

Validating Indicators of Treatment Response: Application to Trichotillomania

Samuel O. Nelson, Kate Rogers, Natalie Rusch, Lauren McDonough, Elizabeth J. Malloy,
Martha J. Falkenstein, Maria Banis, and David A. F. Haaga
American University

Different studies of the treatment of trichotillomania (TTM) have used varying standards to determine the proportion of patients who obtain clinically meaningful benefits, but there is little information on the similarity of results yielded by these methods or on their comparative validity. Data from a stepped-care (Step 1: Web-based self-help; Step 2: Individual behavior therapy; $N = 60$) treatment study of TTM were used to evaluate 7 potential standards: complete abstinence, $\geq 25\%$ symptom reduction, recovery of normal functioning, and clinical significance (recovery + statistically reliable change), each of the last 3 being measured by self-report (Massachusetts General Hospital Hairpulling Scale; MGH-HPS) or interview (Psychiatric Institute Trichotillomania Scale). Depending on the metric, response rates ranged from 25 to 68%. All standards were significantly associated with one another, though less strongly for the 25% symptom reduction metrics. Concurrent (with deciding to enter Step 2 treatment) and predictive (with 3-month follow-up treatment satisfaction, TTM-related impairment, quality of life, and diagnosis) validity results were variable but generally strongest for clinical significance as measured via self-report. Routine reporting of the proportion of patients who make clinically significant improvement on the MGH-HPS, supplemented by data on complete abstinence, would bolster the interpretability of TTM treatment outcome findings.

Keywords: trichotillomania, clinical significance, validity, treatment outcome evaluation

It is widely agreed that statistical significance of between-group mean differences is an insufficient way to determine whether a therapy has clinically meaningful effects (e.g., Lambert, Hansen, & Bauer, 2008). One treatment could be very powerful yet fail to show statistical significance, for instance, if it is compared with a potent alternative treatment or tested in a small study or both. Another could have modest effects yet show statistical significance because it is tested against no treatment, with low within-group variability of results and high sample size. Accordingly, methodologists have advocated supplementing standard significance testing by reporting the proportion of patients who obtain practically meaningful benefits (e.g., Jacobson, Follette, & Revenstorf, 1984). Numerous methods for quantifying practically meaningful effects have been developed. For some disorders or conditions, the most suitable indicator could be complete eradication of the problem behavior. A prominent clinical practice guideline for smoking

cessation, for instance, judges treatments mainly on the basis of the proportion of patients showing 7-day point prevalence abstinence from smoking at 6-month follow-up (Fiore et al., 2008). Smoking rate reduction is considered insufficient to mark treatment success, in part because of health considerations, as even a modest rate of daily smoking (one to four cigarettes/day) significantly predicts heart disease mortality and all-cause mortality (Bjartveit, & Tverdal, 2005).

In other areas of treatment, however, abstinence may be unrealistic (e.g., schizophrenia symptoms) or supernormal and atypical of even the general, nonclinical population (e.g., worry, depressive symptoms). Therefore, various statistical metrics other than complete abstinence are often used instead to index meaningful benefit. A systematic review of such formulae recommended the clinical significance definition developed by Jacobson and colleagues (e.g., Jacobson & Truax, 1991) as provisionally the best choice (Lambert et al., 2008), while noting that few studies have been conducted on the comparative validity of alternate methods. In the Jacobson framework, a clinically significant response requires that the patient make statistically reliable change during treatment *and* recover normal-range functioning by the end of treatment.

The research reported in this article evaluated alternate indicators of practically meaningful treatment response in the context of a stepped-care protocol for treating trichotillomania (TTM). TTM is characterized by the recurrent pulling out of one's own hair, resulting in hair loss. TTM is associated with considerable psychosocial impairment. Relative to healthy control samples, people with TTM report greater disability (Diefenbach, Tolin, Hannan, Crocetto, & Worhunsky, 2005), lower life satisfaction (Diefenbach, Tolin, Hannan, et al., 2005), lower self-esteem (Diefenbach,

This article was published Online First April 7, 2014.

Samuel O. Nelson, Kate Rogers, Natalie Rusch, and Lauren McDonough, Department of Psychology, American University; Elizabeth J. Malloy, Department of Mathematics and Statistics, American University; Martha J. Falkenstein, Maria Banis, and David A. F. Haaga, Department of Psychology, American University.

This research was supported by Grant 1R15MH086852-01 from the National Institute of Mental Health. We are grateful to Charley Mansueto for consultation on this project.

