Cross-validation of short forms of the Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R)

Finkelman, M.D.; Jamison, R.N.; Kulich, R.J.; Butler, S.F.; Jackson, W.C.; Smits, N.; Weiner, S.G.

DOI
10.1016/j.drugalcdep.2017.04.016

Publication date
2017

Document Version
Final published version

Published in
Drug and Alcohol Dependence

Drug and Alcohol Dependence 178 (2017) 94–100

Contents lists available at ScienceDirect

Drug and Alcohol Dependence

journal homepage: www.elsevier.com/locate/drugalcdep

Full length article

Cross-validation of short forms of the Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R)

Matthew D. Finkelman[a,*], Robert N. Jamison[b], Ronald J. Kulich[c,d], Stephen F. Butler[e], William C. Jackson[f], Niels Smits[g], Scott G. Weiner[h]

[a] Department of Public Health and Community Service, Tufts University School of Dental Medicine, 1 Kneeland St., Boston, MA 02111, USA
[b] Departments of Anesthesiology and Psychiatry, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St., Boston, MA 02115, USA
[c] Craniofacial Pain and Headache Center, Tufts University School of Dental Medicine, 1 Kneeland St., Boston, MA 02111, USA
[d] Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, 15 Parkman St., Boston, MA 02114, USA
[e] Inflexxion, Inc., 890 Winter St., Ste. 235, Waltham, MA 02451, USA
[f] Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, 55 Fruit St. #148, Boston, MA 02114, USA
[g] Department of Methods and Statistics, Research Institute of Child Development and Education, University of Amsterdam, Faculty of Social and Behavioural Sciences, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherland
[h] Department of Emergency Medicine, Brigham and Women's Hospital, 75 Francis Street, NH-226, Boston, MA 02115, USA

ARTICLE INFO

Keywords:
Chronic pain
Opioids
Substance abuse
Risk stratification
Short form
Computer-based testing

ABSTRACT

Background: The Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R) is a 24-item assessment designed to assist in the prediction of aberrant drug-related behavior (ADB) among patients with chronic pain. Recent work has created shorter versions of the SOAPP-R, including a static 12-item short form and two computer-based methods (curtailment and stochastic curtailment) that monitor assessments in progress. The purpose of this study was to cross-validate these shorter versions in two new populations.
Methods: This retrospective study used data from patients recruited from a hospital-based pain center (n = 84) and pain patients followed and treated at primary care centers (n = 110). Subjects had been administered the SOAPP-R and assessed for ADB. In real-data simulation, the sensitivity, specificity, and area under the curve (AUC) of each form were calculated, as was the mean test length using curtailment and stochastic curtailment.
Results: Curtailment reduced the number of items administered by 30% to 34% while maintaining sensitivity and specificity identical to those of the full-length SOAPP-R. Stochastic curtailment reduced the number of items administered by 45% to 63% while maintaining sensitivity and specificity within 0.03 of those of the full-length SOAPP-R. The AUC of the 12-item form was equal to that of the 24-item form in both populations.
Conclusions: Curtailment, stochastic curtailment, and the 12-item short form have potential to enhance the efficiency of the SOAPP-R.

1. Introduction

Prescription opioid analgesics are used frequently for patients with pain, and their use has risen at a rapid rate over the past decade, with prescribing levels beginning to stabilize and decrease only recently (Aitken et al., 2016; Kertesz, 2017; Kuehn, 2007). Unfortunately, increased opioid prescribing has also been paralleled by increases in opioid misuse and diversion (Okie, 2010). Recent data show that the rate of opioid abuse has increased, and deaths from opioid overdose have been labeled a national epidemic (Centers for Disease Control Prevention, 2012; Rudd et al., 2016).

To assist providers in determining the risk of aberrant drug-related behavior (ADB) among chronic pain patients, screening questionnaires have been developed. One commonly used questionnaire is the Screener and Opioid Assessment for Patients with Pain—Revised (SOAPP-R), a self-report instrument that classifies respondents as high or low risk for ADB based on a prescribed cutoff (Butler et al., 2008, 2009). The SOAPP-R is a modified version of the original Screener and Opioid Assessment for Patients with Pain (SOAPP) (Butler et al., 2004); the SOAPP-R was empirically derived (as opposed to the SOAPP, which was conceptually derived) and designed to contain more items that are less transparent in their scoring (Butler et al., 2008). The SOAPP-R exhibited sound psychometric characteristics in its validation and cross-validation studies (Butler et al., 2008, 2009).

* Corresponding author.
E-mail addresses: matthew.finkelman@tufts.edu (M.D. Finkelman), rjamison@bwh.harvard.edu (R.N. Jamison), rkulich@mgh.harvard.edu (R.J. Kulich), sfbutler@inflexxion.com (S.F. Butler), wcjackson@mgh.harvard.edu (W.C. Jackson), n.smits@uva.nl (N. Smits), sweiner@bwh.harvard.edu (S.G. Weiner).

http://dx.doi.org/10.1016/j.drugalcdep.2017.04.016
Received 16 January 2017; Received in revised form 17 April 2017; Accepted 26 April 2017
Available online 13 June 2017
0376-8716/

At 24 items, the SOAPP-R's length is manageable for many patients; nevertheless, the introduction of shorter versions may save time and improve utilization rates (Finkelman et al., 2015; Finkelman et al., 2017b). Indeed, the importance of test length is highlighted by the Scientific Advisory Committee of the Medical Outcomes Trust's categorization of respondent and administrative burden as a key attribute of a health questionnaire (Aaronson et al., 2002). While an abbreviated form of the original SOAPP has been developed (Koyyalagunta et al., 2013), shorter versions of the empirically derived SOAPP-R were desired, leading to two recent studies that suggested different approaches to reducing the length of the latter screener. The simpler of the two approaches is to use a static short form containing a subset of the SOAPP-R items; a retrospective study found that such a static short form consisting of 12 items exhibited sensitivity and specificity comparable to those of the full-length SOAPP-R (Finkelman et al., 2017b. The more complex approach is to administer the SOAPP-R via computer, track an individual's responses as he/she proceeds through the test, and stop the assessment if a computer algorithm determines that further items are unnecessary. For instance, if a respondent's item scores are high enough that he/she reaches the cutoff during testing—or low enough that it has become impossible for him/her to reach the cutoff—the test can be terminated and the appropriate classification can be made. This type of stopping rule, which has been well-studied in the psychometric literature, is referred to as *curtailment* or the *countdown method* (Ben-Porath et al., 1989; Butcher et al., 1985; Finkelman et al., 2015; Forbey et al., 2012). A variation on the above approach is to terminate testing if the SOAPP-R's classification ("high risk" or "low risk") has been determined with certainty from a respondent's previous answers, *or* if the classification has been determined up to a specified level of probability. This variation is referred to as *stochastic curtailment* (Finkelman et al., 2015). A retrospective study found that curtailment and stochastic curtailment produce considerable reductions in average test length while maintaining sensitivity and specificity similar to those of the full-length SOAPP-R (Finkelman et al., 2015).

