Self-Report of Waking-State Oral Parafunctional Behaviors in the Natural Environment

Sarah E. F. Kaplan, DDS
Graduate Student
Department of Oral Diagnostic Sciences
University at Buffalo
Buffalo, New York, USA

Richard Ohrbach, DDS, PhD
Associate Professor
Department of Oral Diagnostic Sciences
University at Buffalo
Buffalo, New York, USA

This article is based on the first author's thesis and is in partial fulfillment for Graduation with Thesis Honors at the University at Buffalo School of Dental Medicine. Some of this material has been presented by the first author at the following meetings: 2013 ADA Foundation 49th Annual Dental Students Conference on Research; 2013 IADR/AADR/CADR General Session and Exhibition; 2014 AADR/CADR General Session and Exhibition; 2014 Hinman Student Symposium.

Correspondence to:
Dr Richard Ohrbach
355 Squire
Buffalo, NY 14214, USA
Fax: (716) 829-3554
Email: ohrbach@buffalo.edu

Aims: To determine if retrospective self-report of oral parafunctional behaviors potentially relevant to pain conditions is valid, by comparing oral parafunctional behaviors via a self-report instrument (Oral Behaviors Checklist [OBC]) with in-field reports of oral parafunction. **Methods:** Individuals with a range of oral parafunctional behaviors, as identified by the OBC, were recruited, and 22 completed the field study. Using the Ecological Momentary Assessment paradigm, each subject was randomly prompted about eight times per day, for a target of 7 days, via portable handheld computer to report current behaviors among 11 queried items. Before and after the field study, a paper version of the OBC was administered. Separately, 74 individuals participated in a test-retest study of the paper OBC. Analyses included regression, correlation, intraclass correlation coefficient (ICC) and area under the receiving operating curve (AUC). **Results:** Pre- and postfield study administration of the OBC exhibited substantial reliability (ICC = 0.65), indicating no reactivity during the intervening in-field data collection. Reliability across in-field days was low, indicating high variability in which behavior occurred on which day. Nonobservable behaviors were reported more frequently than observable behaviors. Self-report via OBC was linear with in-field data collection methods (R^2 values ranged from 0.1 to 0.7; most values were within 0.3 to 0.4). The predictive value of the self-report total score was AUC (0.88) relative to the in-field study score. Separate test-retest reliability of the OBC was almost perfect (ICC = 0.88). **Conclusions:** The OBC is a reliable and valid way to predict behaviors in the natural environment and will be useful for further pain research. *J Oral Facial Pain Headache 2016;30:107–119. doi: 10.11607/ofph.1592*

Keywords: *behavior, pain, parafunction, self-report, TMD*

Waking-state overuse behaviors (eg, guarding) are well known in back pain,[1] whereas little is known about waking-state oral overuse, or parafunctional, behaviors; they are assumed to contribute to regional pain disorders.[2,3] Oral parafunctional behaviors are "different from those required for or associated with expected jaw functional demands such as mastication, swallowing, communication, or breathing."[2] Waking-state oral parafunction occurs in many ways, ranging from tooth-contact behaviors of varying intensities to many behaviors not involving tooth contact.[2] Both tooth-contact and noncontact behaviors can be subtle and not observable, and these behaviors generally occur outside conscious awareness, largely unnoticed by the individual and unobserved by others.[4,5] Consequently, substantial difficulty may occur in reliably assessing either the simple presence or extent of these behaviors.

The Oral Behaviors Checklist (OBC) initial item content was developed by the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) Validation Project to address the difficulties associated with assessing and measuring waking-state oral parafunctional behaviors potentially relevant to pain disorders.[6,7] The OBC is a comprehensive self-report instrument that includes a wide range of behaviors; some are self-evident and readily observable (eg, yawning) and some are nonobservable and difficult to detect (eg, guarding of the jaw). The scoring of the OBC uses a five-point ordinal scale ranging from "none" to "all of the time" and refers to the past month. In addition to the self-evident validity of the terms for readily observable behaviors, the terms for nonobservable behaviors are also valid with respect to a common understanding as measured by electromyography.[2,5,8] The OBC is consequently comprised of items with semantic validity that can be used for identifying and quantifying the frequency of oral parafunctional behaviors.[2,8] Given the validation of the OBC terminology as well as several studies demonstrating its potential utility,[9–11] the next step was to address measurement validity of the OBC.

A major limitation of the OBC, like any self-report instrument, is that the targeted behaviors or states generally occur beyond the moment in time or setting that the self-report questionnaire is completed; for example, mood questionnaires address this problem by using a time period that is typically 1 week to 1 month, and reliability is facilitated due to the typically continuous (albeit fluctuating) character of mood states over time. With regard to oral behaviors being discrete (ie, the behaviors may not be occurring at the time of completion of a self-administered questionnaire) and mostly outside conscious awareness, a questionnaire may easily result in underreporting. Thus, the present study investigated the following: Is retrospective self-report of behaviors that occur outside conscious awareness valid? One way to test this was to use a reliable and valid in-field electronic diary method, known as Ecological Momentary Assessment (EMA),[12] in order to acquire momentary self-reports of oral parafunctional behaviors, thereby enabling comparison of reporting of in-field performance with retrospective recall.[5] The specific aim of the study was to determine if retrospective self-report of oral parafunctional behaviors potentially relevant to pain conditions is valid by comparing oral parafunctional behaviors via a self-report instrument (OBC) to in-field reports of oral parafunction.

Materials and Methods

Overview

The RDC/TMD Validation Project was conducted at three study sites (Buffalo, New York; Minneapolis, Minnesota; and Seattle, Washington), with the goal of assessing the reliability and validity of the RDC/TMD.[6,7] The present project was conducted within the Validation Project as two separate substudies (summarized in Table 1). The first substudy—the field study—consisted of clinic and field measurements of parafunctional behaviors; it was conducted at the Buffalo study site. The second substudy—the test-retest reliability assessment—focused on the OBC and was conducted at the Buffalo and Seattle study sites. Ethical committee approvals were provided by the respective institutions.

This study followed the standards for developing psychological tests,[13] in particular, estimates of relevant reliabilities for total scores and subscores, representative subject sample selection, consideration of item context effects, and selection of cut scores on the basis of sound empirical data relating test performance to relevant criteria.

Subjects

As part of the standard protocol in the Validation Project, the initial draft version of the OBC was administered to each participant. During a 1-year period, subjects with high or low frequency of oral parafunctional behaviors, as reported by the OBC, were identified at the Buffalo study site and queried about field study participation. A convenience sample of 23 individuals who could respond to a handheld computer for a week was selected to participate in the field-measurement study. There were no other additional inclusion criteria; the selected subjects could be of any age or gender within the pool of TMD and non-TMD subjects recruited during that period of time into the primary Validation Project at the Buffalo study site. Of the 23 selected subjects, 22 completed the study (age in years: mean = 39.1, standard deviation [SD] = 12.2; female: 20). Informed consent was obtained, and subjects received the following compensation: $25.00 for completing the first OBC questionnaire, $75.00 for completing at least 90% of the electronic diary–prompted questionnaires, and $25.00 for completing the second OBC questionnaire.

The second substudy assessed the 2-week test-retest reliability of the full Validation Project Axis II instrument set; part of this substudy was previously reported.[6] Briefly, based on high or low distress scores (as measured by the General Health Questionnaire [GHQ-28]),[14] 74 subjects (age in years: mean = 39.7, SD = 11.0; female: 59) were identified for recruitment. These individuals completed the full Axis II packet at the time of Validation Study participation and again 2 weeks later. In the present project, only the test-retest of the OBC was of interest. This test-retest reliability assessment of the OBC served as a first control condition for the field study, in that these participants did not engage in the field measurement process, and this allowed the investigators to assess, by contrast, the potential contamination at the second self-report assessment in the field study as a result of electronic data collection. Contamination was construed to be the effects of self-observation as an intervening process.