Correspondence concerning this article should be addressed to David A. F. Haaga, Department of Psychology, Asbury Building, American University, Washington, DC 20016-8062. E-mail: dhaaga@american.edu

Tolin, Hannan, et al., 2005), and lower quality of life (Odlaug, Kim, & Grant, 2010). They often spend a great deal of time on hair pulling and may experience interference with their work and social lives (Wetterneck, Woods, Norberg, & Begotka, 2006). TTM can also have medical consequences, particularly if the patient ingests pulled hairs (McDonald, 2012). In studies of therapy for TTM, various methods have been employed to quantify a clinically meaningful response. Some investigators report the proportion of patients who achieve greater than a particular percentage reduction in symptom severity (e.g., 25%; Mouton-Odum, Keuthen, Wagener, Stanley, & Debakey, 2006). Others report the proportion of patients who recover a normal level of functioning (e.g., van Minnen, Hoogduin, Keijsers, Hellenbrand, & Hendriks, 2003). To our knowledge, no published studies have used the full Jacobson and Truax (1991) criteria, including requiring statistically reliable change. This is a potentially important distinction from recovery of normal functioning because someone could recover normal functioning without actually benefitting much from treatment if his or her pretreatment symptom scores were just barely higher than the cutoff demarcating normal functioning. Each of the three foregoing standards (25% reduction, recovery of normal functioning, and clinical significance per Jacobson and Truax) can, in turn, be based on self-report versus interviewer-rated symptoms, which were examined separately in our study in light of prior research showing that different measures and alternate perspectives often yield varying proportions of clinically significant responding (Lambert et al., 2008). Finally, although previous treatment outcome studies have not generally required it, one could define success in terms of complete abstinence from hair pulling for two reasons. First, while there are no systematic epidemiological studies of truly representative community samples, it seems likely that the majority of people do not pull their hair at all for noncosmetic reasons, making abstinence the modal, statistically “normal” state. Second, post-treatment abstinence has predicted superior 2-year follow-up results after behavior therapy for TTM (Keijsers et al., 2006).

The mere existence and use of differing methods of reporting clinically meaningful response to TTM treatment do not present a problem. Indeed, premature closure on a given approach would entail the risk of fostering a mono-method bias (Cook & Campbell, 1979), leaving researchers unable to detect the flaws of this consensus method. However, it is a problem that researchers do not yet know (a) how the methods compare with one another, nor (b) to the extent to which they differ, whether one is more valid in relation to external criteria than are the others. For reviewers or consumers of the treatment literature, a lack of information on overlap or relative leniency of the methods complicates the task of interpreting cross-study findings. For instance, if Treatment A was successful for 40% of TTM patients in reducing interviewer-rated symptoms by at least 25%, is this a better outcome than Treatment B’s success in another project for 28% of patients, with success defined as complete abstinence from hair pulling? To be sure, ultimately such evaluations should be based on within-study comparisons, but in the meantime, patients are being treated, and provisional decisions about which treatments enjoy the most empirical support need to be made. Likewise, the absence of comparative validation data makes it difficult to select a standard for determining when a course of treatment has been sufficiently successful, which is important for practicing clinicians who need to decide when to

terminate treatment versus when to extend it versus when to perhaps augment it with an additional treatment method.

Attempting to validate metrics for quantifying clinically meaningful response raises the question of what criteria should be used. There is no obvious gold standard criterion to which a valid index of treatment success should relate. Our research was conducted in the context of a stepped-care protocol for treatment of TTM in which patients were given the choice to proceed from Step 1 (Web-based self-help) to Step 2 (in-person individual therapy with habit reversal training). We treated as one external criterion the question of whether those who met a particular standard during Step 1 were less likely to proceed to Step 2. Although this is a fallible standard—a patient who had responded adequately might nevertheless be interested in receiving in-person care, and a patient who had not responded might become discouraged and give up—it seemed likely that a useful marker of clinically meaningful response should at least on average be related to the patient’s behavior in proceeding with further treatment or not. We also collected several kinds of predictive validity evidence for each of the various standards. Clinical response from baseline to the end of our stepped-care protocol (post-Step 2 assessment) was related to follow-up (3 months after post-Step 2 assessment) data on treatment satisfaction, functional impairment related to TTM, and quality of life. We expected that a valid indicator of whether someone has responded well to treatment should be associated with low impairment, high quality of life, and high treatment satisfaction 3 months later. We included diagnosis as a final criterion. One of the few previous studies of the predictive validity of clinical significance estimation methods related multiple statistical methods for quantifying clinical significance to depression relapse over the next 2 years (McGlinchey, Atkins, & Jacobson, 2002). In the present study, we likewise used prediction of TTM diagnostic status at 3-month follow-up as one of the validation criteria.