While previous studies suggested the potential for shorter versions of the SOAPP-R (Finkelman et al., 2015; Finkelman et al., 2017b), their conclusions are limited by the fact that the results of each were based on a single dataset (which was common to both studies). Moreover, the research on the 12-item static short form indicated unstable results with respect to specific cutoffs, and recommended that the preliminary cutoff for this short form ($\geq$10 points) be validated in further study (Finkelman et al., 2017b). Both of the previous studies on short versions of the SOAPP-R emphasized that cross-validation should be conducted in other populations (Finkelman et al., 2015; Finkelman et al., 2017b). The objective of this study was to compare the static short form, curtailment, stochastic curtailment, and the full-length SOAPP-R in two different populations.

2. Material and methods

This retrospective study examined the performance of the full-length SOAPP-R and its short versions using two separate datasets. The Tufts Health Sciences Institutional Review Board granted exempt status or non-human subjects research status for the analysis of each dataset.

2.1. Versions of the SOAPP-R

2.1.1. Full-length form

The 24 items comprising this form are listed in Table 1. Each item can be answered "Never," "Seldom," "Sometimes," "Often," or "Very Often;" the scores for these answer choices are 0, 1, 2, 3, and 4, respectively. The total score on the SOAPP-R is the sum of the item scores; a higher total score indicates greater risk of ADB. The validation and cross-validation studies of the screener recommended a cutoff of $\geq$18 (Butler et al., 2008, 2009).

2.1.2. Curtailment

A curtailment rule is conducted on a fixed order of items. It stops testing once the screener's result ("high risk" or "low risk") has become determined from the respondent's previous answers. Applying this rule to the SOAPP-R with a $\geq$18 cutoff, a computerized version of the questionnaire would stop presenting items (in favor of a "high risk" result) if the respondent's cumulative score reached or exceeded 18. It would also stop presenting items (in favor of a "low risk" result) if the respondent's total score could not reach 18 even if the respondent answered "Very Often" to all remaining items.

Curtailment is a *variable-length testing* method: it produces different test lengths for different respondents. The number of items that a given respondent receives is dependent on his/her answers. In particular, a respondent whose screening result is determined quickly will receive a shorter test length than a respondent whose screening result is not determined until a large number of items have been presented. The maximum number of items that curtailment can administer is equal to the number of items on the full-length screener (24 items for the SOAPP-R). The minimum number of items that curtailment can administer depends on the cutoff; Section 3.3 will present the minimum possible number of items for the particular cutoff used in this study.

2.1.3. Stochastic curtailment

Like curtailment, stochastic curtailment is conducted on a fixed order of items. In stochastic curtailment of the SOAPP-R, early stopping occurs not only if the screener's result has become determined from previous answers, but also if the probability of obtaining one of the results ("high risk" or "low risk") has become adequately high. For the SOAPP-R, previous research (Finkelman et al., 2015) recommended setting the stopping threshold at 99% or 95% (i.e., terminating the screener if the probability of one of the results becomes at least 99%, or if it becomes at least 95%). The use of stochastic curtailment with the former threshold will be referred to as SC-99; its use with the latter threshold will be referred to as SC-95. At each stage of testing, the probability of a "high risk" result, based on the respondent's previous answers, is estimated based on a logistic regression model; see Finkelman et al. (2012) for details. As will be explained in Section 3.3, the set of scores that result in early stopping via stochastic curtailment, at each stage of testing, can be written as a simple look-up table. Finkelman et al. (2015) presented such look-up tables for curtailment, SC-99, and SC-95, but their tables are only applicable when a $\geq$19 cutoff is used. Therefore, in the current study, the data from Finkelman et al. (2015) were re-analyzed to produce look-up tables using the standard $\geq$18 SOAPP-R cutoff.

In sum, stochastic curtailment is a variable-length testing method that is less conservative than curtailment. As in curtailment, the number of items presented to a respondent by stochastic curtailment is based on the respondent's pattern of answers. The maximum number of items that stochastic curtailment can administer, when used in conjunction with the SOAPP-R, is 24; the minimum number of items is dependent on the cutoff. See Section 3.3 for the minimum possible number of items that SC-99 and SC-95 can administer when applied to the SOAPP-R with the cutoff used in this study.

2.1.4. Static short form

The development of the static short form of the SOAPP-R (i.e., the selection of items for this form) was based on both (i) statistical criteria and (ii) a scrutiny of content by an external set of pain practitioners (Finkelman et al., 2017b). The statistical component utilized data from 428 individuals who had taken the full-length SOAPP-R, and had also been classified as "negative" or "positive" for ADB by the Aberrant Drug Behavior Index (ADBI), as part of the screener's original validation study or cross-validation study (Butler et al., 2008, 2009). Using these data, candidate short forms of different lengths were developed and evaluated. In particular, for every test length of fewer than 24 items (i.e., for each test length between one item and 23 items), a candidate

Table 1
Statistics about each SOAPP-R item, the 12-item short form, and the full-length SOAPP-R, by study (n = 84 for Study 1, n = 110 for Study 2).