Electronic Diary

The Experience Sampling Program Version 3.2 (ESP)[15] was used. It is designed to conduct surveys and runs on the Palm Pilot platform, a small handheld touch screen computer. Questions, order, and frequency are among the parameters that can be controlled by the investigator.

Table 1 Study Design

Substudy	Stage 1	Stage 2	Stage 3
Field study	21-OBC$_1$ 11-OBC$_{paper\ 1}$	11-OBC$_{ED}$ (target: 7 days)	21-OBC$_2$ 11-OBC$_{paper\ 2}$
Test-retest reliability assessment	21-OBC$_1$	No in-field data collection	21-OBC$_2$

Instruments

Two different modes of data collection were used: paper-administered OBC and electronic diary. The first mode relied on the self-report paper OBC questionnaire, which was created in two versions: the full 21-item version (termed 21-OBC) and an abbreviated 11-item version (termed 11-OBC$_{paper}$) in which the items matched those administered during the in-field electronic diary portion of the study (see below). The reason for both versions of the OBC was to evaluate whether form length and overall item context affected the responses during completion of the paper version. The remaining 10 items in the 21-OBC—ie, those not included in the 11-OBC$_{paper}$—served as a second control condition: Because those items were not assessed by the electronic diary during the intervening field data collection, the potential for reactivity was reduced when they were administered the second time.

The second mode relied on a set of 11 items administered by the ESP software and termed the electronic diary questions (11-OBC$_{ED}$). The initial software prompt was the request to "Respond to the following questions with respect to when you were prompted." These 11 items, identified in Fig 1, were exactly the same items asked on the 11-OBC$_{paper}$, and were aggregated into two clusters: six of the behaviors are subtle and can potentially evade both self-detection and detection by others, and are therefore termed nonobservable (eg, tightening the jaw, tensing the jaw), while five of the behaviors are readily observable by self and others (eg, leaning on the jaw, cradling a telephone). The measurement of the nonobservable behaviors was of primary interest for the present study, in that these behaviors are hypothesized to occur at varying levels of conscious awareness, thereby making their retrospective recall on self-administered questionnaires potentially more difficult. The observable behaviors were selected to serve as a comparison set of behaviors for assessing the participant's ability to perhaps more readily report obvious behaviors if they occurred. The distinction into nonobservable and observable behavioral clusters was made solely for measurement purposes of the present project.

Because five behaviors were considered incompatible for performing at the same time, they were nested within a single checklist query as question #1 on the electronic diary, with only one response permitted: clenching the teeth, touch or hold the teeth in contact, tighten or tense the muscles without touching the teeth, chewing gum, or biting objects. The first three of these behaviors were nonobservable, and the last two of these behaviors were observable; the single possible response to question #1 was classifiable as either a nonobservable behavior or an observable behavior. The remaining six behaviors (three nonobservable, three observable) were regarded as potentially co-occurring. Consequently, the maximum number of behaviors that could be reported at any recording epoch was four nonobservable behaviors and four observable behaviors. The maximum total number of behaviors was only seven, however, due to overlap in question #1, which included both nonobservable and observable behaviors but with only one response permitted.

To summarize the assessment of parafunction in the field study, each of the 22 participants was administered the 21-OBC, the 11-OBC$_{paper}$, and the 11-OBC$_{ED}$. Only the 21-OBC was used in the test-retest reliability assessment (Table 1).

Procedures

For each of the field study and test-retest study, procedures were explained and informed consent was obtained.

Field Study

The participants of the field study were trained in the use of the Palm Pilot ESP software by way of a short sample assessment item set, and the procedure for determining compensation was explained in relation to requested compliance.

At the first laboratory visit, each of the 21-OBC and 11-OBC$_{paper}$ instruments was administered, in counterbalanced order across subjects; the reference period for these two instruments was the preceding month. The ESP was programmed as follows: A reporting epoch was defined as each time a set of questions was delivered by the electronic diary in the form of the initial prompt. The timing of the reporting epochs was randomized within sequential 2-hour study periods that were determined based on the number of waking hours that the participant reported. The target was approximately eight reporting epochs per day for 7 days. Because state measures—ie, what the participant was doing at the exact time he or she was prompted by the electronic diary—were the goal of the electronic diary assessment, the items were closed and the participant was locked out from responding to that prompt if the participant did not respond to the electronic diary prompt within 90 seconds. The participant left the laboratory with the electronic diary.

At the end of the electronic diary period, the participant returned for a second laboratory visit and, again, completed each of the 21-OBC and 11-OBC_{paper} instruments by using the same counterbalancing sequence. The participant was then debriefed. The reference period for these instruments at the second visit was the previous week in order to intentionally overlap with the preceding electronic diary period and not to overlap with the reference period assessed by the OBC administered at study entry.

Test-Retest Reliability Assessment

Instruments were administered as follows: at the first visit, the 21-OBC was administered, and 2 weeks later at the second visit, the participant again completed the 21-OBC. No intervening electronic diary data were collected.

Data Reduction and Analysis

For the 11-OBC_{ED}, the 11 individual behaviors were nested within seven questions, and these seven questions were multiply nested within reporting epoch, day, and person. Initial data reduction created summary measures at each of the three levels of epoch, day, and person.

The reactivity of the field-study participants to the repeated administration of items during the electronic diary period was evaluated using intraclass correlation coefficients (ICC) of the paper OBC administered at each of the prefield and postfield assessments. Model (1,1), as labeled by Shrout and Fleiss,[16] was selected due to the within-subjects design.[17,18] There was 15.1% missingness in the paper OBC instrument data, which was random across subjects and due to study logistics of this project nested within a larger study; this was of substantial concern due to the removal of 6 of 22 subjects for comparing paper-instrument data versus field-study data. Consequently, values for missing data were imputed from a parallel instrument (21-OBC, 11-OBC_{paper}) if it was available at that same prefield or same postfield assessment. Imputation was performed as a simple replacement of missing data with the value from the other instrument, if available. Because the test-retest statistics for each of these instruments were to be compared, the imputation had the potential to bias the reliability of one instrument versus the other in the examination of contamination and context effects. However, inspection of the reliabilities associated with the nonimputed data disclosed no notable bias. The imputation did not affect the analyses underlying the primary purpose of the study. Internal consistency of behavioral reporting was assessed using Cronbach alpha for each item response to the paper OBC and the electronic diary (where the mean of all responses for a given item for the designated time period was computed as the analysis item).

Electronic diary values at the person level were normalized using two methods. The first method was a direct measure by normalizing based on the number of reporting epochs: a proportion of the number of epochs during which a given behavior was reported, divided by the total number of reporting epochs. The second method is less obvious and relies on latent trait theory by normalizing based on the total parafunctional behavior score as the best available estimate of the parafunctional characteristic or trait: a proportion of the number of epochs during which a given behavior was reported (ie, the sum score of that behavior) divided by the total of all behaviors reported by that person. The latter approach was applicable to both electronic diary data and to paper OBC data, thereby providing a uniform metric between these two different modes of assessing behavioral frequency.