In summary, in a study of stepped care for TTM, we evaluated seven potential standards for judging clinically meaningful response—complete abstinence from pulling as well as (a) $\geq 25\%$ reduction in symptom scores, (b) recovery of normal functioning, or (c) clinical significance, crossed with self-reported or interviewer-rated symptoms—with regard to (a) comparative sensitivity/leniency, (b) convergence with the other standards, (c) concurrent validity in relation to choice of an additional treatment step, and (d) predictive validity in relation to 3-month follow-up treatment satisfaction, functional impairment, quality of life, and diagnosis.

Method

A detailed description of the methods used in this project may be found in Rogers et al. (in press), an initial report focusing on the efficacy of Web-based self-help and on acceptability of stepped care. The method is briefly described with emphasis on the measures used in evaluating clinical significance.

Participants

Sixty adults with TTM (57 women) enrolled in the study. The average age of participants was 33.18 years ($SD = 10.87$). The majority were White (75%), while 17% were African American,

3% Asian, 2% Native Hawaiian/other Pacific Islander, and 3% “other” race/ethnicity. One participant (2%) was Hispanic.

Participants were recruited via newspaper and online ads and clinician referrals. Inclusion criteria were that participants had to be at least 18 years old, to have regular access to the Internet, and to meet *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev., or *DSM-IV-TR*; American Psychiatric Association, 2000) criteria for TTM with the exception that Criteria B (tension before pulling) and Criteria C (pleasure, relief, or gratification when pulling) were not required, so as to not exclude participants with clinically significant hair pulling (Diefenbach, Tolin, Hannan, Maltby, & Crocetto, 2006). *DSM-5* criteria for TTM do not include the former B and C criteria (American Psychiatric Association, 2013).

Prospective participants were excluded if they reported any of the following within the past month: (a) suicidality, (b) major depressive episode, (c) psychosis, (d) severe anxiety, or (e) substance abuse. These are exclusion criteria for users of our Step 1 intervention, StopPulling.com, outside the research context. As in Woods, Wetterneck, and Flessner (2006), prospective participants were also excluded if they were in concurrent psychotherapy for TTM, or if they were taking medication for TTM and had not been on a stable dose for at least 4 weeks.

Materials

Measures of exclusion and inclusion criteria. The Structured Clinical Interview for *DSM-IV-TR* Axis I Disorders, Research Version—Patient Edition With Psychotic Screen (SCID-IP; First, Spitzer, Gibbon, & Williams, 2002) is a semistructured diagnostic interview, used at our initial assessment to evaluate exclusion criteria.

The Trichotillomania Diagnostic Interview (TDI; Rothbaum & Ninan, 2004) was used in diagnosing TTM. As noted earlier, we did not require that the *DSM-IV-TR* B or C criterion be met for a positive diagnosis of TTM. All TDIs were recorded, and a 20% random sample of the videos was coded by a second rater, who was masked to assessment period and treatment condition. Interrater agreement was high (overall agreement = 92%, $\kappa = .77$).

TTM symptoms. The Massachusetts General Hospital Hair-pulling Scale (MGH-HPS; Keuthen et al., 1995) is a seven-item self-report measure of hair-pulling symptoms in the previous week. Each item is rated on a scale from 0 to 4 (total = 0–28), with higher scores reflecting greater severity. Internal consistency reliability is high, and MGH-HPS total scores have been shown to be stable across brief retest intervals, sensitive to change during treatment, and able to discriminate TTM from anxiety and depression (O’Sullivan et al., 1995).

The Psychiatric Institute Trichotillomania Scale (PITS; Winchel et al., 1992) is a six-item, semistructured interviewer-rated measure of TTM symptom severity. Each item is rated on a scale from 0 to 7 (total = 0–42), with higher scores indicating greater severity. Total scores on the PITS have shown convergent validity with other clinician-rated TTM measures, but internal consistency of the PITS is low (Diefenbach, Tolin, Crocetto, Maltby, & Hannan, 2005). Administration of the PITS was recorded, and a 20% random sample of 53 videos was selected for coding by a second rater (masked to treatment condition and assessment point). Item 6 (severity of hair loss) could not be evaluated from videos. For the

sum of Items 1–5, single-rater reliability (Pearson correlation of interviewer rating with video-coder rating) was high ($r = .95$), with no significant difference in mean scores between raters.

