ("positive" or "negative") has been unambiguously determined from prior answers. Previous research has examined whether presenting dichotomously scored items in order from "least to most frequently endorsed" or "most to least frequently endorsed" is more efficient when the countdown method is used. The current study describes the Mean Score procedure, an extension of the above item ordering procedures to polytomously scored items, and evaluates its efficiency relative to the distribution of other possible item orderings in 2 real-data simulations. Both simulations involve item responses to the Posttraumatic Stress Disorder (PTSD) Checklist for *DSM–5* (PCL-5). In the first simulation, items were scored polytomously, and a single cutoff point was used to determine the screening result. In the second simulation, items were converted to dichotomous scores, as well as categorized into 4 clusters; a positive result for the entire assessment was obtained if and only if a positive result was obtained for each cluster. The latter simulation also investigated the effect of reordering the clusters themselves on the efficiency of the countdown method. Results indicated that the Mean Score procedure does not necessarily produce the optimal ordering, but tends to assemble an efficient item ordering relative to the distribution of possible orderings. In the second simulation, reordering the clusters themselves affected efficiency. Future research directions are suggested.

Public Significance Statement
This study suggests that by reordering a questionnaire's items, the burden on a respondent can be reduced when the questionnaire is administered by computer.

Keywords: respondent burden, computerized adaptive testing, countdown method, PCL-5

The last several decades have seen a rise in the number of studies in the medical literature involving questionnaires (Walter, 2010). One of the most common uses of medical and psychological questionnaires is to classify respondents, that is, to place each respondent into one of multiple mutually exclusive categories. Instruments that are used as "screeners" for conditions, disorders, or risk often classify respondents into one of only two categories: "positive" or "negative." Applications of screening instruments are diverse, including the assessment of depression (Radloff, 1977), pathological buying (Müller, Trotzke, Mitchell, de Zwaan, & Brand, 2015), and risk of aberrant medication-related behaviors among pain patients being considered for opioid therapy (Butler, Fernandez, Benoit, Budman, & Jamison, 2008), to name a few examples.

One element to consider when developing or selecting a questionnaire for operational usage is the amount of burden that the questionnaire will place on respondents. In particular, a key practical consideration is the length of an instrument, given that the use of longer questionnaires may result in greater respondent burden (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993). Question-

This article was published Online First March 16, 2017.

Matthew D. Finkelman, Department of Public Health and Community Service, Tufts University School of Dental Medicine; Sarah R. Lowe, Department of Psychology, Montclair State University; Wonsuk Kim, Measured Progress, Inc., Dover, New Hampshire; Oliver Gruebner, Epidemiology and Health Monitoring, Unit Social Determinants of Health, Robert Koch Institute, Berlin, Germany; Niels Smits, Research Institute of Child Development and Education, University of Amsterdam; Sandro Galea, Boston University School of Public Health.

Correspondence concerning this article should be addressed to Matthew D. Finkelman, Department of Public Health and Community Service, Tufts University School of Dental Medicine, 1 Kneeland Street, Boston, MA 02111. E-mail: matthew.finkelman@tufts.edu

naires with more items may also result in increased administrative burden incurred by providers and their colleagues (Dugdale, Epstein, & Pantilat, 1999; Finkelman et al., 2015).

One modern approach to reducing the respondent and administrative burden associated with a screener is to employ *computerized classification testing* (CCT; Thompson, 2007; Weiss & Kingsbury, 1984). As its name suggests, CCT involves the administration of an assessment via computer; additionally, CCT prescribes the tracking of a given respondent's answers as testing is underway in order to determine the appropriate test length for the respondent in real time. Hence, CCT falls under the umbrella of *variable-length testing*, that is, testing that produces different test lengths for different respondents. The test length of a given respondent is typically based on need: those respondents whose classifications can be identified quickly receive shorter tests (thus reducing the burden of the assessment), whereas those respondents who are more difficult to classify receive longer tests. CCT has been shown to reduce test lengths without compromising measurement precision (Thompson, 2007).

Much methodological research on conducting efficient CCTs has focused on assessments that use Item Response Theory (IRT) to model the relation between the trait being measured and the items being administered (e.g., Thompson, 2007; van Groen, 2014; Weiss & Kingsbury, 1984). However, not all tests are based on scoring rules that utilize IRT modeling. In particular, little research has been conducted on how to conduct efficient CCT's alongside *cluster-based scoring,* in which (a) a given assessment is divided into clusters of items that have similar or related content; (b) a respondent is said to receive a positive result for a given cluster if and only if the respondent obtains a certain score (or higher) on that cluster; and (c) the respondent is said to receive a positive result for the assessment as a whole if and only if she receives a positive result for every cluster. The Posttraumatic Stress Disorder (PTSD) Checklist for *DSM–5* (PCL-5), which will be described in a later section, will serve as an example of an assessment alongside which cluster-based scoring can be used.

One "special case" of cluster-based scoring is the use of a test that consists of only one cluster. In this case, a positive result for the entire assessment is obtained if the respondent's total score meets (or exceeds) a specified cutoff point, and a negative result is obtained otherwise. When such a test is administered via computer, CCT may be employed to enhance the efficiency of assessment. In particular, the *countdown method* has been proposed to halt a computer-based test as soon as the outcome of that test has become deterministic (e.g., Ben-Porath, Slutske, & Butcher, 1989; Butcher, Keller, & Bacon, 1985). For example, consider a questionnaire that consists of 25 items, each of which is scored dichotomously as either 0 (*not endorsed*) or 1 (*endorsed*). Suppose that at least 18 of the items must be endorsed in order for a positive result to be obtained. Once a given respondent has endorsed 18 items, the outcome of the test has been determined as positive with certainty; the countdown method would account for this fact by immediately terminating the questionnaire with a positive result. Conversely, once eight "not endorsed" answers have been elicited from a different respondent, the outcome of the test has been determined as negative with certainty; the countdown method would account for this fact by immediately terminating the questionnaire with a negative result. The countdown method is also applicable to tests whose items are *polytomous* (i.e., items that have more than two answer choices, such as those scored on a 0–4 scale). Indeed, in the case of polytomous items, the logic of the countdown method remains the same: the test is terminated with a positive result if the cutoff point has been reached; the test is terminated with a negative result if it has become impossible for the respondent to reach the cutoff point, due to low scores on previous items. Previous studies have suggested the potential of the countdown method, which may also be referred to as a *curtailed test,* to substantially reduce the mean number of items administered to respondents (e.g., Ben-Porath et al., 1989; Forbey, Ben-Porath, & Arbisi, 2012; Roper, Ben-Porath, & Butcher, 1991). As will be seen in a later section, one scoring rule of the PCL-5 treats the instrument's items polytomously and compares a respondent's total score to a single cutoff point.

A natural question is how to order the items within a questionnaire to enhance the degree of efficiency gains offered by the countdown method (Ben-Porath et al., 1989). That is, it may be desired to place the items in the order that provides the greatest reduction in the mean number of items administered. For the Minnesota Multiphasic Personality Inventory-2 (MMPI-2), which contains dichotomous items, Ben-Porath et al. (1989) compared several item orderings based on the logic that a large reduction of items may generally be achieved by either administering items from "least to most frequently endorsed" or "most to least frequently endorsed" (they also took into account considerations that are specific to the MMPI-2 in determining the item order; see Ben-Porath et al. for details). They found that the least to most frequently endorsed ordering provided greater item savings than the most to least frequently endorsed ordering. Based on their results, subsequent research (e.g., Forbey et al., 2012; Forbey, Ben-Porath, & Gartland, 2009; Handel, Ben-Porath, & Watt, 1999; Roper et al., 1991; Roper, Ben-Porath, & Butcher, 1995) utilized the least to most frequently endorsed ordering alongside the MMPI-2 (again allowing for considerations specific to this assessment to be accounted for when finalizing the item ordering). A review paper on computerized adaptive testing in personality assessment (Forbey & Ben-Porath, 2007) also highlighted Ben-Porath et al.'s (1989) comparison of the least to most frequently endorsed and most to least frequently endorsed item ordering methods. Rudick, Yam, and Simms (2013) compared the above two item ordering methods to each other as well as to an individually randomized order and the conventional booklet order of the Schedule for Nonadaptive and Adaptive Personality (SNAP) scales. The authors found that among these specific item-ordering methods, the least to most frequently endorsed ordering produced the greatest item savings for the SNAP. To our knowledge, however, there is a dearth of research focused on evaluating the efficiency of the least to most frequently endorsed and most to least frequently endorsed methods in comparison to the distribution of other possible item orderings, under multiple test conditions, despite the prominence of the issue of item ordering in the literature. Because the above item ordering methods are not always optimal (Yang, Miao, Tian, Liu, & Zhu, 2009), a comprehensive evaluation of their performance compared to other item orderings is warranted.

Although the countdown method has most typically been applied to tests that have only one cluster, it can be extended to the case of multiple clusters. In particular, when an assessment comprises more than one cluster, the countdown method serves two

functions: (a) it determines when a respondent can be "skipped from cluster to cluster" (namely, when the result of one cluster has been unambiguously determined from previous answers), and (b) it determines when the entire assessment can be terminated (namely, when the result of the entire assessment has been unambiguously determined from previous answers). A previous study (Finkelman et al., 2016) showed that the countdown method, when used in the above manner, has the potential to substantially reduce the average test length of the PCL-5, which has four clusters. However, to our knowledge, the effect of item ordering on the countdown method's ability to shorten tests has not been studied in the case of multiple clusters. Indeed, the Finkelman et al. study examined the effectiveness of the countdown method and a similar approach (which they referred to as *curtailment* and *stochastic curtailment,* respectively) to lessen the respondent burden of the PCL-5. However, it only studied these procedures in the context of administering the items in their standard order (i.e., the "conventional booklet" item ordering of the PCL-5). Considering the results of Ben-Porath et al. (1989) for the MMPI-2, as well as the results of Rudick et al. (2013) for the SNAP (both of which indicated that the item ordering affects the item savings of the countdown method), the lack of study on item ordering for the PCL-5 constitutes a gap in the literature for this screener. If it were found that the efficiency of the PCL-5 could be enhanced by reordering its items alongside the countdown method, this information could be used to reduce the respondent burden of individuals taking the assessment. On the other hand, if it were found that the standard PCL-5 ordering were as efficient as other candidate orderings, practitioners could confidently administer the screener in its conventional booklet order. The current study is the first research, both in the single-cluster case and the multiple-cluster case, to examine whether the item ordering influences the efficiency of the PCL-5 when coupled with the countdown method. Therefore, it fills the aforementioned gap in the PCL-5 literature. Moreover, the study extends the item-ordering procedure considered in Ben-Porath et al. (1989) to the case of polytomous items, and comprehensively compares this procedure with other item orderings. Such comprehensive examination and comparison of item orderings has methodological significance that carries implications not only for the PCL-5, but also for other screeners that can be coupled with the countdown method.