Item	12-Item Short Form[a]	Study 1	Study 2
		Mean (SD)	Mean (SD)
1. How often do you have mood swings?		1.7 (1.0)	2.0 (1.1)
2. How often have you felt a need for higher doses of medication to treat your pain?	**YES**	1.8 (1.0)	2.1 (1.2)
3. How often have you felt impatient with your doctors?	**YES**	1.2 (1.1)	1.1 (1.2)
4. How often have you felt that things are just too overwhelming that you can't handle them?	**YES**	1.4 (1.2)	1.9 (1.3)
5. How often is there tension in the home?	**YES**	1.4 (1.0)	1.4 (1.2)
6. How often have you counted pain pills to see how many are remaining?		1.1 (1.1)	1.0 (1.2)
7. How often have you been concerned that people will judge you for taking pain medication?		1.5 (1.4)	1.3 (1.4)
8. How often do you feel bored?		1.6 (1.3)	1.9 (1.3)
9. How often have you taken more pain medication than you were supposed to?	**YES**	1.0 (0.9)	0.7 (0.9)
10. How often have you worried about being left alone?		0.7 (1.1)	1.0 (1.3)
11. How often have you felt a craving for medication?		0.6 (0.8)	0.4 (0.7)
12. How often have others expressed concern over your use of medication?	**YES**	0.8 (1.0)	0.6 (0.9)
13. How often have any of your close friends had a problem with alcohol or drugs?		1.0 (0.9)	0.9 (1.1)
14. How often have others told you that you had a bad temper?		0.5 (0.8)	0.8 (1.1)
15. How often have you felt consumed by the need to get pain medication?		0.6 (0.9)	0.8 (1.1)
16. How often have you run out of pain medication early?	**YES**	0.8 (1.0)	0.6 (0.9)
17. How often have others kept you from getting what you deserve?		0.5 (0.7)	0.7 (1.1)
18. How often, in your lifetime, have you had legal problems or been arrested?	**YES**	0.4 (0.6)	0.6 (0.8)
19. How often have you attended an AA or NA meeting?	**YES**	0.6 (1.0)	0.5 (0.9)
20. How often have you been in an argument that was so out of control that someone got hurt?		0.2 (0.6)	0.2 (0.5)
21. How often have you been sexually abused?	**YES**	0.3 (0.8)	0.4 (0.8)
22. How often have others suggested that you have a drug or alcohol problem?	**YES**	0.4 (0.7)	0.3 (0.7)
23. How often have you had to borrow pain medications from your family or friends?		0.2 (0.6)	0.1 (0.3)
24. How often have you been treated for an alcohol or drug problem?	**YES**	0.3 (0.7)	0.3 (0.7)
Total Score (12-Item Short Form)		10.3 (6.0)	10.3 (5.6)
Total Score (Full-Length SOAPP-R)		20.4 (11.8)	21.3 (10.9)

SD = Standard deviation.
SOAPP-R = Screener and Opioid Assessment for Patients with Pain-Revised.
[a] Items labeled with a "**YES**" are included in the 12-item static short form.

short form was created to best predict the ADBI. The specific statistical methodology utilized to produce these short forms of different lengths was *least absolute shrinkage and selection operator* (LASSO) logistic regression. For each given test length, the LASSO logistic regression procedure selected the specified number of items that, when used in combination with one another, were most predictive of the ADBI. The forms of different lengths were then compared to each other in terms of statistical characteristics such as their sensitivities, specificities, and area under the curve (AUC) values. The use of these statistics is consistent with previous studies that focused on the predictive validity of the SOAPP-R (Butler et al., 2004, 2008, 2009). Based on the statistical results, five candidate short forms (consisting of seven items, nine items, 10 items, 11 items, and 12 items) were chosen for additional evaluation. No form with more than 12 items was chosen because the inclusion of more than 12 items did not result in greater predictive value; in fact, for the dataset of n = 428 individuals, the 12-item form had a higher AUC than every candidate form with more than 12 items. In the second component of evaluation, the five aforementioned candidate short forms, as well as the full-length SOAPP-R, were scrutinized in terms of their content by 12 pain practitioners at a pain care center. The pain practitioners were asked to provide feedback on the content of the different forms, and each practitioner stated which of the six forms he/she would be most likely to use with his/her own patients. The majority of practitioners (nine out of 12) indicated that they would be most likely to use the 12-item form. Participants selecting this form alluded to test length, respondent burden, and/or content coverage when explaining the rationale for their decision. In sum, on the basis of both the statistical results and the evaluation by pain practitioners, the 12-item short form was recommended for further research in different populations, leading to the cross-validation of this form (and comparison with other forms) in the current study. The reader is referred to Finkelman et al. (2017b) for further details on the development of the static short form.

Table 1 indicates the 12 items comprising the static short form of the SOAPP-R. A respondent's total score on the short form is obtained by summing his/her scores on the items. A 50% decrease in the number of items administered, as provided by the SOAPP-R short form in comparison with the full-length form, had been characterized by work in other fields as a substantial reduction in test length (Leidy and Knebel, 2010; Marsh et al., 2005).

2.2. Sources of data

2.2.1. Study 1

Patients who were diagnosed with chronic noncancer neck or back pain with or without radicular pain were recruited for this 6-month study. This was a prospective, randomized, controlled trial designed to test an intervention to improve compliance in chronic pain patients who were misusing their medication (Jamison et al., 2014). All patients were recruited from an urban tertiary university hospital pain center. Enrollment occurred between November 2007 and July 2011. Participants completed multiple questionnaires including the paper-and-pencil version of the full-length SOAPP-R. Similar to the original SOAPP-R studies (Butler et al., 2008, 2009), the patients completed the Prescription Drug Use Questionnaire (PDUQ) to determine self-reported misuse. Urine toxicology screening results were also assessed, and physician-reported aberrant behavior was determined using the Addiction Behavior Checklist (ABC)(Wu et al., 2006). Positive results on any of these tests indicated ADB. See Jamison et al. (2014) for further details.