Client satisfaction. The Client Satisfaction Questionnaire (CSQ-8; Larsen, Attkisson, Hargreaves, & Nguyen, 1979) is an eight-item self-report measure of satisfaction with health services. A sample item is “In an overall, general sense, how satisfied are you with the service you have received?” Total scores range from 8 to 32 (higher = more satisfied). CSQ-8 scores correlated negatively with therapy dropout (Larsen et al., 1979).

Impairment and quality of life. The Sheehan Disability Scale (SDS; Sheehan, 1983) is a three-item self-report measure of impairment in work/school, social life, and family life/home responsibilities. Each item is scored on a scale from 0 (*Not at all*) to 10 (*Extremely*; total = 0–30). In each case, the question deals with the extent to which symptoms (in this study, TTM symptoms) have disrupted one’s life in the indicated domain. In a primary care study, the SDS showed high internal consistency, and total scores were associated with six different psychiatric diagnoses (Leon, Olfson, Portera, Farber, & Sheehan, 1997). Retest reliability of SDS scores is high, as is convergence with clinician-rated functional status (Arbuckle et al., 2009). In a large TTM sample, SDS scores correlated positively with TTM symptom severity (Woods, Flessner, et al., 2006).

The World Health Organization Quality of Life—Brief Version (WHOQOL-BREF) is a 26-item self-report measure of quality of life (past 2 weeks) across four domains: physical health, psychological health, social relationships, and environment. Raw scores are converted to 4–20 (higher = better) domain scores; we used the mean of the domain scores, which has shown good temporal stability (WHOQOL Group, 1998).

Procedure

Design overview. Study procedures were approved by the institutional review board at American University, and participants were treated in accordance with the American Psychological Association ethical code. Those enrolling in the trial were randomized to immediate Step 1 access or to a waitlist (WL) condition; those in the WL condition completed a safety check-in after 5 weeks and a full assessment 10 weeks after baseline, before beginning Step 1. Otherwise, procedures were the same in each condition. Step 1 entailed 10 weeks of access to web-based self-help via StopPulling.com (with midpoint telephone check-in) and was followed by an in-person assessment. At this post-Step 1 assessment, participants chose whether to enter Step 2 HRT. Regardless of what they chose, an additional assessment (post-Step 2) was conducted 8 weeks later. After post-Step 2 assessment, there was no further treatment; the last assessment (follow-up) was 3 months later.

Baseline assessment and randomization. At baseline assessment, after informed consent was obtained, the interviews were administered, followed by the self-report measures described above except for the CSQ-8. Eligible and interested participants were randomly assigned at the end of the baseline assessment to either WL or immediate treatment. However, for the purposes of this report on outcome evaluation methods, the experimental conditions were combined, so the distinction between immediate and wait-list conditions is not discussed further.

Subsequent in-person assessments. The post-WL (for those in the WL condition), post-Step 1, post-Step 2, and follow-up assessments were largely identical; the TDI and PITS interviews were conducted, and all self-reports including the CSQ-8 were administered.

Treatments

Step 1: StopPulling.com. During Step 1, participants were given a code providing 10 weeks of free access to StopPulling.com. This program consisted of assessment, intervention, and maintenance modules. In the assessment module, participants self-monitored each instance of hair pulling or of urges to pull. They were asked to record detailed information about pulling episodes (i.e., motor behaviors, physical sensations, feelings, and thoughts preceding hair pulling, and what was done once the hair was pulled). In the intervention module, these assessment data were used to generate a list of specific interventions relevant to the individual's pulling style. Participants were provided with three suggested strategies per week, setting goals and rewarding themselves for their progress. After meeting their goals for 4 weeks, users proceeded to the maintenance module, in which they continued to record pulling episodes and utilize coping tactics.

Step 2: Habit reversal training (HRT). Those participants who opted to engage in Step 2 received eight weekly sessions of individual HRT with one of seven doctoral student therapists (with experience ranging from first to fifth year of clinical training) in a university outpatient clinic. All HRT sessions were videotaped for supervision and adherence assessment purposes. The HRT manual was a revision of Stanley and Mouton (1996). Changes included (a) adapting group to individual therapy, (b) extending the length of treatment, and (c) increasing the emphasis on stimulus control while decreasing the focus on relaxation. Our protocol thus highlighted components of HRT identified by Bloch, Weisenberger, Dombrowski, Nudel, and Coric (2007): (a) self-monitoring of pulling behavior and urges; (b) awareness training; (c) stimulus control to prevent opportunities to pull; and (d) stimulus-response or competing response training, or learning to substitute activities or physically incompatible behaviors when the urge to pull arises.