Therefore, the objective of this research was to investigate the effect of item ordering on the ability of the countdown method to shorten assessments that use cluster-based scoring. As noted above, the item orderings considered in Ben-Porath et al. (1989) were extended to the case of polytomous items and compared to other item orderings. In simulation studies using data from the PCL-5, two scoring rules were examined: one in which the items were treated as a single cluster (with a single cutoff point), and one in which the items were divided into separate clusters (with a positive result for the entire assessment occurring if and only if each cluster's result was positive). Both of these scoring rules have been suggested for operational use with the PCL-5 (Weathers et al., 2013). For the scoring rule with multiple clusters, the effect of the ordering of the clusters themselves was investigated, in addition to the effect of the ordering of items within each cluster. The results may help practitioners to make informed decisions about item ordering when using the countdown method to enhance the efficiency of an assessment.

Simulation Study 1: The Case of One Cluster

The purpose of Simulation Study 1 was to examine the performance of different item orderings, combined with the use of the countdown method, in the scenario where items are scored as a single cluster. In particular, a real-data simulation was conducted to evaluate the effect of item ordering on the ability of the countdown method to reduce test length. The real-data simulation involved determining the average number of items that would have been presented, if the countdown method had been used, among a set of respondents who had previously taken an assessment. As will be seen, this calculation was repeated for different item orderings, and results were compared.

Method

The assessment used in the analysis was the PCL-5, a 20-item self-report screener for PTSD (Blevins, Weathers, Davis, Witte, & Domino, 2015; Bovin et al., 2016; Weathers et al., 2013; Wortmann et al., 2016). Each item asks how much the respondent was bothered by a specific problem during the last month and is scored from 0 to 4 (*Not at all* = 0, *A little bit* = 1, *Moderately* = 2, *Quite a bit* = 3, *Extremely* = 4). We note here that the PCL-5, being a screening device, is not thought to yield a definitive diagnosis of PTSD, as could be derived from a "gold standard" clinician-administered interview, such as the widely used Clinician Administered PTSD Scale as defined in the *DSM–IV* (CAPS; Weathers, Ruscio, & Keane, 1999). Nonetheless, the PCL for *DSM–IV* (Weathers, Litz, Herman, Huska, & Keane, 1993) has been validated against the CAPS (cf., Blanchard, Jones-Alexander, Buckley, & Forneris, 1996; Goldmann et al., 2011), and used in several epidemiologic studies, including studies in the aftermath of large-scale natural disasters, as a means of determining a provisional or probable PTSD diagnosis (e.g., Horesh, Lowe, Galea, Uddin, & Koenen, 2015; Tracy, Norris, & Galea, 2011). One psychometric evaluation to date sought to determine the cutoff point on the PCL-5 to optimize its sensitivity, specificity, and efficiency relative to the *DSM–IV* PCL (Blevins et al., 2015), suggesting that the PCL-5 is intended to yield provisional or probable PTSD diagnoses as well.

In one scoring rule proposed by Weathers et al. (2013), the individual item scores are summed in order to produce a total symptom severity score, with a minimum possible score of 0 and a maximum possible score of 80. The total symptom severity score is then compared to a cutoff point; a cutoff point of ≥33 has preliminarily been suggested, pending further psychometric study (Weathers et al., 2013). As will be described in a later section, the PCL-5's items can also be grouped into clusters; however, because the above scoring rule of Weathers et al. does not utilize the clusters, they are not relevant to the simulation study presented in this section.

The countdown method is easily applicable to the PCL-5 when the above scoring method is used. For example, suppose that a ≥33 cutoff point is employed. If the respondent's cumulative score reaches 33 or above at any stage of testing, the result of the screener has been unambiguously determined; hence, the countdown method immediately stops the assessment. Additionally, if the respondent's cumulative score at a given point in the assessment is so low that a total score of 33 or higher has become

mathematically impossible (given the respondent's previous answers), the countdown method immediately stops the assessment.

As discussed previously, Ben-Porath et al. (1989) investigated both the ordering of items from least to most frequently endorsed and from most to least frequently endorsed when using the countdown method. These item orderings are appropriate for items that are scored dichotomously as "endorsed" or "not endorsed." For polytomous items (such as those following the 0–4 scale described above), an extension would be to order the items either from "lowest to highest mean score" or "highest to lowest mean score." The former item ordering method is favorable to the latter when a negative result (i.e., a score below the cutoff point) tends to be obtained more quickly than a positive result (i.e., a score at or above the cutoff point). For example, if the cutoff point is very high along the scale, and a high percentage of respondents score below the cutoff point, then it may be hypothesized that more items would be eliminated by ordering the items from lowest to highest mean score (because this item ordering would tend to produce low scores at the beginning of the assessment, thus ruling out a positive result more quickly). Conversely, ordering the items from highest to lowest mean is favorable when a positive result tends to be obtained more quickly than a negative result. For example, if the cutoff point is very low along the scale, and a high percentage of respondents score at or above the cutoff point, then the highest to lowest mean ordering would be expected to eliminate more items (because this ordering would tend to produce high scores at the beginning of the assessment, resulting in respondents reaching the cutoff value quickly). Because either ordering (lowest to highest mean or highest to lowest mean) may potentially provide a greater item reduction than the other, a prudent approach would be to determine the average reduction provided by each, then select the ordering that produces the larger average. Specifically, suppose we have a training (pilot) dataset of respondents who have completed the entire assessment. The following steps are then undertaken, using the training dataset: (a) calculate the mean score of each item; (b) determine the average item reduction that would have been obtained from the countdown method by ordering the items from lowest to highest mean score; (c) determine the average item reduction that would have been obtained from the countdown method by ordering the items from highest to lowest mean score; and (d) select the ordering (either that of step (b) or that of step (c) above) that produced a greater average reduction. The item ordering selected in step (d) is then used operationally for subsequent respondents alongside the countdown method. The above procedure for selecting an item ordering (whereby either the lowest to highest mean score ordering or the highest to lowest mean score ordering is selected) will be referred to as the *Mean Score procedure*. Note that the detail of whether the lowest to highest mean score or the highest to lowest mean score ordering is ultimately chosen by the Mean Score procedure is suppressed in the title of this procedure; the title simply refers to the approach of evaluating both item orderings and selecting the one with the greater average item reduction. A hypothetical example in which the Mean Score procedure does not provide the optimal item ordering is given in the Appendix. The fact that such an example exists illustrates why a comparison of the Mean Score procedure with other item orderings is needed: to determine the ability of this procedure to reduce test lengths relative to the distribution of other possible orderings, and thereby evaluate its effectiveness in enhancing test efficiency alongside the countdown method.

The real-data simulation included responses from $n = 942$ participants from New York City (NYC) who had taken the PCL-5 following Hurricane Sandy. Data were originally obtained for a study on psychological resilience; see Lowe, Sampson, Gruebner, and Galea (2015) for detailed information about the study that produced the item responses. Briefly, adult participants (age ≥18 years) from two sampling zones consisting of NYC census tracts that had been most severely affected by Hurricane Sandy were surveyed via telephone using a computer-assisted interview system. The first zone included tracts in which at least 50% of the area was inundated with floodwater. The second zone included tracts that were adjacent to the tracts of the first zone and/or had less than 50% (but more than 0%) of the area inundated with floodwater. Half of the participants in each zone were recruited through address-based sampling, wherein households were recruited by mail and landline telephone, and one adult from each household was chosen at random for inclusion. The remaining participants were recruited via random-digit dialing of cellular phones; a geographic screening was used to approximate whether potential participants resided within the sampling zones when the hurricane occurred. The Institutional Review Board of Columbia University and the Institutional Review Board at Tufts Medical Center and Tufts University Health Sciences Campus approved the original study and granted exempt status for the retrospective data analysis.

Multiple item orderings were compared in the real-data simulation. The mean and standard deviation of the number of items administered were calculated for (a) the standard PCL-5 item ordering, (b) the ordering produced by the Mean Score procedure, and (c) 10,000 random item orderings. The 10,000 random item orderings were used to gauge the performance of the standard PCL-5 ordering and the Mean Score procedure with respect to other item orderings; obtaining results for all possible orderings was computationally infeasible due to the astronomically high number of possible orderings (namely, *20!* orderings, where the exclamation point represents the "factorial" operation).

To examine the effect of the cutoff point on the comparison of item orderings, results were obtained using six different cutoff points. Five of the cutoff points were selected in order to vary the percentage of positive results that would be obtained in the sample. In particular, the five cutoff points were chosen so that the percentage of positive results would be as close as possible to 5%, 10%, 15%, 20%, and 25%. The sixth cutoff point (≥33) was the value that had been suggested previously for the PCL-5 pending further psychometric study (Weathers et al., 2013). Aside from the ≥33 value (which itself is preliminary), the cutoff points employed herein are for illustration only and are not necessarily suitable for operational usage of the PCL-5. We note that the countdown method was previously applied to the PCL-5 using this dataset (Finkelman et al., 2016), but not for the purpose of comparing different item orderings to one another.

In the real-data simulation, the complete dataset ($n = 942$) was initially used in the training of the Mean Score procedure. Such training involved performing the aforementioned steps that are used to define the ordering selected by the Mean Score procedure: calculating the mean score of each item, determining the average item reduction of the countdown method when ordering items from lowest to highest mean, repeating the previous step with

items ordered from highest to lowest mean, and then selecting the ordering with the larger average reduction. After such training had been conducted, evaluation of the different item orderings was performed, again using the complete dataset (n = 942). In particular, the mean and standard deviation of the number of items administered were calculated for each item ordering. Employing the complete dataset in both training and evaluation enhances the precision of the results by using all available data; however, it may produce the "capitalization on chance" problem, whereby results are biased toward positive findings and may not be generalizable to subsequent test administrations (Hastie, Tibshirani, & Friedman, 2009). Therefore, a second analysis was undertaken whereby a randomly selected subset of two thirds of the data (n = 628) was used to train the Mean Score procedure. The remaining data (n = 314) were used to evaluate the item ordering that had been produced for the Mean Score procedure in the training process (as well as evaluating the other item orderings under comparison). By separating the training and evaluation data sets from each other, the capitalization on chance problem is avoided (Hastie et al., 2009). A computer program was written in R (Version 3.1.2; R Core Team, 2015) to obtain all results.