2.2.2. Study 2

Patients with a diagnosis of chronic nonmalignant pain and treated in one of eight primary care centers were recruited for this 6-month trial (Jamison et al., 2016). Enrollment occurred between September 2012 and September 2014. All subjects completed multiple questionnaires including the paper-and-pencil version of the full-length SOAPP-R. Patient self-reported misuse was assessed using the Current

Opioid Misuse Measure (COMM)(Butler et al., 2010). Physician-reported misuse was assessed using the ABC (Wu et al., 2006), and urine toxicology results were assessed for abnormal findings. A positive ADB classification was made if the urine toxicology results were positive (evidence of tampering, evidence of an opioid that was not prescribed, or an illicit substance such as cocaine), or if both the self-report and physician-reported results were positive (a self-reported COMM of 9 or higher and a physician-reported ABC of 2 or higher). See Jamison et al. (2016) for further details.

2.3. Statistical analysis

A separate analysis was conducted for each of the two datasets, since the datasets arose from different populations. In each analysis, statistical characteristics of the full-length SOAPP-R (sensitivity and specificity in detecting ADB, using the standard ≥18 cutoff) were calculated. Additionally, a *real-data simulation* was conducted to assess the performance of the short versions of the SOAPP-R in each of the two studies. In the real-data simulation, an analysis was undertaken to determine what the sensitivity and specificity *would have been* in each study, if a given short version (curtailment, SC-99, SC-95, or the static short form) had been used. The real-data simulation also determined the mean and standard deviation of test length (i.e., number of items administered) that would have been obtained, if curtailment, SC-99, or SC-95 had been used. The look-up tables providing stopping rules for SC-99 and SC-95 were based on re-analyzing data from Finkelman et al. (2015) using a ≥18 points cutoff, as described in Section 2.1.3. Because previous research on the static short form (Finkelman et al., 2017b) recommended further evaluation of its preliminarily suggested cutoff (≥10 points), statistics for each possible cutoff within two of the preliminary cutoff (i.e., from ≥8 to ≥12 points) were computed.

In addition to the above analysis, receiver operating characteristic (ROC) curves were calculated, and the AUC statistic was computed, to measure the overall ability of each form to predict ADB. Item statistics were also obtained for each study. In particular, the mean score and SD were calculated for each item (as well as for the full-length SOAPP-R and static short form). A measure of effect size—specifically, Cohen's *d*, which had previously been used in the original validation study of the SOAPP-R (Butler et al., 2008)—was calculated by item as well. A higher Cohen's *d* value indicates greater ability for an item to discriminate between individuals with ADB—according to a given study's criterion for ADB—from individuals without ADB, according to the same criterion.

To investigate whether the SOAPP-R items could be reordered to produce greater efficiency in variable-length testing, further real-data simulations were performed. First, 10,000 random orderings of the SOAPP-R items were generated. Then, for each dataset (i.e., the dataset of Study 1 and the dataset of Study 2), the mean test length produced by coupling each random item ordering with the curtailment stopping rule was found by real-data simulation. Curtailment was chosen as the stopping rule, rather than stochastic curtailment, because in the former (but not in the latter), the sensitivity and specificity are always the same regardless of the item ordering. Therefore, with curtailment, the sensitivity and specificity are standardized, so the different orderings can be compared solely by their mean test lengths. The mean test lengths produced by coupling the random item orderings with curtailment were compared to the mean test length produced by using curtailment alongside the standard (i.e., conventional booklet) item ordering of the SOAPP-R. The standard ordering was also compared to other item orderings of interest. In particular, Finkelman et al. (2017a) recently found that when a curtailment stopping rule is used for a given questionnaire, an item ordering that is optimally efficient, or is among the most efficient orderings, can be produced via the following steps: (i) using real-data simulation, determine the mean test length obtained by ordering the items from highest mean score to lowest mean score; (ii) using real-data simulation, determine the mean test length obtained by ordering the items from lowest mean score to highest mean score; (iii) of the two item orderings produced in (i) and (ii) above, select the one that had the smaller mean test length in the real-data simulations. Therefore, for each dataset, the SOAPP-R items were placed in order from the item with the highest mean to the item with the lowest mean, as well as from the item with the lowest mean to the item with the highest mean, and the mean test length was computed for each of these two orderings. Based on the results of Finkelman et al. (2017a) Finkelman et al. (2017a), it was anticipated that one of these two orderings would be among the most efficient orderings examined, if not the most efficient ordering. We note that the standard SOAPP-R item ordering also places the items in approximate order from highest to lowest mean score (Butler et al., 2008); therefore, it was anticipated that the standard item ordering would be approximately as efficient as the "highest mean score to lowest mean score" ordering. Finally, the SOAPP-R items were ordered from highest Cohen's *d* value to lowest Cohen's *d* value, and the mean test length for this ordering was calculated. The latter ordering was examined to investigate whether ordering the items by ability to discriminate between respondents with and without ADB would result in high efficiency. The mean test lengths of all of the above item orderings were then compared. Results were similar in both datasets. In each dataset, the "highest mean score to lowest mean score" ordering was more efficient (had a lower mean test length) than the "lowest mean score to highest mean score" ordering. Consistent with the results of Finkelman et al., 2017aFinkelman et al. (2017a) the better of these two orderings (i.e., "highest mean score to lowest mean score") was more efficient than all 10,000 random orderings, and was also more efficient than ordering the items by Cohen's *d*. As anticipated, the standard SOAPP-R item ordering performed similarly to the "highest mean score to lowest mean score" ordering (their mean test lengths were within 0.13 items of each other in both datasets). The standard SOAPP-R ordering was also more efficient than all 10,000 random orderings, and more efficient than the Cohen's *d* ordering, in both datasets. As the standard SOAPP-R ordering was thus an efficient ordering, and is also the ordering that is typically used in practical administrations of the screener, only the results of this ordering are presented in the sequel.

3. Results

3.1. Demographics and ADB results

3.1.1. Study 1

The mean (SD) age among the $n = 84$ subjects was 49.9 (8.8) years. Forty-five subjects (53.6%) were male. Forty-three (51.2%) were positive for ADB.

3.1.2. Study 2

Of the $n = 110$ subjects, 109 had information on age; among these, the mean (SD) age was 53.4 (9.5) years. Sixty-nine of the 110 subjects (62.7%) were female. Forty of the 110 subjects (36.4%) were positive for ADB.