The primary goal of the study was to assess the validity of parafunctional behaviors as reported on a self-administered instrument referring to a prior time period of, for example, 30 days. The reference standard for assessing that validity was the prospective data collected via the electronic diary. Two approaches to validity were used. The first was based on comparison of the mode of self-report (ie, electronic diary versus self-administered OBC) for each item through the use of graphic depictions, regression coefficients, and R^2 values, as well as using normalized data as previously described. The second approach was to compute area under the receiver-operating characteristic curve (AUC), with the paper instrument serving as the test instrument and the electronic diary data collection serving as the reference standard.

Stata was used for all data analysis. The rejection value for the null hypothesis was set at $P = .05$. Interpretation of ICC was based on a standard guideline, with 0.61 to 0.80 ("substantial") and 0.81 to 1.00 ("almost perfect" to "perfect")[19] as the ranges of interest.

Results

Descriptive results related to the field study are listed in Table 2. Overall, the number of data-collection epochs and the number of reported behaviors were considered sufficient for the purposes of the study. The number of missed epochs represented 15.1% of the available epochs, for which the most commonly stated reason was the participant was driving a car. The mean number of partially completed epochs was 1.0: If the subject responded to the epoch prompt, it was unlikely that the subject would not respond to all items. Partially completed epochs were not included in the scoring.

Reactivity to electronic diary data collection was assessed using the reliability (ICC) of the two self-report instruments (21-OBC; 11-OBC_{paper}), comparing the pre–electronic diary and post–electronic diary administrations of each instrument; results are listed in Table 3. These results are pertinent to the internal validity of the study. First, when comparing the individual item reliabilities on the 11-OBC_{paper} versus those on the 21-OBC, subjects appear to report the same behaviors more variably on the shorter 11-OBC_{paper} instrument than on the 21-OBC. On the 11-OBC_{paper}, 3 of 11 items had higher ICC values (proportion test; P = .033, 2-tailed) compared with the ICC of the corresponding items on the 21-OBC, where the expected value is 5.5 (ie, 50% of 11). This indicates that the self-report instrument length influences the respondent, and that items, when looked at individually, are less reliable on the shorter instrument. However, when considering the reliability of the full scores of the 11-OBC_{paper} and the 21-OBC, the reliability coefficients (ICC = 0.65) are the same for both instruments. Additionally, the reliabilities of the subscale scores for each of the nonobservable and observable behaviors are also equally notable for both instruments. The reliability of the 21-OBC full-scale score in the test-retest reliability assessment was 0.88 (95% confidence limits [CL]: 0.82, 0.92).

Second, reactivity could also be evaluated using the 10 items assessed on the 21-OBC but not during the electronic diary field-collection. Had there been reactivity to the 11-OBC_{ED} items, it would have been expected that the reliability of items not assessed on 11-OBC_{ED} would be significantly higher than that for the 11-OBC_{ED}. The ICC for the subscale of 11-OBC_{ED} items was 0.65 (95% CL: 0.29, 0.85) and it was not substantially different from the ICC for the subscale of the items not assessed on 11-OBC_{ED}, 0.73 (95% CL: 0.42, 0.89). Note that the 95% CL for these ICC values overlap (albeit minimally) with the corresponding 95% CL from the test-retest reliability assessment. It can be concluded that, on the whole, no reactivity across the electronic diary data collection period was observed, supporting the internal validity of the in-field data collection.

Cronbach alpha values computed from electronic diary data collected on each day ranged from 0.12 to 0.63 for raw behavior counts and 0.25 to 0.37 for normalized behavior counts. Cronbach alpha for electronic diary data collected across the entire reporting period was higher, 0.56 for raw item counts and 0.46 for item counts adjusted for the number of reporting epochs. By comparison, for the OBC instruments administered at the first laboratory visit, Cronbach alpha was 0.72 for the 11-OBC_{paper} instrument and was 0.72 for the electronic diary items in the 21-OBC instrument.

Table 2 Descriptive Statistics of In-Field Performance (Mean per Subject, n = 22)

Variable	Mean	SD	Min	Max
Study interval (d)	6.4	1.3	2.9	9.1
Total epochs	41.2	12.0	15	63
No. of missed epochs	7.3	5.3	1	24
No. of unfinished epochs	1	2.0	0	6
Total no. of questions	288.6	83.8	105	441
Total no. reported behaviors	65.5	34.7	8	163

In-field measurement was based on use of the 11-OBC_{ED}.

Table 3 Reliability (ICC) of Different Paper Instruments, Comparing Pre–Electronic Diary to Post–Electronic Diary Administrations

Item	Self-report version	
	ICC: 11-OBC_{paper}	ICC: 21-OBC
Electronic diary items		
Clench teeth together (N)	0.35	0.63
Press, touch, or hold teeth together (N)	0.59	0.65
Hold, tighten, or tense muscles (N)	0.22	0.36
Chew gum (O)	0.85	0.93
Hold or bite objects (O)	0.26	0.83
Hold or jut jaw forward or to the side (N)	0.78	0.57
Bite, chew, or play with soft tissue (N)	0.63	0.36
Press tongue forcibly against teeth (N)	0.74	0.88
Yawn (O)	-0.11	0.33
Lean with hand on jaw (O)	0.62	0.75
Hold telephone between head and shoulders (O)	0.82	0.33
Aggregate subscores		
Any of 6 nonobservable behaviors	0.64	0.67
Any of 5 observable behaviors	0.62	0.79
Adjusted sum score of EMA items	0.67	0.66
Items NOT included in the electronic diary		
Grind or clench teeth while asleep	–	0.90
Sleep in position that strains jaw	–	0.45
Grind teeth during waking hours	–	0.81
Place tongue between teeth	–	0.68
Hold jaw in rigid or tense position	–	0.40
Play musical instrument using mouth or jaw	–	Not computed
Chew food on one side	–	0.84
Eat between meals	–	0.72
Sustained talking	–	0.76
Singing	–	0.58
Score		
Adjusted sum score of non-EMA items	–	0.74

n = 22, with about 10% of data imputed due to missingness related to study logistics.

ICC = intraclass correlation coefficient; N = nonobservable, not readily observable behavior; O = observable, readily observable behavior.

Table 4 Reliability (ICC) of Behaviors Reported on 11-OBC$_{ED}$, Comparing Different Days

	Days					
	1 versus 2		3 versus 4		3 versus 6	
Item	ICC-Raw	ICC-Adj	ICC-Raw	ICC-Adj	ICC-Raw	ICC-Adj
Clench	0.6	0.4	0.5	0.4	0.1	0.2
Touch or hold	–0.1	0.0	0.5	0.5	0.5	0.6
Tight or tense	0.4	0.5	0.8	0.8	0.7	0.6
Chew gum	0.8	0.7	0.9	0.9	–0.1	–0.1
Bite objects	–0.1	–0.1	0.4	0.4	0.4	0.6
Jaw forward	0.6	0.8	0.7	0.8	0.7	0.7
Soft tissue	0.4	0.2	0.4	0.4	0.5	0.5
Press tongue	0.6	0.6	0.8	0.8	0.8	0.8
Yawn	0.3	0.5	0.7	0.7	0	0
Lean jaw	0.5	0.4	0	–0.1	–0.2	–0.1
Cradle telephone	–0.2	–0.1	0.4	0.3	0.6	0.7
Combinations						
Sum of 6 nonobservable behaviors	0.5	0.7	0.7	0.9	0.8	0.8
Sum of 5 observable behaviors	0.2	0.2	0.4	0.2	0.3	0.5
No. of behaviors[a]	0.4	–	0.8	–	0.7	–

[a]ICC-Adj could not be computed for since the numerator and denominator would necessarily be the same.
ICC = intraclass correlation coefficient; Raw = total number of behaviors reported each day; Adj = total number of behaviors reported each day, adjusted for number of sampling epochs.