Results

Descriptive statistics about this dataset were previously provided by Finkelman et al. (2016). Briefly, among the 942 participants, 589 were female (63%). Of the 933 participants with marital status information, 354 were married (38%). The mean (*SD*) age was 50.2 (17.4) years among the 929 participants with age information. The mean (*SD*) PCL-5 total score was 6.0 (10.9); the median (interquartile range) was 1.0 (7.0). Four hundred three participants (42.8%) had a total score of zero. The highest score, which was obtained by one participant (0.1%), was 80.

The five cutoff points chosen to produce positive results in approximately 5%, 10%, 15%, 20%, and 25% of the sample were as follows, in respective order: ≥29 (producing 5.0% positive results in the sample); ≥19 (9.9% positive); ≥14 (14.9% positive); ≥10 (20.5% positive); and ≥7 (25.1% positive). The ≥33 cutoff point of Weathers et al. (2013) produced a positive result for 3.7% of the sample.

Table 1 presents statistics about the performance of different item orderings in the complete PCL-5 dataset. Specifically, the table compares the number of items administered by the standard PCL-5 ordering, the Mean Score procedure, and the best of the 10,000 random orderings (i.e., the random ordering with the lowest mean number of items administered). At four of the six cutoff points (≥7, ≥10, ≥14, and ≥19), the Mean Score procedure selected the highest to lowest mean score ordering, whereas at the other two cutoff points (≥29 and ≥33), it chose the lowest to highest mean score ordering. The Mean Score procedure's mean number of items administered was always lower than the mean number of items administered by the standard PCL-5 ordering; the difference in means between them ranged from 0.05 items to 0.65 items across the six cutoff points. At five of the six cutoff points, the Mean Score procedure exhibited a lower mean number of items administered than all 10,000 random item orderings. The exception was the ≥19 cutoff point, at which the Mean Score procedure's mean number of items administered was 0.01 higher than that of the best random ordering (16.11 vs. 16.10). At all cutoff points, the standard PCL-5 ordering exhibited a higher mean number of items administered than the best random ordering. Overall, when averaging the results of Table 1 across the six cutoff points, the Mean Score procedure reduced the mean number of items administered by an average of 0.28 items when compared to the standard PCL-5 ordering, and an average of 0.03 items when compared to the best random ordering (data not shown).

Table 2 presents results of the training-test analysis. As in the complete-dataset analysis, the Mean Score procedure selected the highest to lowest mean score ordering at the four lowest cutoff points, but not the two highest cutoff points. In the training dataset, the Mean Score procedure exhibited a lower mean number of items administered than both the standard PCL-5 ordering and the best random ordering at all six cutoff points. In the test dataset, the Mean Score procedure still outperformed the standard PCL-5 ordering at all cutoff points; however, the best random ordering exhibited a lower mean number of items administered than the Mean Score procedure at three cutoff points (≥14, ≥19, and ≥33) and tied the Mean Score procedure at an additional cutoff point (≥29). At the ≥14 cutoff, the Mean Score procedure was outperformed by 42 of the 10,000 random orderings (it tied with the 43rd best ordering); at the ≥19 cutoff, the Mean Score procedure was outperformed by 891 random orderings (it tied with the 892nd best ordering); at the ≥33 cutoff, the Mean Score procedure was outperformed by one random ordering (it did not tie with any random ordering). The difference between the Mean Score procedure's mean number of items administered, and that of the best random ordering, was never more than 0.12 items in the test

Table 1
Performance of Different Item Ordering Procedures in the Complete PCL-5 Dataset: Simulation Study 1 (n = 942)

	Standard PCL-5 ordering		Mean Score procedure		Best of 10,000 random orderings	
Cutoff point	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items
≥7	16.73	5.07	16.08	5.81	16.16	5.66
≥10	16.83	3.94	16.47	4.39	16.51	4.30
≥14	16.75	2.89	16.49	3.25	16.55	3.14
≥19	16.16	2.07	16.11	2.28	16.10	2.04
≥29*	14.01	1.70	13.85	1.59	13.87	1.65
≥33*	13.08	1.71	12.89	1.63	12.90	1.67

Note. PCL-5 = Posttraumatic Stress Disorder (PTSD) Checklist-5.
* For the ≥29 and ≥33 cutoff points, the Mean Score procedure ordered items from lowest to highest mean score. For all other cutoff points, the Mean Score procedure ordered items from highest to lowest mean score.

Table 2
Performance of Different Item Ordering Procedures in the Training (n = 628) and Test (n = 314) PCL-5 Datasets: Simulation Study 1

	Standard PCL-5 ordering				Mean Score procedure				Best of 10,000 random orderings			
	Training		Test		Training		Test		Training		Test	
Cutoff point	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items
≥7	16.79	5.04	16.62	5.14	16.23	5.72	15.77	6.00	16.29	5.57	15.87	5.90
≥10	16.89	3.87	16.69	4.08	16.52	4.37	16.31	4.51	16.60	4.10	16.33	4.48
≥14	16.76	2.87	16.75	2.94	16.55	3.17	16.49	3.19	16.59	3.06	16.43	3.37
≥19	16.19	2.05	16.10	2.12	16.13	2.25	16.06	2.33	16.14	2.09	15.94	2.33
≥29*	14.02	1.72	13.99	1.67	13.82	1.59	13.88	1.58	13.86	1.61	13.88	1.65
≥33*	13.07	1.71	13.12	1.73	12.83	1.59	12.96	1.72	12.86	1.58	12.95	1.70

Note. PCL-5 = Posttraumatic Stress Disorder (PTSD) Checklist-5.
* For the ≥29 and ≥33 cutoff points, the Mean Score procedure ordered items from lowest to highest mean score. For all other cutoff points, the Mean Score procedure ordered items from highest to lowest mean score.

dataset. At all six cutoff points, the standard PCL-5 ordering's mean number of items administered was higher than that of the best random ordering, for both the training dataset and the test dataset. When averaging the training-dataset results across all six cutoff points, the Mean Score procedure lowered the mean number of items administered by an average of 0.27 items compared to the standard PCL-5 ordering, and 0.04 items compared to the best random ordering (data not shown). When averaging the test-dataset results across all six cutoff points, the Mean Score procedure lowered the mean number of items administered by an average of 0.30 items compared to the standard PCL-5 ordering; the best random ordering's mean number of items administered was on average lower than that of the Mean Score procedure by 0.01 items. Comparing the results of Table 1 and Table 2, all complete-dataset means of Table 1 were within 0.31 items of the corresponding training-dataset and test-dataset means of Table 2.

One finding evident from Table 2 is that contrary to the trend anticipated from the "capitalization on chance" problem, the mean number of items administered was lower in the test dataset than in the training dataset in some conditions. This phenomenon occurred at the four lowest cutoff points when comparing the Mean Score procedure's performance in the test dataset with the same procedure's performance in the training dataset. An analogous result occurred for the standard PCL-5 ordering at the five lowest cutoff points. This finding may have been due to random chance: the best random ordering (whose performance is unaffected by the capitalization on chance problem) also exhibited a lower mean number of items administered in the test dataset than in the training dataset at the four lowest cutoff points. The latter result suggests that at low cutoff points, respondents randomly selected for the test dataset may have been able to be classified more quickly than respondents randomly selected for the training dataset.

Figure 1 supplements Tables 1 and 2 by showing histograms of the mean number of items administered by the 10,000 random orderings, separated by dataset and cutoff point. The figure demonstrates that the number of items administered tended to be greater for the cutoff points ≥7, ≥10, ≥14, and ≥19 than for the cutoff points ≥29 and ≥33. The spread of the plotted values also varied by cutoff point; the ≥7 cutoff point exhibited the largest range (minimum = 16.16 items, maximum = 17.24 items for the random orderings evaluated in the complete dataset), and the ≥19 cutoff point exhibited the smallest range (minimum = 16.10 items, maximum = 16.32 items for the random orderings evaluated in the complete dataset). The mean number of items administered by the standard PCL-5 ordering and the Mean Score procedure are shown on each plot for purposes of comparison. Figure 1 was made using MATLAB 2015a (MathWorks, 2015).

Simulation Study 2: The Case of Multiple Clusters

As in Simulation Study 1, the aim of Simulation Study 2 was to compare the performance of different item orderings alongside the countdown method. However, in Simulation Study 2, a cluster-based scoring rule with multiple clusters was investigated.

Method

The PCL-5 consists of four clusters: Intrusion (e.g., "feeling very upset when something reminded you of the stressful experience"), Avoidance (e.g., "avoiding external reminders of the stressful experience"), Negative Alterations in Cognitions and Mood (e.g., "loss of interest in activities that you used to enjoy"), and Alterations in Arousal and Reactivity (e.g., "feeling jumpy or easily startled"; Weathers et al., 2013). These clusters are referred to as Cluster B, Cluster C, Cluster D, and Cluster E, respectively. As mentioned previously, each item in each cluster is scored from 0 to 4; under one scoring rule described by Weathers et al. (2013) and used in the current section, an item is considered to be "endorsed" if its score is 2 or above. This method has been used to determine provisional PTSD cases in prior epidemiologic studies using the PCL for *DSM–IV* PTSD (e.g., Goldmann et al., 2011; Tracy et al., 2011). A positive result is considered to be obtained for Cluster B if at least one of the five items from that cluster is endorsed. The rules for defining a positive result for the other clusters are as follows: for Cluster C, at least one of the cluster's two items must be endorsed; for Cluster D, at least two of the cluster's seven items must be endorsed; for Cluster E, at least two of the cluster's six items must be endorsed. A provisional PTSD diagnosis is then made if and only if all four clusters are positive; hence, this scoring rule is an example of cluster-based scoring with multiple clusters.