3.2. Item and test statistics

Table 1 presents item statistics, as well as statistics for the static 12-item short form and the full-length SOAPP-R, for both studies. The mean score on the static short form was the same for Study 1 and Study 2. The mean score on the full-length SOAPP-R was slightly higher for Study 2 than for Study 1.

For Study 1, the items with the highest Cohen's *d* values were Item 16 ($d = 0.790$), Item 11 ($d = 0.786$), Item 9 ($d = 0.714$), and Item 19 ($d = 0.591$). For Study 2, the items with the highest values were Item 23 ($d = 0.618$), Item 9 ($d = 0.589$), Item 8 ($d = 0.575$), and Item 18 ($d = 0.480$)(data not shown).

Table 2
Stopping boundaries of curtailment, SC-99, and SC-95, using a cutoff of ≥18.

	Curtailment		SC-99		SC-95	
Stage of Testing	Stop for "Low Risk" Result	Stop for "High Risk" Result	Stop for "Low Risk" Result	Stop for "High Risk" Result	Stop for "Low Risk" Result	Stop for "High Risk" Result
1	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	X = 0	X ≥ 7
3	NA	NA	NA	X ≥ 11	X ≤ 1	X ≥ 9
4	NA	NA	NA	X ≥ 13	X ≤ 2	X ≥ 11
5	NA	X ≥ 18	X ≤ 1	X ≥ 14	X ≤ 3	X ≥ 12
6	NA	X ≥ 18	X ≤ 2	X ≥ 15	X ≤ 4	X ≥ 13
7	NA	X ≥ 18	X ≤ 3	X ≥ 16	X ≤ 5	X ≥ 14
8	NA	X ≥ 18	X ≤ 4	X ≥ 18	X ≤ 6	X ≥ 15
9	NA	X ≥ 18	X ≤ 4	X ≥ 18	X ≤ 7	X ≥ 16
10	NA	X ≥ 18	X ≤ 5	X ≥ 18	X ≤ 7	X ≥ 17
11	NA	X ≥ 18	X ≤ 6	X ≥ 18	X ≤ 8	X ≥ 18
12	NA	X ≥ 18	X ≤ 6	X ≥ 18	X ≤ 8	X ≥ 18
13	NA	X ≥ 18	X ≤ 8	X ≥ 18	X ≤ 10	X ≥ 18
14	NA	X ≥ 18	X ≤ 9	X ≥ 18	X ≤ 11	X ≥ 18
15	NA	X ≥ 18	X ≤ 9	X ≥ 18	X ≤ 11	X ≥ 18
16	NA	X ≥ 18	X ≤ 9	X ≥ 18	X ≤ 11	X ≥ 18
17	NA	X ≥ 18	X ≤ 10	X ≥ 18	X ≤ 12	X ≥ 18
18	NA	X ≥ 18	X ≤ 10	X ≥ 18	X ≤ 12	X ≥ 18
19	NA	X ≥ 18	X ≤ 11	X ≥ 18	X ≤ 13	X ≥ 18
20	X ≤ 1	X ≥ 18	X ≤ 12	X ≥ 18	X ≤ 14	X ≥ 18
21	X ≤ 5	X ≥ 18	X ≤ 13	X ≥ 18	X ≤ 15	X ≥ 18
22	X ≤ 9	X ≥ 18	X ≤ 14	X ≥ 18	X ≤ 15	X ≥ 18
23	X ≤ 13	X ≥ 18	X ≤ 14	X ≥ 18	X ≤ 16	X ≥ 18
24	X ≤ 17	X ≥ 18	X ≤ 17	X ≥ 18	X ≤ 17	X ≥ 18

NA = Not Applicable. A cell marked "NA" indicates that there is no possibility of early stopping associated with that cell.
SC–95 = Stochastic curtailment with a stopping threshold of 95%.
SC–99 = Stochastic curtailment with a stopping threshold of 99%.
X = Respondent's cumulative score at a given stage of testing.

3.3. Look-up tables

Table 2 presents the stopping rules of curtailment, SC-99, and SC-95, based on the re-analysis of previously examined SOAPP-R data (Finkelman et al., 2015) using the traditional SOAPP-R cutoff of ≥18. The table shows the set of cumulative scores that result in early stopping at each stage of testing. For example, at the seventh stage of testing (i.e., after the respondent has answered seven items), curtailment never stops early for a "low risk" result (as indicated by an "NA" in the table). However, this method stops early for a "high risk" result if the respondent's cumulative score after seven items (denoted "X" in the table) is 18 or above.

As demonstrated in Table 2, the minimum number of items that curtailment can administer (when applied to the SOAPP-R with a cutoff of ≥18) is 5, and the maximum possible number of items administered is 24. For SC-99, the set of possible test lengths ranges from 3 to 24; for SC-95, it ranges from 2 to 24. The minimum possible test lengths stated above are consistent with curtailment's status as the most conservative variable-length testing method, and SC-95's status as the most liberal.

3.4. Comparison of SOAPP-R forms

3.4.1. Study 1

As shown in Table 3, the sensitivity and specificity of the full-length SOAPP-R (using a ≥18 cutoff) were 0.67 and 0.59, respectively. Curtailment and SC-99 exhibited the same sensitivity and specificity as the full-length SOAPP-R; however, curtailment reduced the mean test length to 16.8 items (SD = 6.3), while SC-99 reduced the mean test length to 13.1 items (SD = 5.8). SC-95 had a sensitivity 0.02 lower than that of the full-length SOAPP-R, and a specificity 0.03 lower; the mean test length of SC-95 was 9.6 items (SD = 5.6). The static 12-item short form exhibited a sensitivity equal to that of the full-length SOAPP-R, and a specificity 0.03 lower, when a ≥9 cutoff was used. The previously suggested cutoff for the short form (≥ 10) resulted in a sensitivity 0.07 lower than that of the full-length SOAPP-R, and a specificity 0.02 higher. The AUC of the full-length SOAPP-R, curtailment, SC-99, and the static short form was 0.71; the AUC of SC-95 was 0.69.