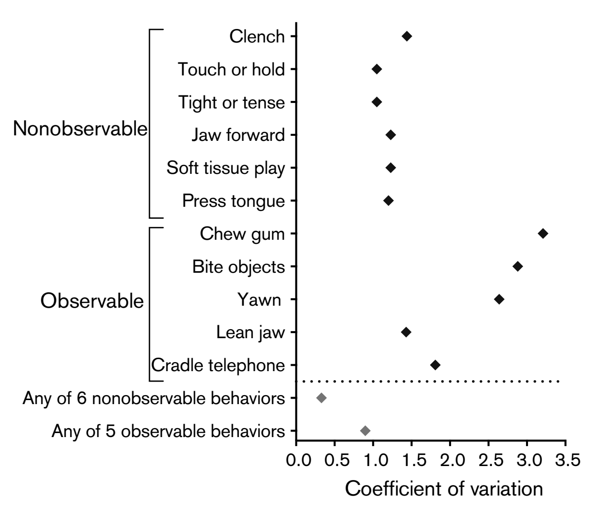

Fig 1 Variability across days for each behavior or groups of behaviors. Variability is expressed by the coefficient of variation computed from the mean and SD of the behavioral count of each targeted behavior, for each of the first 6 days. The aggregate measure for each of the nonobservable behaviors and of the observable behaviors was computed based on any of the component behaviors being reported during an in-field measurement epoch.

Figure 1 shows the variability, presented as coefficient of variation (CV; standard deviation divided by the mean), of each behavior or groups of behaviors across 6 days. The CV controls for differences in mean values and was computed as follows: The sum of each reported behavior by each subject over each day was computed, a mean value for each day was obtained across participants, and a grand mean and SD based on 6 days was computed. These results indicate that the occurrence of each behavior differed considerably each day.

Table 4 provides detail underlying the data illustrated in Fig 1, describing the reliability across days comparing different intervals and scoring methods and gives further detail about each sampling period. These days for comparison were selected because they represent either contiguous days at the beginning or middle, or a broader interval from the middle to the end of the in-field period of data collection, and they could highlight findings about data collection methods. Scoring methods (ie, raw count versus count adjusted for number of sampled epochs) did not affect the reliability, which ranged from low to high, depending on the behavior but not the time interval. The ICCs of the sum of all nonobservable behaviors (range, 0.6 to 0.9) demonstrated a high level of consistency of these behaviors, whereas the relatively low ICCs of the sum of the observable behaviors (range, 0.2 to 0.4) most likely reflected their relatively low occurrence and thereby lower variance. Overall, the results indicate a more consistent pattern of occurrence for the nonobservable behaviors and a more sporadic pattern for the observable behaviors across the identified time periods and across subjects.

Figure 2 shows the rating scale characteristics of the OBC as a descriptive box-and-whisker plot of the electronic diary reports. The 21-OBC asked how often an activity occurred in the last month and indicated to choose the higher option if the frequency varied. All behaviors on the 21-OBC with a given rating scale response were aggregated, and the mean value of the electronic diary reports for those same behaviors was computed within participant. If a participant did not use a given rating scale value on the 21-OBC, that individual would not be represented in the corresponding box-and-whisker; consequently, the number of participants contributing to each box-and-whisker varied, and there was no statistic for this unbalanced data. Figure 2 graphically addresses whether the 21-OBC rating scale worked as intended. The electronic diary–reported behavior medians marched upward across the paper rating scale levels, indicating that as the rating scale endorsement increased so did the median number of behaviors reported in the field. The whiskers indicate that at least one person overestimated their behaviors on the 21-OBC (whiskers at the floor for each of the rating scale values) and that at least one person underestimated their behaviors on the OBC (whiskers at or near the ceiling for each of the rating scale values).

Tables 5 and 6 describe the proportion of behaviors of each subject. Because this was a convenience sample selected for exhibiting a range of behaviors, the results should be interpreted only relatively across behaviors. In Table 5, the proportion of behaviors per reporting epoch (ie, number of behaviors per epoch) ranged from a mean of 0.02 (SD = 0.04) for chewing gum to a mean of 0.26 (SD = 0.24) for tighten or tense the muscles; any nonobservable behaviors occurred at a mean of 0.76 (SD = 0.21) and any observable behavior occurred at a mean of 0.24 (SD = 0.14). In Table 6, the proportion of behaviors per total behavioral count ranged from a mean of 1.3% (SD = 2.9) for chew gum to 15.3% (SD = 14.7) for tighten or tense the muscles; any nonobservable behaviors occurred at a mean of 52.5% (SD = 13.7) and any observable behavior occurred at a mean of 16.5% (SD = 12.2). Every behavior was completely absent in at least one person (ie, min = 0), while some behaviors were reported far more frequently than others. When any of six nonobservable behaviors are considered, a mean of 76% of sampled epochs exhibit one of these behaviors, and the maximum value of 100% indicates that at least one participant was always exhibiting one of these behaviors when prompted by the electronic diary. In contrast, when any of the five observable behaviors were considered, the mean, minimum, and maximum were lower, indicating that these behaviors are less likely to be reported at the time of diary prompting.

Figure 3 shows the values for each participant's behavior, as presented in Tables 5 and 6. The data, as indicated by regression statistics beneath each plot, clearly conform to a linear model; the plots and R^2 values ranging from 0.45 to 0.96 are consistent with an acceptable to good linear fit. These results indicate that a relative proportion of electronic diary–reported behavior, as defined by the number of

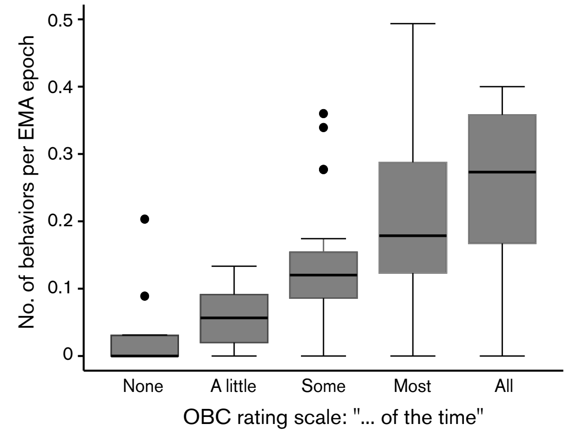

Fig 2 Rating scale characteristics. For each person, the number of electronic diary–based reports was averaged for a given rating scale response (none, a little, some, etc) on the paper OBC, regardless of specific behaviors. Each bar then represents one value from each of a maximum of 22 persons. Some bars have fewer than 22 persons if a person did not use that particular rating for any behaviors on the 11-OBC_{paper}. Standard whiskers and outliers are shown.

occurrences of that behavior adjusted for the total number of behaviors, is a sufficient proxy for the more intuitively correct sampling proportion of the behavior.

Figure 4 illuminates similarities between the initial self-report via paper questionnaire and self-report via electronic diary data collection on the 11-OBC_{ED}. Using the same metric, a proportion of the total score, the electronic diary data were compared with the corresponding items on the 21-OBC, as the overall goal was to validate the 21-OBC. Each scatter plot includes all of the participants. A linear fit existed for the general trend of each scatter plot, and 9 of 11 behaviors exhibited significant linearity in the comparison of the two methods of assessing the behavioral frequency. The more/less the participant reported the behavior on paper, the more/less the behavior was reported in the field. R^2 values ranged from low (0.1) to high (0.7), but most were in the range of 0.3 to 0.4, which would be consistent with the daily variability (Fig 1) and incorporates the outliers, which can influence the R^2 values. Plots demonstrating good, if not better, linear fit between proportion of behavior per electronic diary epochs (as shown on the x-axis in Fig 3) and proportion of each behavior to the total behavioral score from the 21-OBC (as shown on the x-axis in Fig 4) were generated (but are not shown) and presented the same finding.