As discussed in the Introduction, the countdown method performs two functions in the case of multiple clusters. First, it stops

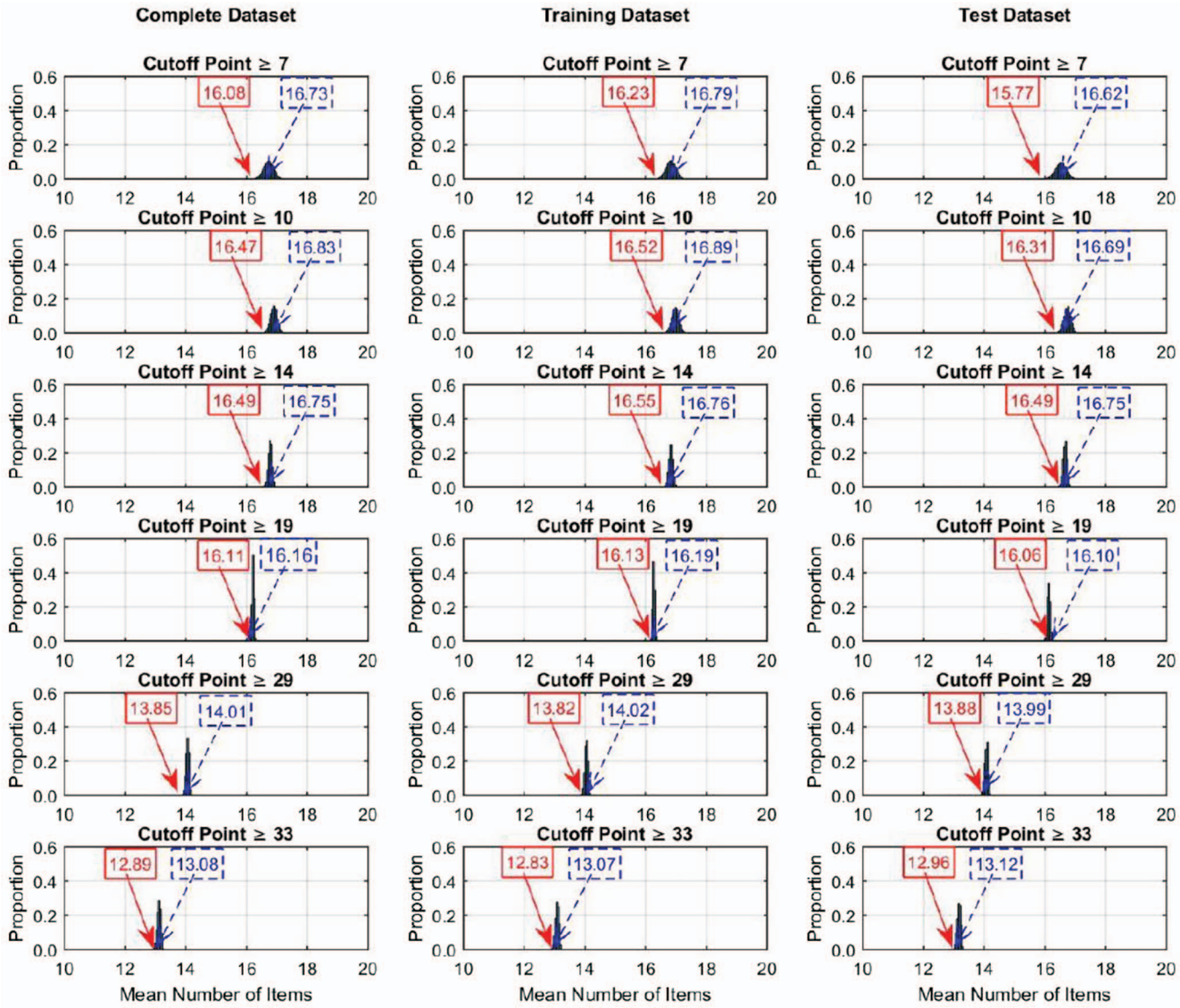

Figure 1. Mean number of items administered for 10,000 random orderings, the standard Posttraumatic Stress Disorder Checklist-5 (PCL-5) item ordering, and the Mean Score procedure: Simulation Study 1 (n = 942). The mean number of items administered by the standard PCL-5 item ordering is indicated on each plot by a dashed arrow. The mean number of items administered by the Mean Score procedure is indicated on each plot by a solid arrow. See the online article for the color version of this figure.

presenting items from the current cluster—and skips the respondent ahead to the next cluster—if the positive or negative result of the current cluster has become deterministic based on the respondent's previous answers. For instance, if a respondent endorses the first item of Cluster B (e.g., with a score of 3), then it has become deterministic that a positive result will be obtained for that cluster (since only one endorsed item from Cluster B is required in order for a respondent to be positive on that cluster). Therefore, for the purpose of test efficiency, the countdown method prescribes that the respondent be skipped past the remaining items in Cluster B (Items 2–5 of that cluster) and moved forward to Cluster C. In similar fashion, the respondent may be skipped from Cluster C to Cluster D, or from Cluster D to Cluster E, as appropriate. The second function of the countdown method is to stop the assessment as a whole if its result has become deterministic. For instance, if a respondent receives a positive result from Cluster B, but a negative result from Cluster C, then the result of the assessment has been determined to be negative (since a positive result for the assessment cannot be obtained if any cluster's result is negative). Therefore, the countdown method prescribes that the test be terminated for a negative result following Cluster C. Similarly, early stopping can occur after Cluster B or Cluster D (and naturally, stopping always occurs after Cluster E, the final cluster).

As in the case of one cluster, one may hypothesize that when multiple clusters are present, the ordering of items would affect the countdown method's ability to reduce the average length of a test.

One practical consideration is that when selecting an item ordering for a test with multiple clusters, it may be desired to limit the search to orderings in which each cluster is presented as an intact unit (and in fact, the standard ordering of items for the PCL-5 does administer the clusters as intact units). If indeed each cluster is to be presented as a continuous block of items, there are two decisions to make about the item ordering: how to order the items within each cluster, and how to order the clusters themselves, in order to enhance the countdown method's ability to shorten the assessment. Each of these decisions will be discussed in turn.

To order the items within a given cluster, the Mean Score procedure may be used, with the individual cluster (as opposed to the test as a whole) serving as the unit to be shortened. That is, the Mean Score procedure operates as follows to select the item ordering for a specific cluster: using training data, the average number of items administered in the cluster (before the countdown method terminates testing in that cluster) is calculated for the case where items are ordered from lowest to highest mean score. This calculation is repeated for the case where items are ordered from highest to lowest mean score. The Mean Score procedure then selects the ordering that results in the lower average number of items administered, that is, the greater average item reduction. Note that if all items in the cluster are scored dichotomously, then the two orderings considered by the Mean Score procedure revert to the least to most frequently endorsed ordering and most to least frequently endorsed ordering of Ben-Porath et al. (1989). This is due to the fact that the mean of a dichotomous variable is equal to the proportion of endorsement of that variable, when *endorsement* is coded as 1 and *lack of endorsement* is coded as 0.

To order the clusters themselves, several different approaches may be considered. One possible method would be to begin with the cluster that is most likely to result in termination of the entire assessment. In particular, because cluster-based scoring prescribes a positive result for the entire assessment if and only if a respondent receives a positive result for every cluster, one may start with the cluster that is most likely to produce a negative result (as such a result would terminate the assessment as a whole, per the countdown method). Likewise, the remaining clusters can then be placed in order of their probability to produce a negative result (from highest probability to lowest probability). One potential drawback of this approach is that the clusters with the highest probabilities of a negative result are not necessarily the shortest clusters (i.e., the clusters with the fewest items). For example, consider two hypothetical clusters, which will be referred to as Cluster 1 and Cluster 2. Suppose that the probability of a negative result is 41% for Cluster 1 and 40% for Cluster 2, so that the above ordering approach would administer Cluster 1 prior to Cluster 2. However, if Cluster 1 contained 10 items, whereas Cluster 2 contained only three items, it might be more efficient to administer Cluster 2 first, in order to have a 40% chance of terminating the test after only three items (or fewer, if early stopping occurred during the presentation of Cluster 2). Following this logic, one may suggest ordering the clusters from shortest to longest; however, such a procedure might result in presenting clusters that are short, but are nevertheless very unlikely to terminate the assessment, at the beginning of the test. If the number of clusters is relatively small, it may be feasible to evaluate the average item reduction provided by the countdown method (using the Mean Score procedure to order the items within a cluster) for every possible cluster ordering, then select the cluster ordering that produces the greatest average item reduction. As the PCL-5 only contains four clusters, and therefore only $4! = 24$ possible cluster orderings, each cluster ordering may easily be evaluated without undue computational burden.

Multiple analyses were undertaken in the real-data simulation, which utilized the same dataset as Simulation Study 1 ($n = 942$). First, the effect of item ordering within a given cluster was examined for each of the four PCL-5 clusters separately. Specifically, for each cluster, the mean number of items administered was calculated for all possible orderings of items in that cluster. In other words, the mean was calculated for every possible item ordering in a given cluster, assuming that all participants entered the cluster and that the countdown method was used to determine how many items were administered in the cluster. We note that the use of 10,000 random item orderings (as was done for Simulation Study 1) was not necessary in the analysis being described currently. This is because each cluster was short enough that the total number of within-cluster item orderings was adequately modest ($5! = 120$ possible orderings in Cluster B, $2! = 2$ in Cluster C, $7! = 5{,}040$ in Cluster D, and $6! = 720$ in Cluster E) for every ordering to be analyzed in each cluster. For each cluster, the ordering producing the greatest average item reduction (hereafter referred to as the ordering obtained by *comprehensive search*) was compared to the PCL-5's standard ordering for that cluster, as well as to the ordering produced by the Mean Score procedure for that cluster. This investigation was conducted in order to shed light on the Mean Score procedure's ability to select efficient within-cluster orderings, relative to other orderings including the best possible ordering (that of the comprehensive search).

The above analysis focused on each cluster as a separate unit. Additional analyses were then conducted to compare the performance of different orderings in their ability to shorten the entire test alongside the countdown method. First, each cluster's items were placed in order according to a given within-cluster ordering procedure—either the standard PCL-5 ordering, the Mean Score procedure, or comprehensive search. Then, each of the above within-cluster ordering procedures was coupled with every possible ordering of the clusters themselves to produce an overall ordering of items. For example, in the analysis of the Mean Score procedure, items within each cluster were placed in order according to this procedure; then, the prescribed within-cluster orderings were paired with all possible orderings of the clusters themselves ("B, C, D, E", "B, C, E, D," "B, D, C, E," etc.) for evaluation. In particular, the mean and standard deviation of the number of items administered was computed for every such combination of *within-cluster ordering* and *between-cluster ordering* under study. This design allowed for the comparison of within-cluster orderings as well as between-cluster orderings. The percentage of times that each cluster produced a negative result (which would thereby terminate the entire assessment, per the countdown method) was also calculated. For parsimony, only one cutoff point was examined for each cluster: the cutoff point suggested by Weathers et al. (2013). As mentioned previously, the cutoff points were ≥ 1 for Cluster B, ≥ 1 for Cluster C, ≥ 2 for Cluster D, and ≥ 2 for Cluster E.