3.4.2. Study 2

The sensitivity and specificity of the full-length SOAPP-R were 0.68 and 0.44, respectively (Table 3). As in Study 1, curtailment and SC-99 had sensitivity and specificity values identical to those of the full-length SOAPP-R. SC-95 had the same sensitivity as the full-length SOAPP-R, and a specificity 0.03 higher. Mean (SD) test lengths were 15.9 (6.6), 12.8 (6.2), and 8.8 (6.5) items for curtailment, SC-99, and SC-95, respectively. Using a ≥9 cutoff, the static short form had a sensitivity equal to that of the full-length SOAPP-R, and a specificity 0.02 higher. Using a ≥10 cutoff, the short form's sensitivity was 0.08 lower than that of the full-length SOAPP-R, while its specificity was 0.07 higher. The AUC of the full-length SOAPP-R, curtailment, and the static short form was 0.62; the AUC of SC-99 was 0.61, and the AUC of SC-95 was 0.60.

4. Discussion

Previous research, based on the analysis of a single dataset, indicated the potential of curtailment, stochastic curtailment, and the static 12-item short form to enhance the efficiency of the SOAPP-R (Finkelman et al., 2015; Finkelman et al., 2017b. However, prior to the current study, there was no evidence to indicate whether such findings would generalize to different populations. The mean item savings of curtailment and stochastic curtailment observed herein, as well as the methods' ability to maintain sensitivity and specificity close or equal to those of the SOAPP-R, were similar to previous study (Finkelman et al., 2015). Moreover, the fact that stochastic curtailment's stopping rules could be derived from one dataset and applied to other populations, without unduly compromising sensitivity and specificity, confirms the method's robustness as applied to the SOAPP-R. Turning to the static short form, prior findings (Finkelman et al., 2017b that this form's overall discriminatory power, as measured by the AUC, is comparable to that of the full-length SOAPP-R were also supported by the current study; in fact, the AUC of the short form was equal to that of the full-length SOAPP-R in both populations considered herein. In sum, the results of this study suggest that the short versions exhibit screening characteristics comparable to those of the full-length SOAPP-R in different settings, while providing considerably reduced test lengths.

The discriminatory power of the different SOAPP-R forms (including the full-length form) was modest in the current study. In both of the populations, the AUC value of the full-length SOAPP-R was lower than the AUCs observed in the form's validation and cross-validation studies (Butler et al., 2008, 2009). Additionally, the AUC values of both the full-length SOAPP-R and the static short form were lower than those obtained in the study in which the static short form was developed (Finkelman et al., 2017b, and the sensitivity and specificity of curtailment and stochastic curtailment were generally lower than those found in previous study (Finkelman et al., 2015). One possible reason for this finding is that the definitions of ADB used in the current study differed from those of prior research. As originally derived, the SOAPP-R was designed to detect aberrant behavior as defined by the ADBI, which was composed of triangulated data from patient self-reported misuse on the PDUQ, physicians' reporting of patient aberrant behavior on the Prescription Opioid Therapy Questionnaire (POTQ), and urine toxicology screening showing evidence of illicit substances. Our outcome measures were different. Indeed, in our studies, similar constructs were assessed to determine a positive or negative classification for ADB as were used in the original validation studies (patient self-report, physician report, and urine toxicology results), but the specific set of validated measures

Table 3
Sensitivities, specificities, areas under curve, and test length statistics, by study (n = 84 for Study 1, n = 110 for Study 2).

	Study 1				Study 2			
	Sensitivity	Specificity	AUC	Mean (SD) Test Length	Sensitivity	Specificity	AUC	Mean (SD) Test Length
Full-Length SOAPP-R (≥18 Cutoff)	0.67	0.59	0.71	24.0 (0.0)	0.68	0.44	0.62	24.0 (0.0)
Curtailment	0.67	0.59	0.71	16.8 (6.3)	0.68	0.44	0.62	15.9 (6.6)
SC-99	0.67	0.59	0.71	13.1 (5.8)	0.68	0.44	0.61	12.8 (6.2)
SC-95	0.65	0.56	0.69	9.6 (5.6)	0.68	0.47	0.60	8.8 (6.5)
12-Item SOAPP-R (≥8 Cutoff)	0.81	0.44	0.71	12.0 (0.0)	0.78	0.40	0.62	12.0 (0.0)
12-Item SOAPP-R (≥9 Cutoff)	0.67	0.56	0.71	12.0 (0.0)	0.68	0.46	0.62	12.0 (0.0)
12-Item SOAPP-R (≥10 Cutoff)	0.60	0.61	0.71	12.0 (0.0)	0.60	0.51	0.62	12.0 (0.0)
12-Item SOAPP-R (≥11 Cutoff)	0.60	0.66	0.71	12.0 (0.0)	0.55	0.56	0.62	12.0 (0.0)
12-Item SOAPP-R (≥12 Cutoff)	0.51	0.78	0.71	12.0 (0.0)	0.48	0.66	0.62	12.0 (0.0)

AUC = Area under the curve.
SC–95 = Stochastic curtailment with a stopping threshold of 95%.
SC–99 = Stochastic curtailment with a stopping threshold of 99%.
SD = Standard deviation.
SOAPP-R = Screener and Opioid Assessment for Patients with Pain-Revised.

and rules to determine this classification were not identical. Different definitions of aberrant behaviors (i.e., "the target to be predicted") can have a significant impact on estimates of the predictive validity of the SOAPP-R. The lack of a single "gold standard" measure of ADB is, therefore, a limitation of the current study; however, the fact that the short versions of the SOAPP-R consistently performed well in comparison to the complete screener, even when different measures of ADB were used, may provide evidence of these versions' utility as brief alternatives to the full-length SOAPP-R.