Table 7 builds on the relationships shown in Fig 4; the receiver operating characteristics associated with the self-report OBC questionnaire as a predictor

Table 5 Sampling Proportion of Behaviors per Subject

Variable	Mean[a]	SD	Min	Max
Clench	0.14	0.15	0	0.50
Touch, press, hold	0.21	0.16	0	0.56
Tighten, tense	0.26	0.24	0	0.86
Chew gum	0.02	0.04	0	0.18
Bite objects	0.04	0.06	0	0.25
Move jaw forward or side	0.24	0.25	0	0.75
Bite soft tissue	0.22	0.20	0	0.75
Press tongue	0.25	0.25	0	0.71
Yawn	0.03	0.05	0	0.18
Lean on jaw	0.10	0.07	0	0.21
Cradle telephone	0.07	0.09	0	0.40
Composites				
Any of 6 nonobservable behaviors	0.76	0.21	0.35	1.00
Any of 5 observable behaviors	0.24	0.14	0	0.51

[a]Mean value is the number of instances of the behavior per reporting epoch. A mean of 0.14, using clench as an example, can also be interpreted as equivalent to the report of 14 clenches across 100 reporting epochs, indicating that, on average, each subject was exhibiting clenching 14% of the time, generalizing from the random EMA sampling, with a minimum of 0% by at least one subject and 50% of the time by at least one subject.

Table 6 Relative Proportion of Behaviors During Electronic Diary Collection

Variable	Mean[a]	SD	Min	Max
Clench	10.6	12.4	0	50.0
Touch, press, hold	15.3	14.9	0	64.7
Tighten, tense	15.3	14.7	0	64.2
Chew gum	1.3	2.9	0	12.9
Bite objects	2.5	3.5	0	12.3
Move jaw forward or side	13.1	11.9	0	38.7
Bite soft tissue	14.6	13.7	0	52.6
Press tongue	12.8	11.8	0	37.6
Yawn	2.4	4.2	0	14.7
Lean on jaw	6.5	4.40	0	14.7
Cradle telephone	5.6	8.6	0	41.0
Composites				
Any of 6 nonobservable behaviors	52.5	13.7	31.9	75.0
Any of 5 observable behaviors	16.5	12.2	0	52.5

[a]Mean values are percent of the total behaviors reported during the electronic diary period.

of the electronic diary data are presented. The electronic diary data were regarded as the reference standard since they were collected via the EMA paradigm, which is generally regarded as substantially better than recalled behavior.[20] In other words, how good is retrospective self-report at predicting the actual behaviors performed in the field? The final validity coefficient was computed as AUC by using sensitivity-specificity analysis. The electronic diary score was dichotomized using a cutoff derived from other research suggesting that there is a threshold effect in the amount of oral parafunction that matters clinically. From the 21-OBC instrument, a score of approximately 26 or more predicts first-onset TMD and is strongly associated with chronic TMD.[9,11] The score of 26, relative to the maximum score of 84 for the 21-OBC, was converted to a proportion that was applied to the electronic diary total score, thereby creating a dichotomous outcome. The rationale for this approach was that there is, at present, no published data using a multi-item approach such as the OBC identifying a minimum level of these behaviors that is meaningful with regard to putative disease effects. Therefore, the approach here is an initial estimate. The sensitivity, specificity, AUC, and likelihood ratio (positive and negative) values were computed for each type of OBC score obtained at the first study visit, referencing the prior 30 days. The cutpoint of the selected predictor instrument was based on maximizing each of the above statistics, starting with AUC and then using the other statistics for refining that choice. The final test evaluating the items of the initially offered 21-OBC that were not tested in the field study showed an expected drop in the AUC, but the value at AUC = 0.70 was still large, again consistent with parafunction as a latent construct. That is, parafunction is a latent construct that can be identified by items highly related to the behaviors assessed in the field or it can also be identified by the 10 OBC items not related to behaviors assessed in the field study.

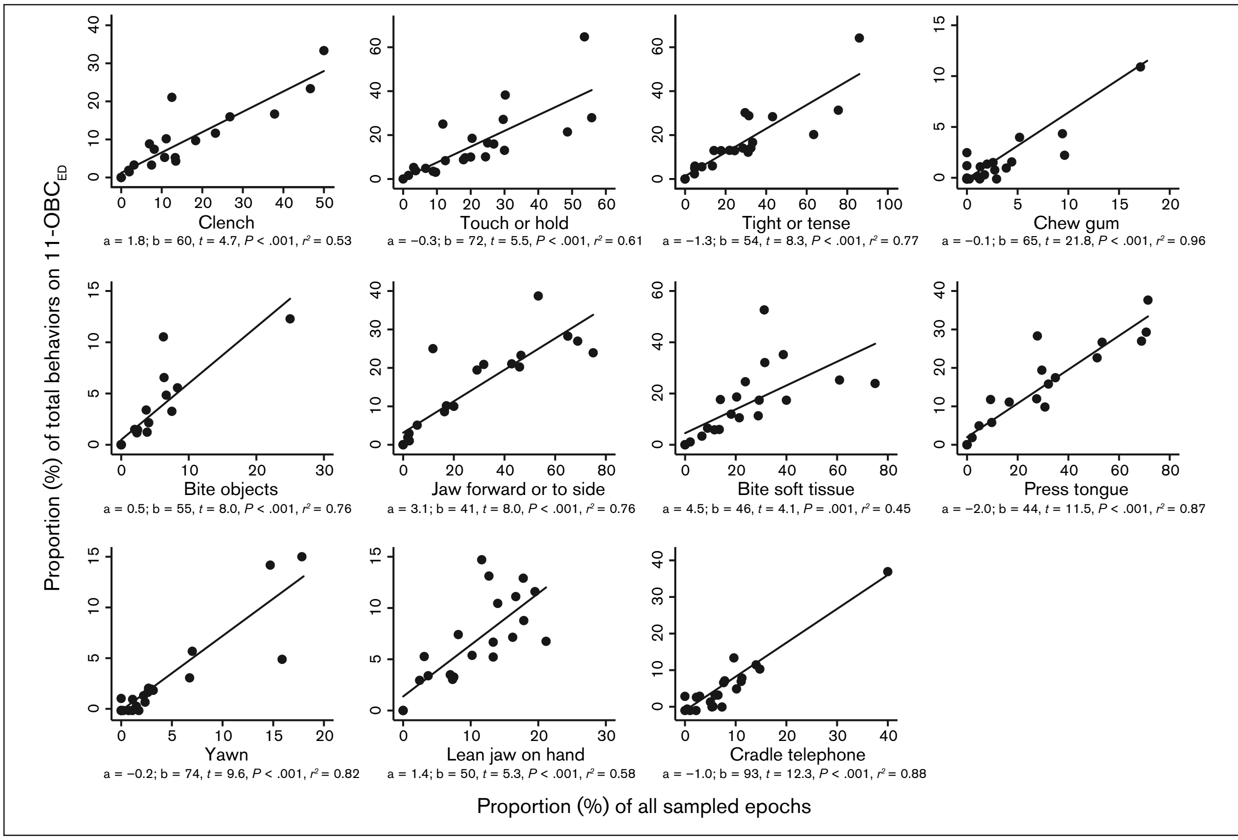

Fig 3 Percentage of total in-field behaviors versus number of behaviors per epoch. In-field behaviors were measured using 11-OBC$_{ED}$.