As in Simulation Study 1, both a complete-dataset analysis ($n = 942$) and a training-test analysis were undertaken. The same training dataset ($n = 628$) and test dataset ($n = 314$) as used in

Simulation Study 1 were used in Simulation Study 2. Results were obtained via a computer program written in R (Version 3.1.2; R Core Team, 2015).

Results

As the group of participants examined in Simulation Study 2 was the same as that of Simulation Study 1, all demographic information was identical. Among the 942 participants, 37 (3.9%) had a positive result on the PCL-5 when the aforementioned cluster-based scoring rule with four clusters was used.

Figure 2 presents histograms of the mean number of items administered within a given cluster, separated by dataset and cluster, for all possible within-cluster item orderings. For example, the plots labeled *Cluster D* (second row of Figure 2) show the distribution of the mean number of items administered in Cluster D only when the ordering of items within Cluster D is varied, assuming that each participant enters that cluster. Results are not shown graphically for Cluster C because that cluster, which has two items, has only two possible within-cluster item orderings; see the footnote to Figure 2 for Cluster C's results. The mean number of items administered by the standard PCL-5 ordering (dashed arrow) and the mean number of items administered by the Mean Score procedure (solid arrow) are also displayed on each plot. In all cases, the Mean Score procedure selected the most frequently endorsed to least frequently endorsed ordering, as opposed to the converse. Furthermore, in all cases, the Mean Score procedure achieved a greater reduction in the mean number of items administered than the standard PCL-5 item ordering. In the complete dataset, the Mean Score procedure produced the best possible within-cluster item ordering (i.e., the ordering producing the smallest mean number of items administered in that cluster, compared to all other within-cluster orderings), or an ordering that was tied for the best performance, in three of the four clusters. The exception was Cluster D, in which the Mean Score procedure's mean number of items administered was 5.82; this was tied for the 195th best ordering out of all possible 5,040 orderings in that cluster (the best ordering averaged 5.80 items, and the worst averaged 5.91). In the training dataset, the Mean Score procedure tied for the best order-

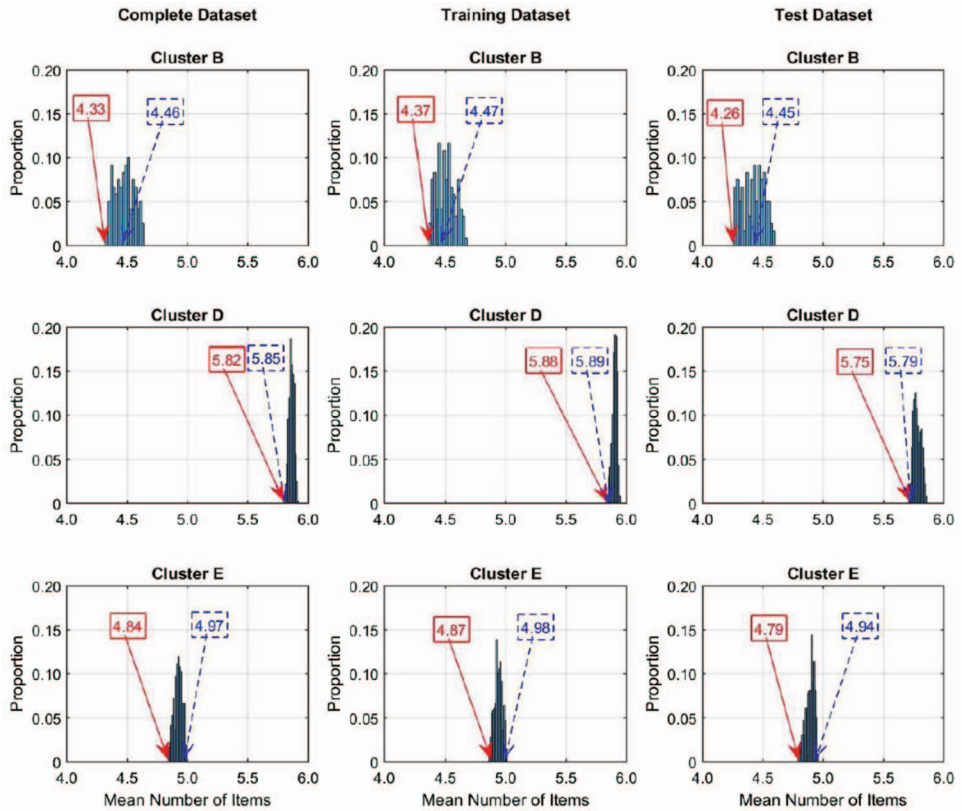

Figure 2. Mean number of items administered for all possible within-cluster item orderings,* the standard Posttraumatic Stress Disorder Checklist-5 (PCL-5) ordering, and the Mean Score procedure: Simulation Study 2 ($n = 942$). The mean number of items administered by the standard PCL-5 item ordering is indicated on each plot by a dashed arrow. The mean number of items administered by the Mean Score procedure is indicated on each plot by a solid arrow. * Results are not shown for Cluster C because there are only two possible item orderings for Cluster C. Means of the number of items administered for Cluster C were as follows: Complete dataset: 1.93 for the standard PCL-5 ordering, 1.92 for Mean Score procedure; Training dataset: 1.93 for the standard PCL-5 ordering, 1.92 for Mean Score procedure; Test dataset: 1.92 for the standard PCL-5 ordering, 1.91 for Mean Score procedure. See the online article for the color version of this figure.

ing (out of 120) in Cluster B, produced the best ordering (out of only two) in Cluster C, tied for the 589th best ordering in Cluster D, and tied for the fifth best ordering (out of 720) in Cluster E. In the test dataset, the Mean Score procedure produced the best ordering in both Clusters B and C, tied for the 795th best ordering in Cluster D, and tied for the best ordering in Cluster E. Consistent with the fact that each cluster contains a small number of items (two to seven), the difference between the least efficient item ordering and the most efficient item ordering in each cluster was relatively slight. In the complete dataset, the difference in the mean number of items administered between the least and most efficient orderings ranged from 0.01 items (Cluster C) to 0.30 items (Cluster B), with similar results in the training and test data sets.

Table 3 shows results for different combinations of between-cluster and within-cluster orderings, using the complete PCL-5 dataset. The means and standard deviations presented in Table 3 refer to the total number of items administered in all clusters when using the countdown method, as opposed to results for each cluster separately. All orderings in which the first cluster presented was Cluster C were found to exhibit mean test lengths between 2.58 and 2.78 items. The smallest mean test length displayed in the table, 2.58 items, was observed when the cluster ordering was C, E, B, D and the within-cluster orderings were determined by comprehensively searching through every possible item ordering within each cluster. Use of the Mean Score procedure alongside the C, E, B, D between-cluster ordering resulted in a mean test length of 2.59 items; the combination of using the standard PCL-5 within-cluster ordering and the C, E, B, D between-cluster ordering resulted in a mean of 2.70 items. When Cluster B was presented first, all mean test lengths were between 5.11 and 5.86 items. The corresponding ranges were 5.26 to 5.75 items for orderings in which Cluster E was presented first and 6.23 to 6.57 items for orderings in which Cluster D was presented first. The combination of the standard PCL-5 between-cluster ordering (B, C, D, E) and the standard PCL-5 within-cluster ordering resulted in a mean test length of 5.35 items. For any given between-cluster ordering, the difference in means among the three within-cluster orderings under study (the standard PCL-5 ordering, Mean Score procedure, and comprehensive search) was never more than 0.27 items. This value was obtained for between-cluster ordering B, E, D, C ($M = 5.78$ items for the standard PCL-5 ordering, $M = 5.51$ items for comprehensive search). When averaging the results of Table 3 across all 24 between-cluster orderings, the Mean Score procedure reduced the mean number of items administered by an average of 0.16 items when compared to the standard PCL-5 ordering (data not shown). Comprehensive search's mean number of items administered was on average lower than that of the Mean Score procedure by 0.01 items.

Table 4 presents results of the training-test analysis comparing different combinations of between-cluster and within-cluster orderings. The mean test length ranged from 2.53 to 2.85 items when Cluster C was presented first. The corresponding ranges were 5.08 to 5.96 items when Cluster B was presented first, 5.21 to 5.99 items when Cluster E was presented first, and 6.23 to 6.65 items when Cluster D was presented first. Therefore, the presentation of Cluster C first had the greatest impact on mean test length, with other differences in ordering having a small effect on efficiency. The combination of the standard PCL-5 between-cluster ordering

Table 3
Performance of Different Combinations of Between-Cluster Orderings and Within-Cluster Orderings in the Complete PCL-5 Dataset ($n = 942$): Simulation Study 2

	Standard PCL-5 ordering within each cluster		Mean Score procedure within each cluster		Comprehensive search within each cluster	
Cluster ordering	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items
B, C, D, E	5.35	1.63	5.17	1.35	5.17	1.34
B, C, E, D	5.30	1.53	5.11	1.13	5.11	1.11
B, D, C, E	5.77	1.85	5.58	1.45	5.58	1.43
B, D, E, C	5.86	2.06	5.65	1.57	5.64	1.55
B, E, C, D	5.65	1.68	5.40	1.07	5.40	1.05
B, E, D, C	5.78	2.06	5.52	1.36	5.51	1.33
C, B, D, E	2.78	2.39	2.67	2.05	2.67	2.04
C, B, E, D	2.73	2.27	2.61	1.84	2.60	1.82
C, D, B, E	2.78	2.37	2.69	2.09	2.68	2.06
C, D, E, B	2.77	2.35	2.69	2.08	2.68	2.06
C, E, B, D	2.70	2.19	2.59	1.78	2.58	1.77
C, E, D, B	2.71	2.21	2.60	1.83	2.60	1.81
D, B, C, E	6.46	1.38	6.36	1.12	6.33	1.06
D, B, E, C	6.56	1.62	6.42	1.24	6.40	1.18
D, C, B, E	6.36	1.26	6.25	1.05	6.23	1.00
D, C, E, B	6.36	1.25	6.25	1.04	6.23	.99
D, E, B, C	6.57	1.68	6.44	1.27	6.41	1.20
D, E, C, B	6.52	1.51	6.38	1.14	6.36	1.08
E, B, C, D	5.59	1.63	5.40	1.05	5.40	1.03
E, B, D, C	5.72	2.03	5.52	1.35	5.51	1.32
E, C, B, D	5.44	1.40	5.27	.95	5.26	.92
E, C, D, B	5.44	1.42	5.28	1.00	5.27	.98
E, D, B, C	5.75	2.07	5.57	1.48	5.56	1.43
E, D, C, B	5.71	1.92	5.52	1.34	5.51	1.29

Note. PCL-5 = Posttraumatic Stress Disorder (PTSD) Checklist-5.