Which short version of the SOAPP-R to use in practice may depend on the specific administration at hand. For instance, if computer-based testing is not available at a given setting, then the static 12-item screener (which can be administered via paper-and-pencil) is the only feasible short form. On the other hand, if it is possible for the screener to be given via computer and stopped sequentially according to the rules of curtailment and stochastic curtailment, then these techniques may provide greater concordance with the full-length SOAPP-R than the static short form while still offering item savings. Indeed, curtailment is guaranteed to provide the same result ("high risk" or "low risk") as the full-length SOAPP-R for every subject, and SC-99 had the same sensitivity and specificity as the full-length SOAPP-R in each of the two specific populations considered herein. SC-95, however, exhibited the greatest average item savings, reducing the mean test length by 60% to 63% in the two studies (as compared with 30% to 34% for curtailment, 45% to 47% for SC-99, and a constant 50% for the static short form). When deciding between short versions, practitioners should consider the tradeoff between item savings and concordance with the full-length SOAPP-R, with curtailment being the most conservative option and SC-95 stopping most liberally. We note that when a ≥18 cutoff is used for the full-length SOAPP-R, as was recommended in the screener's validation and cross-validation studies (Butler et al., 2008, 2009) and was done herein, curtailment has a greater maximum potential for item savings among "high risk" respondents than among "low risk" respondents. Indeed, as can be seen in Table 2, curtailment can stop after as few as five items (i.e., the fifth stage of testing) for a "high risk" result when a ≥18 cutoff is used; it cannot stop until at least the twentieth stage of testing for a "low risk" result when this same cutoff is employed. Table 2 also shows that SC-99 can stop as early as the third stage of testing for a "high risk" result, and the fifth stage for a "low risk" result; SC-95 can stop as early as the second stage of testing for both a "high risk" result and a "low risk" result. The mean test lengths provided in Table 3 can be used to gauge the overall level of item savings of each method when applied to both "high risk" and "low risk" respondents in the two studies.

In settings where the static short form is used, the decision of which cutoff to employ alongside that form is clearly consequential. Although a ≥10 cutoff was preliminarily suggested in a previous study for the static short form (Finkelman et al., 2017b, the results of the current study suggest that using a ≥9 cutoff for this form may lead to sensitivity and specificity more similar to those of the full-length SOAPP-R (when the standard ≥18 cutoff is used for the latter). Employing a ≥9 cutoff rather than a ≥10 cutoff could only improve the sensitivity of the form; while this improvement would come at the expense of its specificity, previous research on the SOAPP-R has stressed the importance of sensitivity relative to specificity (Butler et al., 2004, 2008, 2009). Therefore, one inference from the current study's results is that the proposed cutoff of the static short form, which had been mentioned specifically as needing further validation in previous research (Finkelman et al., 2017b may need to be modified. Further comparison of the ≥9 cutoff and the ≥10 cutoff may be valuable.

Aside from the lack of a gold standard measure of ADB, limitations of the current study include its retrospective nature. Indeed, results obtained via real-data simulation do not necessarily generalize to live administrations of a given screener. The static short form, for instance, was not administered prospectively to subjects as a contiguous block of items; rather, this form was evaluated post-hoc, from the data of subjects who had completed the full-length SOAPP-R, by identifying the items comprising the short form and calculating their sensitivity and specificity as if they had been presented as a unit. It is possible that due to context effects, the screening characteristics of the short form would be different when its items are presented as a contiguous block. Additionally, results obtained when a screener is administered via paper-and-pencil do not necessarily apply when the screener is given via computer, and vice versa. Prospective studies should be conducted to validate the results found in the real-data simulation. A further limitation of the current study was the relatively small sample size in each of the datasets employed. However, the use of two datasets from different populations, each of which independently suggested the potential of the short versions of the SOAPP-R, may mitigate this limitation to some degree. Finally, an additional limitation is that data on the length of time taken on each item and form were not recorded in either study. Therefore, the average amount of time spent on the full-length SOAPP-R could not be computed, nor could the amount of time that would be saved by administering a shorter version. Reduction in test length itself is a commonly used measure of the savings provided by shortened instruments (e.g., Finkelman et al., 2015; Leidy and Knebel, 2010; Rudick et al., 2013; Smits et al., 2011; Thompson, 2007, 2011); nevertheless, future studies should record participants' response times so that time savings can be calculated as well. Even a small time savings for some individual patients would cumulatively reduce administrative burden.

From a clinician's perspective, any opportunity to shorten screening

tools is likely to be welcome, given the increasing time constraints. Computerized testing is likely to revolutionize screening procedures as well. Further study will elucidate patient preference for shorter versus longer forms, computer- or paper-based screeners, and the true impact of curtailment, stochastic curtailment, and the static short form on the ultimate goal: identification of high-risk patients that need further intervention.

5. Conclusions

Results suggest the potential of curtailment, stochastic curtailment, and the 12-item static short form of the SOAPP-R to enhance the efficiency of this screener in two different patient populations. The short versions, including computerized implementation with real-time curtailment and stochastic curtailment techniques, should be tested in prospective studies.

Conflict of interest

SFB is an employee of Inflexxion, Inc. Inflexxion holds the copyright for the Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP®-R).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Role of funding source

Nothing declared.

Contributors

MDF made a significant contribution to the study concept and design, the analysis and interpretation of data, and the drafting/revising of the manuscript for important intellectual content. RNJ made a significant contribution to the study concept and design, the acquisition of data, the interpretation of data, and the drafting/revising of the manuscript for important intellectual content. RJK made a significant contribution to the study concept and design, the interpretation of data, and the drafting/revising of the manuscript for important intellectual content. SFB made a significant contribution to the study concept and design, the interpretation of data, and the drafting/revising of the manuscript for important intellectual content. WCJ made a significant contribution to the interpretation of data, as well as the drafting/revising of the manuscript for important intellectual content. NS made a significant contribution to the analysis and interpretation of data, as well as the drafting/revising of the manuscript for important intellectual content. SGW made a significant contribution to the study concept and design, the interpretation of data, and the drafting/revising of the manuscript for important intellectual content. All authors approved the final article.

References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K.N., Patrick, D.L., Perrin, E., Stein, R.E., 2002. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual. Life Res. 11, 193–205.