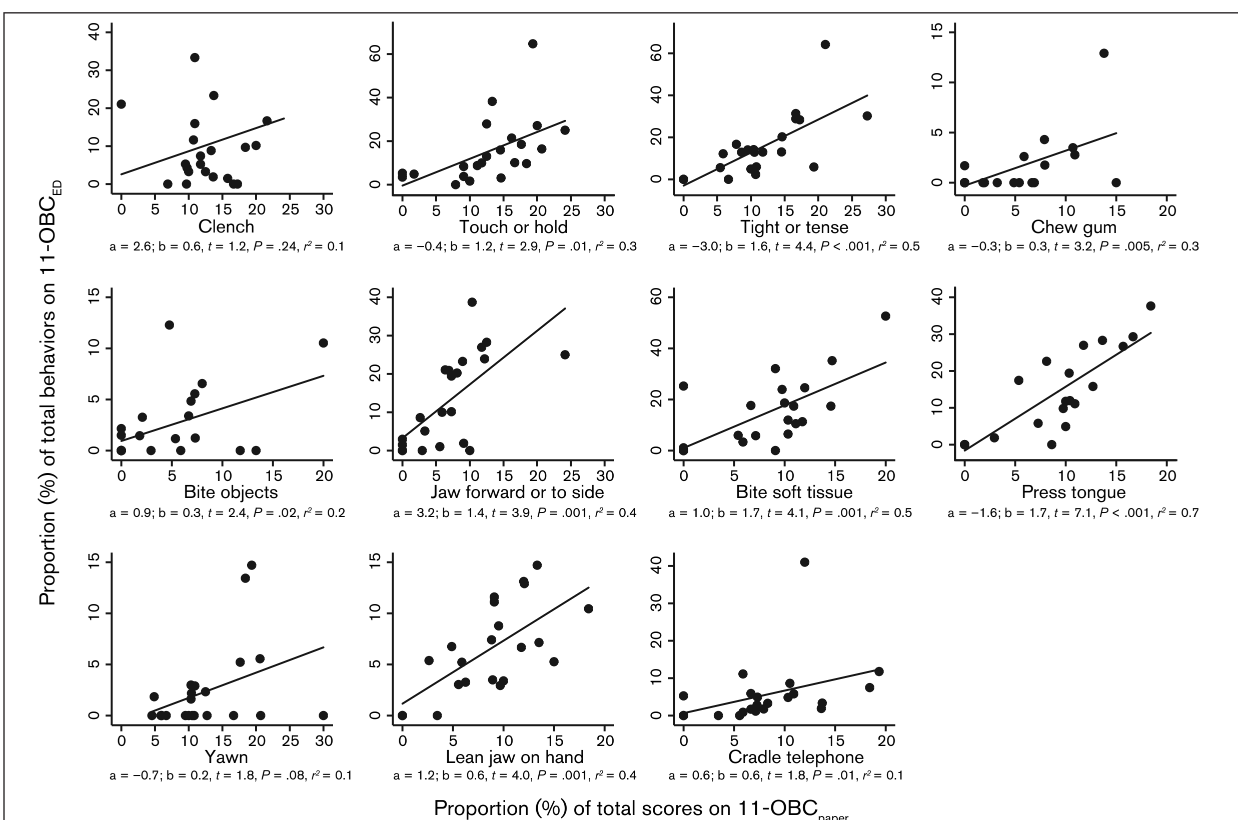

Fig 4 Percentage of total in-field behaviors versus percentage of corresponding behaviors via paper self-report. In-field behaviors were measured using 11-OBC$_{ED}$. Paper self-report was measured using 11-OBC$_{paper}$.

Table 7 Prediction Results for Different Self-Report Measures or Different Scoring Methods

Description of baseline measure	n	AUC	SE	Cutpoint	Sens	Spec
11-OBC$_{paper}$, adjusted for missing data	22	0.88	0.09	1.6	86	88
11-OBC$_{paper}$	16	0.86	0.09	1.3	67	86
Items not tested on 11-OBC$_{paper}$	16	0.70	0.15	1.6	67	100
21-OBC (full scale score)	16	0.80	0.12	1.5	67	100

Outcome (reference standard) variable is behaviors recorded during the electronic diary period, and the rate of reporting across all sampled behaviors was dichotomized using a cutpoint of 1.16, as derived from OPPERA data.[9] The optimal cutpoint for each prediction variable was selected based on maximum percent classified correctly and either maximal LR+ or minimum LR–.
AUC = area under the receiver operating characteristic curve; SE = standard error of AUC; Sens = sensitivity; Spec = specificity.

Discussion

The present results indicate that actual oral parafunctional behaviors do map to the behaviors individuals report (eg, via a self-administered instrument such as the OBC). For each of the behaviors, the subject's responses to the paper and in-field assessments had direct, linear relationships. The occurrence of individual behaviors during the in-field measurement varied across time and persons. When like behaviors are aggregated together (here, nonobservable or observable clusters), the variability becomes more stable, consistent with the Spearman-Brown prophecy formula.[21,22] Hence, the construction of potential subscales within a latent variable, parafunction, will be important for further development of scoring a paper self-report instrument with clinical utility.

These data also indicate both substantial under- and overestimation, based on the OBC, of the actual extent of behavior occurrence. Both clinician and patient will benefit by having a more precise understanding of the actual extent of the putative parafunctional behaviors. An instrument such as the OBC might be an excellent screening tool, but it must be accompanied in the clinical setting by further evaluation (eg, history, in-field monitoring over a sufficient number of days, follow-up review)[23] in order to base treatment on the actual occurrence of any particular behavior. At the latent variable level, the full-score reliability of ICC = 0.88 in the test-retest reliability assessment of the 21-OBC corresponds to ICC = 0.86 as reported elsewhere,[24] denoting that the longer time base of the paper instrument may well be a strong advantage, as it taps into memory of actual behavior. The fact that the nonobservable behaviors, as a cluster, occurred at three times the rate of the observable behaviors, as a cluster, is also important to note in terms of potential clinical application of these findings.

A major strength of the present study was the use of the EMA paradigm. Electronic data collection in the field is time stamped, can only be answered at specific times, and therefore cannot be answered at a later time with potentially unreliable information.[25] Consequently, the reported behaviors during the electronic diary phase represent good estimates of the occurrences of these behaviors, assuming sufficient sampling and response rates. Some other studies have used the EMA paradigm for assessing a few oral parafunctional behaviors,[26–29] with the work of Glaros and colleagues using an EMA paradigm as perhaps the only systematic studies of these oral behaviors.[28,30]

Because collection of self-report data by using the EMA paradigm leads to immediate recall of the respondent's experience in the moment of inquiry, while going about daily tasks, bias and inaccuracy are minimized.[12] A different form of bias associated with the EMA paradigm is the reactive measure, which can potentially contaminate in-field measurement.[25,30,31] If subjects had been reactive to the in-field data collection with a resultant increase or decrease in the behavior, the self-rated frequency of the behaviors at the post–electronic diary assessment would have decreased, thereby resulting in a lower ICC. The reliability coefficients of the pre–electronic diary and post–electronic diary administrations of the paper instruments, at 0.65, fell into the "substantial" range for the subscale scores, which is often the best a given instrument might achieve for many types of measurement. In addition, the data indicate that performance across days is consistent for each behavior through the electronic diary data-collection period, again suggesting reactivity and behavioral change due to the diary reporting did not occur. Similar findings with regard to lack of reactivity have been reported elsewhere.[32–35]

The reliability results suggest that the items on either instrument may be sufficient indicators of a latent variable, oral parafunction. The Cronbach alpha from the electronic diary data are weak, probably because the behavior is discrete and not sampled for a long enough time, whereas the results from the 21-OBC point to some minimal level of acceptability of these behaviors serving as a basis for measuring a latent variable. The relatively high CV and low reliability across days underscores the complexity associated with measuring these behaviors, either in-field or on paper as a retrospective report. A high CV may be indicative of more variability in circumstances that could control the behavior. For instance, whether the participant is chewing gum at the time of electronic diary reporting might be dependent on access or a situation in which it is appropriate to chew gum. Additionally, some behaviors may have been missed if sampled on the wrong day, and more frequent sampling (at the possible expense of creating a reactive measurement) or sampling for more days would be beneficial.