Table 4
Performance of Different Combinations of Between-Cluster Orderings and Within-Cluster Orderings in the Training (n = 628) and Test (n = 314) PCL-5 Datasets: Simulation Study 2

	Standard PCL-5 ordering within each cluster				Mean Score procedure within each cluster				Comprehensive search within each cluster			
	Training		Test		Training		Test		Training		Test	
Cluster ordering	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items	Mean # items	*SD* # items
B, C, D, E	5.31	1.62	5.41	1.67	5.17	1.38	5.18	1.28	5.16	1.33	5.19	1.36
B, C, E, D	5.25	1.42	5.40	1.73	5.09	1.08	5.15	1.25	5.08	1.05	5.16	1.26
B, D, C, E	5.74	1.81	5.82	1.92	5.59	1.49	5.59	1.36	5.57	1.44	5.59	1.41
B, D, E, C	5.81	1.99	5.96	2.21	5.65	1.61	5.66	1.53	5.63	1.55	5.67	1.55
B, E, C, D	5.58	1.54	5.77	1.92	5.39	1.03	5.43	1.14	5.38	1.01	5.44	1.17
B, E, D, C	5.70	1.91	5.92	2.34	5.50	1.32	5.59	1.50	5.48	1.28	5.59	1.47
C, B, D, E	2.74	2.35	2.85	2.47	2.64	2.06	2.72	2.04	2.63	2.01	2.74	2.11
C, B, E, D	2.68	2.15	2.84	2.50	2.57	1.76	2.70	2.00	2.56	1.73	2.70	2.00
C, D, B, E	2.75	2.34	2.84	2.42	2.66	2.07	2.75	2.12	2.64	2.00	2.77	2.17
C, D, E, B	2.74	2.32	2.84	2.41	2.65	2.05	2.76	2.15	2.63	1.97	2.77	2.20
C, E, B, D	2.64	2.06	2.82	2.43	2.54	1.69	2.69	1.98	2.53	1.65	2.70	1.99
C, E, D, B	2.66	2.13	2.81	2.37	2.55	1.73	2.71	2.03	2.54	1.69	2.71	2.06
D, B, C, E	6.46	1.38	6.47	1.38	6.38	1.18	6.35	1.00	6.33	1.09	6.34	1.02
D, B, E, C	6.53	1.57	6.61	1.72	6.43	1.29	6.42	1.18	6.39	1.20	6.41	1.16
D, C, B, E	6.36	1.25	6.37	1.29	6.28	1.06	6.25	.97	6.24	.97	6.24	1.04
D, C, E, B	6.35	1.24	6.37	1.28	6.27	1.03	6.25	1.01	6.23	.93	6.24	1.08
D, E, B, C	6.53	1.59	6.65	1.85	6.43	1.29	6.47	1.30	6.39	1.18	6.46	1.28
D, E, C, B	6.50	1.49	6.56	1.57	6.39	1.17	6.39	1.09	6.35	1.08	6.39	1.10
E, B, C, D	5.48	1.45	5.79	1.92	5.33	.92	5.55	1.28	5.32	.87	5.56	1.31
E, B, D, C	5.60	1.85	5.94	2.34	5.44	1.23	5.71	1.60	5.42	1.18	5.71	1.58
E, C, B, D	5.38	1.27	5.56	1.63	5.23	.86	5.35	1.11	5.21	.81	5.37	1.14
E, C, D, B	5.39	1.34	5.55	1.56	5.24	.91	5.37	1.17	5.22	.86	5.39	1.21
E, D, B, C	5.64	1.91	5.99	2.35	5.48	1.34	5.79	1.77	5.45	1.26	5.78	1.75
E, D, C, B	5.61	1.81	5.90	2.10	5.44	1.20	5.71	1.59	5.41	1.13	5.70	1.58

Note. PCL-5 = Posttraumatic Stress Disorder (PTSD) Checklist-5.

(B, C, D, E) and the standard PCL-5 within-cluster ordering had a mean test length of 5.31 items in the training dataset and 5.41 items in the test dataset. In the training dataset, the smallest mean test length (2.53 items) was produced by the combination of the C, E, B, D between-cluster ordering and the within-cluster ordering found by comprehensive search; the Mean Score procedure's mean test length with the same between-cluster ordering was 2.54 items. When applying to the test dataset the within-cluster orderings that had been found by the Mean Score procedure and comprehensive search in the training dataset, the Mean Score procedure's mean test length alongside the C, E, B, D between-cluster ordering (2.69 items) was slightly lower than that of comprehensive search (2.70 items). The value of 2.69 represented the lowest mean test length presented in Table 4 for the test dataset. For any given between-cluster ordering, the difference in means among the three within-cluster orderings under study was never more than 0.34 items. When averaging the training-dataset results across all 24 between-cluster orderings, the Mean Score procedure lowered the mean number of items administered by an average of 0.13 items compared to the standard PCL-5 ordering (data not shown). Comprehensive search's mean number of items administered was on average lower than that of the Mean Score procedure by 0.02 items. When averaging the test-dataset results across all 24 between-cluster orderings, the Mean Score procedure lowered the mean number of items administered by an average of 0.19 items compared to the standard PCL-5 ordering, and 0.003 items compared to comprehensive search. Comparing the results of Table 3 and Table 4, all complete-dataset means of Table 3 were within 0.24 items of the corresponding training-dataset and test-dataset means of Table 4.

Regarding the percentage of times that each cluster produced a negative result, the highest such percentage in the complete dataset was for Cluster C (840 of 942 participants, or 89.2%, were negative in this cluster). Eight hundred thirty-nine participants (89.1%) were negative in Cluster D, 835 (88.6%) were negative in Cluster E, and 765 (81.2%) were negative in Cluster B. All percentages in the training and test data sets were within 4% of their values in the complete dataset.

Discussion

In their landmark paper, Ben-Porath et al. (1989) compared the least to most frequently endorsed and most to least frequently endorsed item ordering procedures in the context of using the countdown method for the MMPI-2. Further research (e.g., Forbey et al., 2009, 2012; Handel et al., 1999; Roper et al., 1991, 1995; Rudick et al., 2013; Yang et al., 2009) also acknowledged the effect of item ordering when employing the countdown method. The current research extended the approach of Ben-Porath et al. (1989) to the case of polytomous items, via the Mean Score procedure described herein. Additionally, comprehensive simulation studies were performed to evaluate the Mean Score procedure

to other item orderings in both the case of one cluster and the case of multiple clusters. Because the Mean Score procedure reduces to selecting the better of the least to most frequently endorsed and most to least frequently endorsed orderings in the case of dichotomously scored items, the results of Simulation Study 2 (which used dichotomous scoring) have implications for Ben-Porath et al.'s approach to choosing an item ordering.

Although the methodologies examined in this research can be applied generally to assessments that employ cluster-based scoring, the study also has implications for the PCL-5 specifically. Simulation Study 1 suggested that when the one-cluster scoring rule is used for this assessment, the effect of item ordering on the gains provided by the countdown method was modest, that is, there was a small difference between the item orderings in terms of the mean number of items administered. Indeed, despite the finding that the Mean Score procedure's mean number of items administered was lower than the mean number of items administered by the standard PCL-5 ordering, the difference in means was small enough to be of limited practical significance. Simulation Study 2 suggested that when the four-cluster scoring rule is used for the PCL-5, the between-cluster ordering had a greater effect on the mean test length than the within-cluster ordering. In particular, the fact that a mean test length of fewer than three items was found whenever Cluster C was presented first (as opposed to a mean test length of over five whenever another cluster was presented first, including a mean of 5.35 items for the conventional booklet PCL-5 ordering in the complete dataset) may be useful to practitioners who are considering using the countdown method for this screener. This result is understandable given that Cluster C was both the shortest cluster (two items) and that which that produced the most negative results (just ahead of Clusters D and E). The relative lack of effect of the within-cluster ordering on mean test length is highlighted by Figure 2, which illustrates that the maximum difference in mean test length (i.e., the mean test length of the least efficient ordering minus the mean test length of the most efficient ordering) never exceeded 0.30 items for any cluster. Additionally, it shows that the mean difference between the standard PCL-5 ordering and the ordering resulting from the Mean Score procedure never exceeded 0.19 items for any cluster. These results suggest that from the standpoint of practical significance, the limited benefit of using the Mean Score procedure as a within-cluster ordering method for the PCL-5 may not warrant the additional work to determine and implement its revised ordering of items within each cluster. Indeed, practitioners using the PCL-5 to screen respondents for PTSD may infer from the results that maintaining the standard PCL-5 ordering within each cluster yields adequate efficiency in comparison with the Mean Score procedure. In sum, the results of Simulation Study 2 suggest that the between-cluster ordering (specifically, the presentation of Cluster C first) is of greater consequence than the within-cluster ordering in terms of efficiency. Practitioners seeking to reduce the mean test length of the PCL-5 may therefore give a greater level of focus to the between-cluster ordering than the within-cluster ordering when employing the countdown method alongside cluster-based scoring with four clusters. This distinction between the impact of the between-cluster ordering and that of the within-cluster ordering was not known prior to this study.

We emphasize that any finding obtained herein does not necessarily generalize to other populations. For instance, results may depend on the percentage of respondents in the given population who are positive on the screener. In the current study, the percentage of positive results was 3.7% in Simulation Study 1 (using a cutoff point of ≥33); the percentage of positive results in Simulation Study 2 was 3.9%.

The observed effect of item ordering on the efficiency of the countdown method was smaller than that of previous research using the MMPI-2 (Ben-Porath et al., 1989). One reason for this discrepancy was likely the difference in test length between the instruments: as the MMPI-2 is substantially longer than the PCL-5, the former presents a greater opportunity for the effect of item ordering to manifest itself than the latter. Additionally, in the current study, over 40% of respondents had a total score of zero; since these respondents had the same score (0) on every item, their test lengths did not change based on the within-cluster item ordering. Therefore, their presence dampened the effect of the within-cluster item ordering on test efficiency.