Aitken, M., Kleinrock, M., Pennente, K., Lyle, J., Nass, D., Caskey, L., 2016. Medicines Use and Spending in the U.S.: A Review of 2015 and Outlook To 2020. IMS Institute for Healthcare Informatics, Parsippany, NJ.

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

Butcher, J.N., Keller, L.S., Bacon, S.F., 1985. Current developments and future directions in computerized personality assessment. J. Consult. Clin. Psychol. 53, 803–815. http://dx.doi.org/10.1037/0022-006X.53.6.803.

Butler, S.F., Budman, S.H., Fernandez, K., Jamison, R.N., 2004. Validation of a screener and opioid assessment measure for patients with chronic pain. Pain 112, 65–75. http://dx.doi.org/10.1016/j.pain.2004.07.026.

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). J. Pain 9, 360–372. http://dx.doi.org/10.1016/j.jpain.2007.11.014.

Butler, S.F., Budman, S.H., Fernandez, K.C., Fanciullo, G.J., Jamison, R.N., 2009. Cross-validation of a screener to predict opioid misuse in chronic pain patients (SOAPP-R). J. Addict. Med. 3, 66–73. http://dx.doi.org/10.1097/adm.0b013e31818e41da.

Butler, S.F., Budman, S.H., Fanciullo, G.J., Jamison, R.N., 2010. Cross validation of the Current Opioid Misuse Measure to monitor chronic pain patients on opioid therapy. Clin. J. Pain 26, 770–776. http://dx.doi.org/10.1097/ajp.0b013e3181f195ba.

Centers for Disease Control and Prevention (CDC), 2012. CDC grand rounds: prescription drug overdoses − a U.S. epidemic. MMWR Morb. Mortal Wkly. Rep. 61, 10–13.

Finkelman, M.D., Smits, N., Kim, W., Riley, B., 2012. Curtailment and stochastic curtailment to shorten the CES-D. Appl. Psych. Meas. 36, 632–658. http://dx.doi.org/10.1177/0146621612451647.

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 Med. 16, 2344–2356. http://dx.doi.org/10.1111/pme.12864.

Finkelman, M.D., Lowe, S.R., Kim, W., Gruebner, O., Smits, N., 2017a. Item ordering and computerized classification tests with cluster-based scoring: an investigation of the countdown method. Psychol. Assess. http://dx.doi.org/10.1037/pas0000470. (in press).

Finkelman, M.D., Smits, N., Kulich, R.J., Zacharoff, K.L., Magnuson, B.E., Chang, H., Dong, J., Butler, S.F., 2017b. Development of short-form versions of the Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R): A proof-of-principle study. http://dx.doi.org/10.1093/pm/pnw210. (in press).

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. Psychol. Assess. 24, 628–639. http://dx.doi.org/10.1037/a0026509.

Jamison, R.N., Martel, M.O., Edwards, R.R., Qian, J., Sheehan, K.A., Ross, E.L., 2014. Validation of a brief Opioid Compliance Checklist for patients with chronic pain. J. Pain 15, 1092–1101. http://dx.doi.org/10.1016/j.jpain.2014.07.007.

Jamison, R.N., Scanlan, E., Matthews, M.L., Jurcik, D.C., Ross, E.L., 2016. Attitudes of primary care practitioners in managing chronic pain patients prescribed opioids for pain: a prospective longitudinal controlled trial. Pain Med. 17, 99–113. http://dx.doi.org/10.1111/pme.12871.

Kertesz, S.G., 2017. Turning the tide or riptide? The changing opioid epidemic. Subst. Abus. 38, 3–8. http://dx.doi.org/10.1080/08897077.2016.1261070.

Koyyalagunta, D., Bruera, E., Aigner, C., Nusrat, H., Driver, L., Novy, D., 2013. Risk stratification of opioid misuse among patients with cancer pain using the SOAPP-SF. Pain Med. 14, 667–675. http://dx.doi.org/10.1111/pme.12100.

Kuehn, B.M., 2007. Opioid prescriptions soar: increase in legitimate use as well as abuse. JAMA 297, 249–251. http://dx.doi.org/10.1001/jama.297.3.249.

Leidy, N.K., Knebel, A., 2010. In search of parsimony: reliability and validity of the Functional Performance Inventory-Short Form. Int. J. Chron. Obstruct. Pulmon. Dis. 5, 415–423. http://dx.doi.org/10.2147/COPD.S13389.

Marsh, H.W., Ellis, L.A., Parada, R.H., Richards, G., Heubeck, B.G., 2005. A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. Psychol. Assess. 17, 81–102. http://dx.doi.org/10.1037/1040-3590.17.1.81.

Okie, S., 2010. A flood of opioids, a rising tide of deaths. N. Engl. J. Med. 363, 1981–1985. http://dx.doi.org/10.1056/NEJMp1011512.

Rudd, R.A., Aleshire, N., Zibbell, J.E., Gladden, R.M., 2016. Increases in drug and opioid overdose deaths—United States, 2000–2014. MMWR Morb. Mortal. Wkly Rep. 64, 1378–1382. http://dx.doi.org/10.15585/mmwr.mm6450a3.

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

Smits, N., Cuijpers, P., van Straten, A., 2011. Applying computerized adaptive testing to the CES-D scale: a simulation study. Psychiatry Res. 188, 147–155. http://dx.doi.org/10.1016/j.psychres.2010.12.001.

Thompson, N.A., 2007. A practitioner's guide for variable-length computerized classification testing. Prac. Asmnt. Res. Eval. 12 (Available at: http://pareonline.net/pdf/v12n1.pdf).

Thompson, N.A., 2011. Termination criteria for computerized classification testing. Prac. Asmnt. Res. Eval. 16 (Available at: http://pareonline.net/pdf/v16n4.pdf).

Wu, S.M., Compton, P., Bolus, R., Schieffer, B., Pham, Q., Baria, A., Van Vort, W., Davis, F., Shekelle, P., Naliboff, B.D., 2006. The Addiction Behaviors Checklist: Validation of a new clinician-based measure of inappropriate opioid use in chronic pain. J. Pain Symptom Manage. 32, 342–351. http://dx.doi.org/10.1016/j.jpainsymman.2006.05.010.