The absence of reactivity also provides evidence for the challenge that many people find in trying to change behavior: If self-monitoring via computer prompt multiple times per day, over many days, does not change the behavior that most people intuitively know probably is not good, then promoting therapeutic change will likely require additional methods beyond simple self-observation.

There is considerable evidence that self-report instruments, especially when administered in the context of space for personal reflection, allow individuals to reliably and accurately report previously unexamined personal states.[31,36,37] Because the OBC is a self-report instrument that assesses recall of behavior frequency, a major source of error in the reported frequency resides in the unconscious aspects of the behavior and associated memory retrieval. In other words, how can an individual, unaware of whether the behavior occurs or not, accurately report the frequency of the behavior? The equivalent reliabilities for each nonobservable or observable cluster on the OBC are notable: While the nonobservable behaviors are typically produced in a state of low conscious awareness, individuals are able—if using an instrument such as the OBC—to recall those behaviors as reliably as they do the more obvious behaviors.

Electronic diary data collection allows investigation into the temporal relationships of complex patterns of variables, eg, between recalled pain and real-time measured pain. Stone et al found, as did previous researchers, that momentary measures of pain intensity were lower than recall measures of that same pain.[38] In comparison to daily-recorded data, pain-free episodes were often discounted, while pain-related episodes were often embellished during clinical interviews when subjects recalled information about their pain.[38] The reported absence of a potentially causal relationship between OBC scores and clinical pain, denoted by a low correlation,[24] may be due entirely to the problem highlighted by Stone et al: If oral parafunctional behaviors do contribute in a state-like manner to pain intensity, then measures of momentary pain intensity, rather than recalled pain intensity, will likely need to be linked to the momentary measures of parafunctional behaviors. And, indeed, Glaros et al already have provided strong evidence through an experience sampling method of exactly that: Momentary processes related to stress, pain, and parafunction appear to be causally linked in time.[28] In addition, an individual case report and analysis presented plausible causal relations across time between these variables.[39]

Examining the present data yields several other notable findings. Overall, item reliability was variable and oftentimes quite low, indicating that no two days are identical; multiple sampling days are essential. The results also showed that each subject was different; for example, some subjects always performed specific oral parafunctional behaviors and never performed other behaviors. In general, trying to evaluate the true rate of a single behavior within one day is extremely difficult, if not impossible, because people exhibit high variability in what they do each day (at least, as sampled per this study design). Extrapolating from the findings of Glaros et al,[28] variability in personal stress reactivity across days would likely affect the production of certain behaviors, thereby decreasing the measurement reliability of the behaviors within the electronic diary measurement framework.

While a number of studies have attempted to estimate the prevalence of oral parafunctional behaviors, little is actually known. Helkimo assessed the prevalence of different parafunctional behaviors by using self-report.[40] Each of the different behaviors was reported in five age groups, ranging from 15 to 65 years. In the 35- to 44-year-old group, for example, 7% reported biting on objects, 9% tongue lip or cheek biting, 14% pressing of tongue, and 42% clenching and grinding of teeth. Across all behaviors and ages, the reported range in occurrence was 0% to 42%.

In a Medline search done by the authors using the terms “parafunction” and “prevalence,” only 27 citations were found, and abstracts yielded 19 papers relevant to the current project. For the assessment of the prevalence of parafunction, these 19 studies relied upon either dentist observation of wear facets of the teeth, which is not a reliable assessment,[41] or simple questions such as apparently used by Helkimo.[40] None used any sampling methods. Determination of parafunction by only a few items appears to result in false-negative reports, and consequently even the estimates provided by Helkimo[40] probably represent underreporting. Even less is known about the rate, or frequency, at which the behaviors occur within an individual. Glaros and Williams noted that individuals with head pain report, via questionnaire, that they clench and contact the teeth 41% and 65% of the time, respectively, whereas individuals with any chronic pain clench and contact at 31% and 48%, respectively.[42] The between-group (those with head pain versus any chronic pain) differences in the self-estimated rates of clench (41% versus 31%) and contact (65% versus 48%) may point to the critical level where the behaviors are problematic in terms of pain.

Because of the convenience sample used in the present study, prevalence of the different behaviors could not be determined, but relative frequency indicated that clenching is probably less frequent than even more subtle behaviors such as tensing or tightening the muscles. These findings highlight the need for sensitive measures to identify the presence of behaviors such as tensing or tightening in order to determine whether they are clinically important. This convenience sample does provide information about one aspect of prevalence: All participants, even those selected for study inclusion because the initial assessment indicated low probability of reporting behaviors to the electronic diary, reported some type of parafunctional behavior at some point. Similarly, everyone seems to exhibit sleep bruxism, if only a little.[43] If everyone reports some level of waking oral parafunctional behaviors, then the simple report of the occurrence (yes versus no) of parafunction may have no clinical meaningfulness. Rather, the clinical meaningfulness is the threshold at which these behaviors contribute to disease, for example, TMD.

In this study, the AUC (0.88) for the EMA-relevant self-administered items to predict in-field behaviors is important, both for its magnitude as well as for the threshold used to create the binary outcome variable, the in-field behaviors. That threshold was based on previously published regression models in which OBC scores above that threshold were strongly associated with both chronic TMD and first onset-TMD.[9,11] This threshold requires further evaluation, as its role here was solely to create a binary reference standard against which to test the paper OBC as a first step. Glaros and Williams reported that tooth contact, compared with clenching, is a better predictor of pain[42] and, as such, point to different behaviors as potentially having different importance, or weight, with regard to their contribution to any consequences, such as pain or first-onset TMD. The binary reference standard in the present study was based on a dimensional scale derived from unit weighting of all reported behaviors, and weighting was used in the scoring of the paper OBC. Two directions for further research are *(1)* to examine the assumption that equal weighting best represents the biologic correlates associated with any consequences of parafunctional behaviors and *(2)* to determine how best to clinically estimate the actual extent of behavior.

The present study had several limitations. First, the convenience sample prevented generalizations about proportions associated with the various behaviors; this sample was selected solely for its value in assessing measurement validity. Second, missing data across paper instruments, which did not affect observed correlations, may nevertheless have affected the reproducibility of the present results. Third, a larger sample size would increase the power of the study and likely improve the fit and precision of estimates of the linear models depicted in the plots. It is unlikely that a larger sample size would cause a substantial increase in the primary outcome statistic, the AUC, in that that value is already respectable. And fourth, a longer electronic diary period, such as a minimum of 2 weeks, would illuminate additional information about behaviors that occur less frequently and would improve the precision of the estimates of occurrence of those behaviors.