Given that the effect of item ordering on test efficiency varies from study to study, the relative performance of the ordering methods is of importance in addition to the magnitude of the differences. Indeed, although the current research suggested a small magnitude of difference between orderings for one specific combination of test and respondent population, a finding that the Mean Score procedure performed well relative to other orderings could portend the method's utility in scenarios where the item ordering has a greater effect on efficiency. Results of both simulation studies suggested that although the Mean Score procedure does not necessarily select the optimal ordering (i.e., the ordering that produces the smallest mean number of items administered alongside the countdown method), it tends to select an item ordering with a low mean relative to the distribution of possible orderings. Therefore, it has potential to assist in reducing the respondent and administrative burden associated with testing when the countdown method is used. A more complex approach to finding the optimal ordering, such as the use of a genetic algorithm (Holland, 1973) or another search algorithm, could be used but would be far more computationally intensive than the Mean Score procedure. Indeed, in the case of one cluster, the Mean Score procedure requires the evaluation of only two item orderings (lowest to highest mean score and highest to lowest mean score) in the selection of its ordering. To define the within-cluster orderings for an assessment with multiple clusters, the Mean Score procedure requires the evaluation of only two item orderings per cluster. The decision of whether to use a heuristic method like the Mean Score procedure, or a more complicated approach like a genetic algorithm, can be made on a case-by-case basis depending on the relative importance of finding the optimal ordering versus maintaining computational simplicity. Differences in performance between these ordering methods might depend on the assessment and cutoff point used.

Another matter to consider when reordering questionnaire items is that a respondent's answer to a given item might not be invariant to which items preceded it in the assessment. That is, the possible presence and consequences of context effects should be investigated when altering the item ordering (Ortner, 2008). Because the current study involved retrospectively analyzing data from participants who had previously answered test items in their standard order, the results assume that respondents' answers would have remained the same if the items had been reordered. Further pro-

spective studies could be performed to evaluate the methods described herein in the case where items are reordered prior to presentation to respondents. See Roper et al. (1991, 1995) for pertinent previous work on the comparability of booklet and computerized adaptive forms in the context of the MMPI-2.

In addition to the limitations mentioned previously (the use of a single dataset whose results might not generalize to other populations, and the study's retrospective nature), another limitation was the absence of a gold standard diagnosis of PTSD in the dataset. Due to this lack of a gold standard, the sensitivity and specificity of the PCL-5 could not be evaluated in this study. However, the objectives of the study—extending the item orderings of Ben-Porath et al. (1989) to polytomous items, and comparing the efficiencies of different item orderings alongside the countdown method and cluster-based scoring—did not require a gold standard diagnosis. Furthermore, because the countdown method always produces the same classification (positive or negative) as the full-length screener for every respondent, the countdown method's sensitivity and specificity always match those of the full-length screener, irrespective of the gold standard being used for comparison. Therefore, it is unnecessary to compare the countdown method's sensitivity and specificity to those of a full-length screener such as the 20-item PCL-5. We also note that in some assessment contexts, a continuous score may be desired, as opposed to a dichotomous classification. Indeed, there may be benefits of employing dimensional assessments in lieu of a categorical model. See Helzer et al. (2008) for a comprehensive discussion, including advantages and disadvantages of different approaches.

The simulations performed in the current research corresponded to the "classification" version of the countdown method, in which the assessment is stopped early as soon as the respondent's classification has become deterministic. Ben-Porath et al. (1989) also described a second version of the countdown method, the full scores on elevated scales (FSES) procedure, which only stops early if a negative result has become deterministic (it continues testing after a deterministic positive result). It is useful for scenarios in which full scores are desired for respondents who score positively. Future study could compare the Mean Score procedure to the distribution of possible item orderings, including for assessments with polytomously scored items, in the case where the FSES procedure is used. Because the FSES procedure does not provide item savings for respondents who reach the cutoff point quickly, it would be anticipated that the lowest to highest mean score ordering would tend to be selected, rather than the highest to lowest mean score ordering, in the Mean Score procedure. This anticipation is consistent with Ben-Porath et al.'s finding that for the dichotomously scored items of the MMPI-2, the least to most frequently endorsed ordering was more efficient than the most to least frequently endorsed ordering alongside the FSES.

In the current paper, both the item clusters and their cutoff points were based on the diagnostic rules for PTSD under *DSM–5*; that is, they were based upon clinical theory. It is noteworthy that item clusters and their cutoff points may also be deduced empirically. Recent research (Fokkema, Smits, Kelderman, Carlier, & van Hemert, 2014) applied an alternative version of the countdown method that used a decision tree to select clusters and their associated cutoff points for optimal diagnostic accuracy. Evidently, for such an approach a large dataset from the target population is needed with not only scores on the items of the screening instrument, but also a diagnosis based on a gold standard.

Possible directions of future research involve conducting variations on the work presented herein. The current study investigated the countdown method, which only stops early if the result of the assessment is unambiguous. It is also possible, however, to stop early if one of the potential test results (either positive or negative) becomes disproportionately probable; such a rule, mentioned briefly in the Introduction as stochastic curtailment, is more aggressive than the countdown method (Butcher et al., 1985; Finkelman et al., 2016). The current study could be repeated with this more aggressive stopping procedure. An additional possibility would be to combine the countdown method with adaptive item selection (i.e., to allow items to be presented in different orders for different respondents, depending on the respondents' previous answers). The study of such an approach is beyond the scope of the objectives of the current research, which focused on static item ordering procedures because they have dominated the previous work on the countdown method and have been used in practice alongside it (Ben-Porath et al., 1989; Forbey et al., 2012, 2009; Handel, Ben-Porath, & Watt, 1999; Roper et al., 1991, 1995). Indeed, as stated by Rudick et al. (2013, p. 770), the countdown method "administers items in a static order until scale elevation above a predetermined clinical threshold is either certain or impossible."

In sum, the study carries practical implications for the PCL-5, as well as methodological implications for other screeners that can be paired with the countdown method. Given the importance of minimizing respondent burden among individuals being screened for PTSD, as well as the importance of lessening administrative burden, an empirically supported approach to reducing the test length of the PCL-5 is significant. The current study is the first to propose that the screener's item clusters be reordered to improve efficiency. The results of the data analysis suggest that such between-cluster reordering is effective in reducing the PCL-5's mean test length alongside the countdown method and cluster-based scoring with four clusters. In particular, the use of a simple modification of the screener whereby Cluster C's items are presented first may save time and alleviate burden. The finding that reordering the clusters themselves is more influential in enhancing the PCL-5's efficiency than reordering the items within each cluster is a novel result. Indeed, the improvement derived from administering Cluster C first, while intuitive, had no empirical basis prior to the current study, nor had the idea been presented in the literature. With regard to methodological implications, the current study makes a dual contribution. First, as noted previously, the pioneering work of Ben-Porath et al. (1989) recommended examination of the least to most frequently endorsed and most to least frequently endorsed item orderings to enhance the efficiency of the countdown method. Their recommendation guided the item ordering of further studies employing the countdown method (Forbey et al., 2012, 2009; Handel, Ben-Porath, & Watt, 1999; Roper et al., 1991, 1995). Yet, to our knowledge, previous study had not comprehensively examined the Ben-Porath et al. (1989) approach in comparison to the distribution of possible item orderings. Therefore, it was unknown whether the reduction in mean test length derived from their approach was effective relative to other orderings. Because the Mean Score procedure is equivalent to taking the better of the least to most frequently endorsed ordering and the

most to least frequently endorsed ordering in the case of dichotomously scored items, Simulation Study 2 provided a comparative investigation of the Ben-Porath et al. approach. The results indicated that while the approach is not always optimal, its relative performance (i.e., its rank in comparison with other orderings) tended to place it at or near the top of the distribution. This result carries ramifications for other assessments, aside from the PCL-5, for which the countdown method can be used. For example, the MMPI-2 is an assessment for which the item ordering has been shown to be consequential (Ben-Porath et al., 1989). Practitioners may note that our results support the effectiveness of Ben-Porath et al.'s procedure, and thus an exhaustive search algorithm to find the best ordering may not be necessary. The second methodological contribution made by the current study was the extension of Ben-Porath's item ordering procedure to polytomously scored items (via the Mean Score procedure). Through this extension, the study increased the universe of assessments that can take advantage of item ordering to improve efficiency alongside the countdown method. Our comprehensive comparative examination of the Mean Score procedure using polytomous items (Simulation Study 1) found that its efficiency was consistently at or near the top of the distribution of orderings. While the magnitude of the difference tended to be modest in the particular application examined, the Mean Score procedure's performance relative to other orderings (i.e., its rank) suggests its potential utility in general. Hence, both the application to the PCL-5 and the methodological component of the study promote the efficiency of the countdown method among tests that utilize cluster-based scoring.

As a final note, it is worth emphasizing that, as mentioned in the Methods section, the PCL-5 is not intended to provide definitive information on PTSD diagnostic status. Nonetheless, the previous version of this measure for *DSM–IV* has been frequently used to determine provisional or probable PTSD diagnosis in research studies (e.g., Horesh et al., 2015; Tracy et al., 2011). Recent work evaluating the psychometric properties of potential cutoff points and cluster-based conceptualizations of the PCL (Blevins et al., 2015), as well as recommendations for using cutoff points and cluster-based methods to estimate provisional or probable PTSD diagnoses (Weathers et al., 2013), suggests that the PCL-5 is intended to be used this way as well, at least in research settings. In the aftermath of a mass traumatic event, such as the context of the current study, PTSD screeners such as the PCL-5 could be utilized for rapid assessment of survivors' mental health needs. Increasing the efficiency of such screeners, such as through the methods employed in the current study, could help to ensure that mass trauma survivors with likely PTSD diagnoses are connected with services in a timely manner. Once referrals are made, diagnoses obtained via brief screeners should be validated with more time-intensive clinical interviews.

Conclusions

Within the limitations of the study, we observe that: (a) although the Mean Score procedure is not always optimal for use alongside the countdown method, it tends to produce an efficient item ordering relative to the distribution of other possible orderings. The difference between orderings may be modest in some cases; (b) for tests with multiple clusters, both the within-cluster ordering and between-cluster ordering affect efficiency to some degree, with the between-cluster ordering having a greater effect in the results presented herein; and (c) further studies should be conducted using additional assessments and a prospective design.