Operationalization of Response Criteria

Abstinence. Consistent with Keijsers et al. (2006), we indexed complete abstinence from hair pulling as a score of 0 on Item 4 of the MGH-HPS, which assesses frequency of pulling. Post-Step 1 abstinence was used for the concurrent validation analyses (relating response to Step 1 to the choice to enter Step 2 treatment), whereas post-Step 2 abstinence was used for all other analyses.¹

Symptom reduction of 25% or greater. Separately for self-report (MGH-HPS) and interview (PITS), we calculated the percentage of reduction in total scores as [for concurrent validity analyses] the baseline score minus post-Step 1 score, divided by the baseline score. If this value was equal to or greater than .25, the patient was deemed to have responded. For all other analyses (prevalence, predictive validity), we made parallel calculations of symptom reduction from baseline to post-Step 2 assessment.

Recovery of normal-range functioning. The participant was considered to have recovered on the PITS if she or he at post-Step

1 (for concurrent validity analysis) or post-Step 2 (for other analyses) obtained a total score of 14 or lower. For the MGH-HPS, a total score of 9 or lower was required.² In each case, the cutoff for return to normal functioning was based on Definition A from Jacobson and Truax (1991); in other words, the score had to be at least 2 standard deviations below the dysfunctional population mean. There are no norms on TTM symptom measures from a nondysfunctional population, so this is the only possible definition to use. We used our baseline sample to estimate the dysfunctional population mean. Previous TTM treatment studies reporting recovery of normal functioning based on the MGH-HPS have used cutoffs from 6.70 (Diefenbach et al., 2006; van Minnen et al., 2003) to 12.74 (Woods, Wetterneck, & Flessner, 2006), so our cutoff of 9 was in the middle of the range.

Clinical significance. Finally, clinical significance per Jacobson and Truax (1991) required both (a) recovery of normal-range functioning as defined earlier and (b) reliable change, meaning that the reliable change index (RCI) was equal to or greater than 1.96. RCI is the difference score (from baseline to post-Step 1 for concurrent validity analyses; from baseline to post-Step 2 for other analyses) divided by the standard error of the difference score. Calculating the standard error of the difference score requires a reliability estimate; for each measure, we used internal consistency, as recommended by Lambert et al. (2008). In our baseline sample, alpha was .74 for the MGH-HPS and was quite low for the PITS ($\alpha = .37$). Reliable change required a decrease of at least 10 points on the PITS and at least 6 points on the MGH-HPS.

Results

Attrition

Of the 60 participants enrolled at baseline, 54 completed post-Step 1 assessment (one of these was missing the MGH-HPS), and 50 completed post-Step 2 (six of these were missing the PITS interview). The six participants who missed post-Step 1 assessment did not differ significantly in baseline TTM symptoms (PITS $M = 24.33$, $SD = 3.01$; MGH-HPS $M = 16.33$, $SD = 4.03$) from those who completed the assessment (PITS $M = 23.76$, $SD = 4.65$; MGH-HPS $M = 16.98$, $SD = 3.71$). Likewise, the 10 participants who missed post-Step 2 assessment did not differ significantly in baseline TTM symptoms (PITS $M = 23.30$, $SD = 2.71$; MGH-HPS $M = 16.90$, $SD = 3.90$) from those who completed it (PITS $M = 23.92$, $SD = 4.78$; MGH-HPS $M = 16.92$, $SD = 3.72$).

Comparative Sensitivity or Lenience

Table 1 shows the proportion of eligible cases at each time point meeting each standard for clinically meaningful response. Several

¹ Patients would have been excluded from calculations of abstinence as an indicator of treatment response if they were already abstinent at baseline, but none was, so this stipulation eliminated no one from any analyses.

² Participants were ineligible for recovery if their baseline scores were already in the normal functioning range. For the PITS, this eliminated no participants, whereas for the MGH-HPS, this stipulation eliminated two participants from all analyses involving recovery as well as analyses of clinical significance, which subsumes recovery.