In conclusion, the OBC is a reliable and valid way to predict parafunctional behaviors in the natural environment that have relevance to pain conditions. This prediction, however, will be improved with the addition of in-field monitoring or patient inquiry, as determined by circumstances.

Acknowledgments

The authors declare no conflict of interest related to the present article. This study was supported by the University at Buffalo, School of Dental Medicine, Dental Student Research Fellowship for Summer Research, and the Participating Fund for Dental Education and the Centennial Fund. Data collection was funded by NIH/NIDCR U01-DE013331. The RDC/TMD Validation Project clinical investigator team was composed of Eric Schiffman, PI, and Gary Anderson (University of Minnesota); Ed Truelove and Earl Sommers (University of Washington); and Richard Ohrbach and Yoly Gonzalez (University at Buffalo), all of whom contributed equally to the development of the initial items comprising the OBC.

References

1. Prkachin KM, Schultz IZ, Hughes E. Pain behavior and the development of pain-related disability: The importance of guarding. Clin J Pain 2007;23:270–277.
2. Ohrbach R, Markiewicz MR, McCall WD Jr. Waking-state oral parafunctional behaviors: Specificity and validity as assessed by electromyography. Eur J Oral Sci 2008;116:438–444.
3. Ohrbach R, Blasberg B, Greenberg MS. Temporomandibular disorders. In: Glick M (ed). Burket's Oral Medicine, ed 12. Shelton, CT: PMPH-USA, 2015:263–308.
4. Glaros AG. EMG Biofeedback as an Experimental Tool for Studying Pain. Biofeedback 2007;50–53.
5. Ohrbach R. Assessment and further development of RDC/TMD Axis II biobehavioural instruments: A research programme progress report. J Oral Rehabil 2010;37:784–798.
6. Ohrbach R, Turner JA, Sherman JJ, et al. The Research Diagnostic Criteria for Temporomandibular Disorders. IV: Evaluation of psychometric properties of the Axis II measures. J Orofac Pain 2010;24:48–62.

7. Schiffman EL, Truelove EL, Ohrbach R, et al. The Research Diagnostic Criteria for Temporomandibular Disorders. I: Overview and methodology for assessment of validity. J Orofac Pain 2010;24:7–24.
8. Markiewicz MR, Ohrbach R, McCall WD Jr. Oral behaviors checklist: Reliability of performance in targeted waking-state behaviors. J Orofac Pain 2006;20:306–316.
9. Ohrbach R, Fillingim RB, Mulkey F, et al. Clinical findings and pain symptoms as potential risk factors for chronic TMD: Descriptive data and empirically identified domains from the OPPERA case-control study. J Pain 2011;12:T27–T45.
10. Michelotti A, Cioffi I, Festa P, Scala G, Farella M. Oral parafunctions as risk factors for diagnostic TMD subgroups. J Oral Rehabil 2010;37:157–162.
11. Ohrbach R, Bair E, Fillingim RB, et al. Clinical orofacial characteristics associated with risk of first-onset TMD: The OPPERA prospective cohort study. J Pain 2013;14:T33–T50.
12. Shiffman S. Real-time self-report of momentary states in the natural environment: Computerized ecological momentary assessment. In: Stone AA, Turkkan JS, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS (eds). The Science of Self-Report: Implications for Research and Practice. Mahwah, NJ: Lawrence Erlbaum Associates, 2000:277–296.
13. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: AERA, 2014.
14. Goldberg D, Williams P. A User's Guide to the General Health Questionnaire. Windsor, UK: NFER-Nelson, 1988.
15. Barrett LF, Barrett DJ. ESP: The Experience Sampling Program. http://www.experience-sampling.org/. Accessed 1 June, 2015.
16. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–428.
17. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19:231–240.
18. Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: Assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. Clin J Pain 2005;21:73–82.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
20. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1–32.
21. Ghiselli EE, Campbell JP, Zedeck S. Measurement Theory for the Behavioral Sciences. New York: WH Freeman, 1981.
22. Anastasi A. Psychological Testing. New York: Macmillan, 1988.
23. Dworkin SF, Ohrbach R. Biobehavioral assessment and treatment of temporomandibular disorders. In: Bays RA, Quinn PD (eds). Temporomandibular Disorders. Philadelphia: Saunders, 2000:389-409.
24. van der Meulen MJ, Lobbezoo F, Aartman IH, Naeije M. Validity of the Oral Behaviours Checklist: Correlations between OBC scores and intensity of facial pain. J Oral Rehabil 2014;41:115–121.
25. Aaron LA, Turner JA, Mancl L, Brister H, Sawchuk CN. Electronic diary assessment of pain-related variables: Is reactivity a problem? J Pain 2005;6:107–115.
26. Glaros AG, Hanson AH, Ryen CC. Headache and oral parafunctional behaviors. Appl Psychophysiol Biofeedback 2014;39:59–66.
27. Chen CY, Palla S, Erni S, Sieber M, Gallo LM. Nonfunctional tooth contact in healthy controls and patients with myogenous facial pain. J Orofac Pain 2007;21:185–193.
28. Glaros AG, Williams K, Lausten L. The role of parafunctions, emotions and stress in predicting facial pain. J Am Dent Assoc 2005;136:451–458.
29. Glaros AG, Williams K, Lausten L, Friesen LR. Tooth contact in patients with temporomandibular disorders. Cranio 2005; 23:188–193.
30. Glaros AG. Temporomandibular disorders and facial pain: A psychophysiological perspective. Appl Psychophysiol Biofeedback 2008;33:161–171.
31. Steinberg L. Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. J Pers Soc Psychol 1994;66:341–349.
32. Scollon CN, Kim-Prieto C, Diener E. Experience sampling: Promises and pitfalls, strengths and weaknesses. J Happiness Studies 2003;4:5–34.
33. Hufford MR, Shields AL, Shiffman S, Paty J, Balabanis M. Reactivity to ecological momentary assessment: An example using undergraduate problem drinkers. Psychol Addict Behav 2002;16:205–211.
34. Conner TS, Tennen H, Fleeson W, Barrett LF. Experience Sampling Methods: A modern idiographic approach to personality research. Soc Personal Psychol Compass 2009;3: 292–313.
35. Johnson EI, Grondin O, Barrault M, et al. Computerized ambulatory monitoring in psychiatry: A multi-site collaborative study of acceptability, compliance, and reactivity. Int J Methods Psychiatr Res 2009;18:48–57.
36. Hamilton JC, Shuminsky TR. Self-awareness mediates the relationship between serial position and item reliability. J Pers Soc Psychol 1990;59:1301–1307.
37. Knowles ES, Byers B. Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? J Pers Soc Psychol 1996;70:1080–1090.
38. Stone AA, Broderick JE, Shiffman SS, Schwartz JE. Understanding recall of weekly pain from a momentary assessment perspective: Absolute agreement, between- and within-person consistency, and judged change in weekly pain. Pain 2004;107:61–69.
39. van Selms MK, Lobbezoo F, Wicks DJ, Hamburger HL, Naeije M. Craniomandibular pain, oral parafunctions, and psychological stress in a longitudinal case study. J Oral Rehabil 2004; 31:738–745.
40. Helkimo M. Studies on function and dysfunction of the masticatory system. IV. Age and sex distribution of symptoms of dysfunction of the masticatory system in Lapps in the north of Finland. Acta Odontol Scand 1974;32:255–267.
41. Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 1990;63:574–579.
42. Glaros AG, Williams K. Tooth contact versus clenching: Oral parafunctions and facial pain. J Orofac Pain 2012;26:176–180.