References

Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. *Psychological Assessment, 1,* 18–22. http://dx.doi.org/10.1037/1040-3590.1.1.18

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist (PCL). *Behaviour Research and Therapy, 34,* 669–673. http://dx.doi.org/10.1016/0005-7967(96)00033-2

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The Posttraumatic Stress Disorder Checklist for *DSM–5* (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress, 28,* 489–498. http://dx.doi.org/10.1002/jts.22059

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2016). Psychometric properties of the PTSD checklist for Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (PCL-5) in veterans. *Psychological Assessment, 28,* 1379–1391. http://dx.doi.org/10.1037/pas0000254

Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology, 53,* 803–815. http://dx.doi.org/10.1037/0022-006X.53.6.803

Butler, S. F., Fernandez, K., Benoit, C., Budman, S. H., & Jamison, R. N. (2008). Validation of the revised Screener and Opioid Assessment for Patients with Pain (SOAPP-R). *The Journal of Pain, 9,* 360–372. http://dx.doi.org/10.1016/j.jpain.2007.11.014

Dugdale, D. C., Epstein, R., & Pantilat, S. Z. (1999). Time and the patient–physician relationship. *Journal of General Internal Medicine, 14*(Suppl. 1), S34–S40. http://dx.doi.org/10.1046/j.1525-1497.1999.00263.x

Finkelman, M. D., Kulich, R. J., Zacharoff, K. L., Smits, N., Magnuson, B. E., Dong, J., & Butler, S. F. (2015). Shortening the Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R): A proof-of-principle study for customized computer-based testing. *Pain Medicine, 16,* 2344–2356. http://dx.doi.org/10.1111/pme.12864

Finkelman, M. D., Lowe, S. R., Kim, W., Gruebner, O., Smits, N., & Galea, S. (2016). Customized computer-based administration of the PCL-5 for the efficient assessment of PTSD: A proof-of-principle study. *Psychological Trauma: Theory, Research, Practice, and Policy*. Advance online publication. http://dx.doi.org/10.1037/tra0000226

Fokkema, M., Smits, N., Kelderman, H., Carlier, I. V. E., & van Hemert, A. M. (2014). Combining decision trees and stochastic curtailment for assessment length reduction of test batteries used for classification. *Applied Psychological Measurement, 38,* 3–17. http://dx.doi.org/10.1177/0146621613494466

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment, 19,* 14–24. http://dx.doi.org/10.1037/1040-3590.19.1.14

Forbey, J. D., Ben-Porath, Y. S., & Arbisi, P. A. (2012). The MMPI-2 computerized adaptive version (MMPI-2-CA) in a Veterans Administration medical outpatient facility. *Psychological Assessment, 24,* 628–639. http://dx.doi.org/10.1037/a0026509

Forbey, J. D., Ben-Porath, Y. S., & Gartland, D. (2009). Validation of the MMPI-2 computerized adaptive version (MMPI-2-CA) in a correctional intake facility. *Psychological Services, 6,* 279–292. http://dx.doi.org/10.1037/a0016195

Goldmann, E., Aiello, A., Uddin, M., Delva, J., Koenen, K., Gant, L. M., & Galea, S. (2011). Pervasive exposure to violence and posttraumatic

stress disorder in a predominantly African American urban community: The Detroit Neighborhood Health Study. *Journal of Traumatic Stress, 24,* 747–751. http://dx.doi.org/10.1002/jts.20705

Handel, R. W., Ben-Porath, Y. S., & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment, 11,* 369–380. http://dx.doi.org/10.1037/1040-3590.11.3.369

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Helzer, J. E., Kraemer, H. C., Krueger, R. F., Wittchen, H.-U., Sirovatka, P. J., & Regier, D. A. (Eds.). (2008). *Dimensional approaches in diagnostic classification: Refining the research agenda for DSM-V*. Arlington, VA: American Psychiatric Publishing.

Holland, J. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing, 2,* 88–105. http://dx.doi.org/10.1137/0202009

Horesh, D., Lowe, S. R., Galea, S., Uddin, M., & Koenen, K. C. (2015). Gender differences in the long-term associations between posttraumatic stress disorder and depression symptoms: Findings from the Detroit Neighborhood Health Study. *Depression and Anxiety, 32,* 38–48. http://dx.doi.org/10.1002/da.22267

Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the CES-D depression symptoms index. *Journal of Aging and Health, 5,* 179–193. http://dx.doi.org/10.1177/089826439300500202

Lowe, S. R., Sampson, L., Gruebner, O., & Galea, S. (2015). Psychological resilience after Hurricane Sandy: The influence of individual- and community-level factors on mental health after a large-scale natural disaster. *PLoS ONE, 10*(5), e0125761. http://dx.doi.org/10.1371/journal.pone.0125761

MathWorks. (2015). MATLAB 2015a [Computer software]. Natick, MA: Author.

Müller, A., Trotzke, P., Mitchell, J. E., de Zwaan, M., & Brand, M. (2015). The Pathological Buying Screener: Development and psychometric properties of a new screening instrument for the assessment of pathological buying symptoms. *PLoS ONE, 10*(10), e0141094. http://dx.doi.org/10.1371/journal.pone.0141094

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16,* 249–257. http://dx.doi.org/10.1111/j.1468-2389.2008.00431.x

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. http://dx.doi.org/10.1177/014662167700100306

R Core Team. (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. *Journal of Personality Assessment, 57,* 278–290. http://dx.doi.org/10.1207/s15327752jpa5702_7

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1995). Comparability and validity of computerized adaptive testing with the MMPI-2. *Journal of Personality Assessment, 65,* 358–371. http://dx.doi.org/10.1207/s15327752jpa6502_10

Rudick, M. M., Yam, W. H., & Simms, L. J. (2013). Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment, 25,* 769–779. http://dx.doi.org/10.1037/a0032541

Thompson, N. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation, 12*(1). Retrieved from http://pareonline.net/getvn.asp?v=12&n=1

Tracy, M., Norris, F. H., & Galea, S. (2011). Differences in the determinants of posttraumatic stress disorder and depression after a mass traumatic event. *Depression and Anxiety, 28,* 666–675. http://dx.doi.org/10.1002/da.20838

van Groen, M. M. (2014). *Adaptive testing for making unidimensional and multidimensional classification decisions* (Unpublished doctoral dissertation). University of Twente, Enschede, the Netherlands.

Walter, O. B. (2010). Adaptive tests for measuring anxiety and depression. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 123–136). New York, NY: Springer.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993, October). *The PTSD checklist: Reliability, validity, and diagnostic utility*. Paper presented at the 9th Annual Meeting of the International Society for Traumatic Stress Studies, San Antonio, TX.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD Checklist for *DSM–5* (PCL-5). Available from the National Center for PTSD at www.ptsd.va.gov

Weathers, F. W., Ruscio, A. M., & Keane, T. M. (1999). Psychometric properties of nine scoring rules for the Clinician-Administered Posttraumatic Stress Disorder Scale. *Psychological Assessment, 11,* 124–133. http://dx.doi.org/10.1037/1040-3590.11.2.124

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wortmann, J. H., Jordan, A. H., Weathers, F. W., Resick, P. A., Dondanville, K. A., Hall-Clark, B., . . . Litz, B. T. (2016). Psychometric analysis of the PTSD Checklist-5 (PCL-5) among treatment-seeking military service members. *Psychological Assessment, 28,* 1392–1403. http://dx.doi.org/10.1037/pas0000260

Yang, Y., Miao, D., Tian, J., Liu, X., & Zhu, X. (2009, June). *A real data simulation study of computerized adaptive testing of Chinese Soldier Personality Questionnaire*. Presented at the 3rd International Conference on Bioinformatics and Biomedical Engineering. Beijing, China.

Appendix

Example in Which the Mean Score Procedure Does not Provide the Optimal Ordering

Table A.1
Probability Mass Functions of Four Items Comprising a Hypothetical Assessment for Which the Mean Score Procedure Does Not Produce the Optimal Item Ordering

Item	P(Score = 0)	P(Score = 1)	P(Score = 2)	P(Score = 3)	P(Score = 4)	Mean Score
1	0%	40%	0%	60%	0%	2.2
2	0%	0%	100%	0%	0%	2.0
3	0%	0%	100%	0%	0%	2.0
4	0%	100%	0%	0%	0%	1.0

Table A.1 shows the probability mass functions of four items that comprise a hypothetical assessment. The items are ordered within the table from highest to lowest mean score. Item 1 exhibits a 40% probability of a score of 1, and a 60% probability of a score of 3, for a mean score of 2.2. Both Item 2 and Item 3 have a 100% chance of being scored 2, and therefore each has a mean score of 2. Item 4 has a 100% chance of being scored 1, and therefore has a mean score of 1. Now suppose that a cutoff point of ≥ 4 is utilized for this test. Because the maximum possible score on each item is also 4, it is always possible for a respondent to reach the cutoff point up until the final item. Hence, the countdown method never stops early for a negative result; it only stops early if a respondent's cumulative score reaches 4 or above. Next, note that if the items are ordered from lowest to highest mean score, then three items will always be administered (since there is 100% probability that a score of 1 will be obtained on Item 4, a score of 2 will then be obtained on Item 3, and a score of 2 will then be obtained on Item 2, thus passing the cutoff point after three items with a cumulative score of 5). If the items are ordered from highest to lowest score, there is a 40% chance of three items being administered (a score of 1 on Item 1, a score of 2 on Item 2, and a score of 2 on Item 3, for a cumulative score of 5) and a 60% chance of two items being administered (a score of 3 on Item 1 and a score of 2 on Item 2, for a cumulative score of 5). Hence, the average number of items administered by the highest to lowest mean score ordering is $(0.40)(3) + (0.60)^{*}(2) = 2.4$, which is lower than the average number of items (3) administered by the lowest to highest mean score ordering. Thus, the Mean Score procedure selects the highest to lowest ordering and results in an average of 2.4 items administered. However, this ordering is not optimal: if Item 2 were administered first and Item 3 were administered second (or vice versa), the number of items administered would always be two (because each of these items is scored 2 in 100% of cases, the cutoff point is always reached after two items with a cumulative score of 4). An ordering with Item 2 administered first and Item 3 administered second (or vice versa) would therefore exhibit an average of two items administered, better than the average provided by the Mean Score procedure. Hence, depending on the items' probability mass functions, as well as the assessment's cutoff point, it is possible that both the highest to lowest and lowest to highest item orderings are outperformed by another ordering. While this hypothetical example of four items is unlikely to resemble a practical testing situation, it does demonstrate the potential for the Mean Score procedure to be suboptimal.

Received October 20, 2016
Revision received January 11, 2017
Accepted February 7, 2017 ■