Table 1
Percentage of Eligible Cases Showing Clinically Meaningful Response

Standard	Baseline to post-Step 1	Baseline to post-Step 2
Abstinence (change from ≥ 1 to 0 on MGH-HPS Item 4)	6	26
$\geq 25\%$ reduction MGH-HPS	24	68
$\geq 25\%$ reduction PITS	18	57
Recovery on MGH-HPS (change from ≥ 10 to ≤ 9)	6	46
Recovery on PITS (change from ≥ 15 to ≤ 14)	11	39
Clinical significance on MGH-HPS (recovery + decrease of at least 6)	6	40
Clinical significance on PITS (recovery + decrease of at least 10)	2	25

Note. MGH-HPS = Massachusetts General Hospital-Hairpulling Scale; PITS = Psychiatric Institute Trichotillomania Scale.

trends are apparent in these descriptive data. Not surprisingly, the proportions are higher for the period from baseline to post-Step 2, which subsumes the earlier (baseline to post-Step 1) interval, reflecting the response of additional patients during Step 2 (habit reversal training). Also, response rates were a little higher for the self-report MGH-HPS than for the interviewer-rated PITS (the only exception being recovery of normal functioning during Step 1). Finally, it is quite clear that choice of response metric makes a sizable difference in outcome. Considering the entire stepped-care protocol (baseline to post-Step 2) and focusing, for instance, on the MGH-HPS, one could justify success rates ranging from 26% (abstinence), to 40% (clinical significance), 46% (recovery of normal-range functioning), or even 68% (at least 25% symptom reduction). Given that many studies select and report only one such metric, consumers of the research could derive very different perspectives on probability of successful treatment depending on which is chosen.

Convergent Validity of Treatment Response Indicators

Table 2 shows the convergence among our seven indicators of clinically meaningful response to the entire stepped-care protocol (baseline to post-Step 2). Pairwise agreement rates varied substan-

tially, from 56% (clinically significant response on the PITS with 25% or greater improvement on the MGH-HPS) to 95% (clinically significant response on the PITS with abstinence). Kappa coefficients were all significantly greater than zero and ranged from .26 (clinical significance on the PITS and 25% or greater improvement on the MGH-HPS) to .88 (clinical significance on the PITS and abstinence). Verbal classification of these values is of course somewhat arbitrary, but in the interobserver agreement context, Landis and Koch (1977) suggested the following labels for characterizing kappa coefficients: 0 or less = poor; .01-.20 = slight; .21-.40 = fair; .41-.60 = moderate; .61-.80 = substantial; and .81 or more = almost perfect. Both percentage-of-improvement indicators showed lower median kappas in relation to other indicators (PITS .47, MGH-HPS .37) than did recovery-of-normal-functioning indicators (PITS .62, MGH-HPS .56), clinical significance indicators (PITS .55, MGH-HPS .64), or abstinence (.58).

Concurrent Validity of Step 1 Response Indicators With Deciding Not to Enter Step 2

Table 3 shows the association of Step 1 response indicators with the decision to stop treatment and not enter Step 2 (habit reversal training). All indicators were positively associated with stopping treatment; in other words, treatment responders (by any standard) were more likely to conclude that they had received enough treatment, whereas Step 1 nonresponders were more likely to enter Step 2. Kappa coefficients significantly exceeded zero only for abstinence, improvement (25% or greater) on the MGH-HPS, and recovery on the PITS, but agreement of the response indicators with treatment discontinuation fell in a narrow range from 74% to 85%.

Validity of Step 2 Response Indicators for Predicting 3-Month Follow-Up Status

Table 3 also shows predictive validation evidence. TTM-related impairment at follow-up was on average ($M = 6.51$, $SD = 6.41$) quite similar to the impairment reported on the same measure by a TTM sample in Diefenbach, Tolin, Hannan, et al. (2005; $M = 6.14$, $SD = 6.17$). Individual variation in impairment scores, however, was not significantly predicted by any indicator. In absolute value, most associations were small by conventional effect size standards, with the exception of clinical significance on the MGH-HPS, which approached a medium effect ($\eta = -.26$).

The only significant predictor of 3-month follow-up diagnosis (i.e., not being judged by the interviewer as meeting TTM criteria)

Table 2
Convergent Validity (% Agreement (κ)) for Baseline to Post-Step 2 Response Indicators

Variable	1	2	3	4	5	6	7
1. Abstinence	—						
2. $\geq 25\%$ MGH-HPS	.58 (.28)	—					
3. $\geq 25\%$ PITS	.72 (.47)	.74 (.47)	—				
4. Recovery MGH-HPS	.79 (.57)	.77 (.56)	.76 (.52)	—			
5. Recovery PITS	.81 (.58)	.60 (.27)	.73 (.47)	.83 (.65)	—		
6. Clinical significance MGH-HPS	.85 (.67)	.71 (.46)	.73 (.48)	.94 (.87)	.85 (.68)	—	
7. Clinical significance PITS	.95 (.88)	.56 (.26)	.68 (.40)	.76 (.49)	.86 (.69)	.83 (.61)	—

Note. Every pairwise association in this table was significant ($p < .05$) by Fisher's exact test. MGH-HPS = Massachusetts General Hospital Hairpulling Scale; PITS = Psychiatric Institute Trichotillomania Scale.

Table 3
Concurrent and Predictive Validity of Response Indicators

Response indicator	Decision to not enter Step 2 (% agreement, κ) Prevalence = 24%	3-month follow-up			
		Treatment satisfaction (η) $M = 28.00$, $SD = 4.49$	Impairment (η) $M = 6.51$, $SD = 6.41$	Quality of life (η) $M = 15.85$, $SD = 2.28$	Absence of TTM diagnosis (% agreement, κ) Prevalence = 33%
Abstinence	80 (.29)**	.14	-.09	.14	69 (.26)
≥25% MGH-HPS	85 (.59)***	.40**	-.11	.29*	50 (.10)
≥25% PITS	74 (.26)	.18	-.05	.16	61 (.26)
Recovery MGH-HPS	78 (.19)	.27	-.15	.23	61 (.19)
Recovery PITS	78 (.29)*	-.03	-.02	.21	66 (.27)
Clinical significance MGH-HPS	78 (.19)	.28	-.26	.38**	63 (.20)
Clinical significance PITS	76 (.10)	.10	-.01	.10	70 (.32)*

Note. Decision to not enter Step 2 was analyzed in relation (concurrently) to response indicators from baseline to post-Step 1. The four follow-up criteria in this table were analyzed in relation (predictively) to response indicators from baseline to post-Step 2. TTM = treatment of trichotillomania; MGH-HPS = Massachusetts General Hospital Hairpulling Scale; PITS = Psychiatric Institute Trichotillomania Scale.

* $p < .05$. ** $p < .01$. *** $p < .001$.

was a clinically significant response on the interviewer-rated PITS (70% hit rate, $\kappa = .32$). For follow-up treatment satisfaction, MGH-HPS improvement (25% or greater from baseline to post-Step 2) was a significant predictor ($\eta = .40$), and for follow-up quality of life both MGH-HPS improvement ($\eta = .29$) and MGH-HPS clinical significance ($\eta = .38$) were significant predictors.

Discussion

Knowing that a treatment helps on average is important, but being able to determine whether, or more likely how often, it works well enough to achieve practically meaningful benefits is important from several perspectives. For instance, Miller and Manuel (2008) introduced a survey method for determining how strong an effect a treatment must have to make a difference to treatment providers. A sample finding was that substance use treatment providers would be interested in learning a new method if it made a difference of at least 10 points in the percentage of patients arrested for driving while intoxicated.

Other constituencies needing information on how often practically meaningful effects are achieved include patients, family members, therapists needing to decide whether a treatment has helped sufficiently versus needs extension or augmentation, and reviewers interested in synthesizing the research literature on efficacy of interventions. Some indicators of practical benefit have a directly interpretable, obvious importance in a specific realm of intervention (e.g., 5-year survival rates in cancer treatment). In working with conditions in which life or death is not at issue, there is more of a need for a logical and empirical rationale for the preferred standard. Social validation approaches entail determining whether treatment-related improvements are sufficiently striking as to be confirmed by those close to the patient (teachers, parents, and significant others; Kazdin, 1977). These methods seem ideal for, say, child aggression; if a child's parents, teachers, and classmates view him or her as less aggressive than before, and no more aggressive than an average child, these reports carry substantial weight.

For trichotillomania, however, social validation may have limited utility. Hair pulling behavior often occurs in private, and many people with TTM take pains to hide hair loss effects from people around them. Accordingly, we used data from a stepped-care study

(Web-based self-help followed at the patient's discretion by in-person individual habit reversal training) to evaluate TTM treatment response indicators derived from self-report and interview measures. The descriptive data (Table 1) make clear that the choice of indicator is consequential, with baseline-to-post-Step 2 response rates ranging from 25% to 68%. We consider each in turn and make recommendations for future research and practice.

Abstinence from hair pulling could usefully be reported in TTM treatment studies. It is a conservative metric, yielding the lowest response rate at post-Step 2 assessment other than clinical significance on the PITS (Table 1). Independent of any future findings, one can be confident that an abstainer is within the normal range for hair pulling. Finally, a study with a longer follow-up than ours (Keijsers et al., 2006) showed lower TTM symptom scores 2 years later for posttreatment abstainers. By the same token, we would argue against reporting solely abstinence. The Keijsers et al. (2006) finding was from one small ($N = 28$) sample; to our knowledge was not replicated; and used a subsequent administration of the MGH-HPS (same measure incorporating the abstinence item) as the criterion, maximizing common method effects. In our study, abstinence was concurrently related to treatment discontinuation after Step 1 but was not significantly related to follow-up criteria. More generally, abstinence as the sole indicator of clinically meaningful response would (a) lack comparability to metrics used across disorders and (b) lack content validity in that it does not take into account urges to pull hair, difficulty resisting pulling, or distress about pulling.

The improvement rate indicators (25% or greater) cast the success of the stepped-care protocol in the most favorable light (see Table 1), but we do not recommend their general use. They converged the least well with other standard indicators of TTM treatment success in our study (see Table 2). Also, in contrast to recovery or clinical significance, they do not entail achievement of a healthy end-state. A participant with the maximum total score (28) on the MGH-HPS at baseline, for example, could show a 25% reduction to 21 and still be reporting more severe symptoms than our baseline sample mean (16.92). It seems unlikely that most therapists or research reviewers would be content with this progress as evidence of success.

Recovery and, by extension, *clinical significance* metrics share the limitation that they are not applicable to patients who begin treatment with symptom scores in the normal range. In our study, this was a minor issue, affecting no participants on the PITS and two on the MGH-HPS (see Footnote 2). It could be eliminated as a concern in future studies by establishing minimum symptom severity standards outside the normal range as an inclusion criterion. Such an a priori minimum could be established in TTM if functional-population norms were collected or if a meta-analysis of clinical studies established a general-purpose dysfunctional-population norm. As indicated by Lambert et al. (2008), nonapplicability to less symptomatic patients is likely to be a more common issue in routine clinical practice or in effectiveness studies conducted in such settings than in efficacy trials, which often use fairly stringent symptom severity cutoffs as inclusion criteria.

Even if applicable to all patients, *recovery* criteria have limitations, and we do not recommend their routine use to characterize success rates in TTM treatment. As noted in the introduction, recovery is less conservative than clinical significance in the sense that it does not require the patient to have made statistically reliable change. A reduction of just 1 point, for example, on the MGH-HPS, from 10 to 9 would constitute recovery. From Table 1, it is clear that this phenomenon, though not common, does exist, and inflated success rates relative to clinical significance (from 40% to 46% on the MGH-HPS at post-Step 2; from 25% to 39% on the PITS). Moreover, neither recovery indicator was significantly predictive of any of the follow-up criteria (Table 3).

Consistent with the review by Lambert et al. (2008) with regard to psychotherapy in general, we concluded that for treatment of TTM *clinical significance* per Jacobson and colleagues is the best metric for success. More specifically, though validity results were not completely consistent in our study, we recommend that TTM researchers routinely report clinical significance per the MGH-HPS. The low internal consistency reliability of the PITS meant that a more substantial change on that measure was needed for a reliable change index of at least 1.96, and therefore, we believe the lower clinical significance proportion on the PITS (25% vs. 40% for MGH-HPS at post-Step 2) reflects insensitivity of the PITS to actual change rather than excessive leniency of the MGH-HPS. Also, though not statistically significant in most cases, clinical significance on the MGH-HPS generally showed medium effects in validity analyses and significantly predicted ($\eta = .38$) 3-month follow-up quality of life. Effect sizes were greater for clinical significance on the MGH-HPS than for clinical significance on the PITS in relation to four of five criterion-related validity analyses (Table 3).

Method Limitations and Directions for Future Research

Method limitations of this research include the relatively brief (3-month) follow-up and modest sample size. Conversely, strengths include incorporation of multiple response metrics, allowing for an initial comparison of stringency and several types of validation evidence (convergent, concurrent, and predictive) for standards used in TTM treatment research. Those conducting future research would do well to replicate these findings in larger samples, with longer follow-ups, and with additional validation criteria such as the social and economic consequences of TTM

(Wetterneck et al., 2006). It would be useful in such research to include other TTM severity measures as well. For instance, the National Institute of Health Trichotillomania Severity Scale (NIMH-TSS; Swedo et al., 1999) could be included as an alternative to the PITS, which showed weak internal consistency in our study. The NIMH-TSS and the PITS correlated .75 with one another and fared similarly in other psychometric analyses in Diefenbach, Tolin, Crocetto, et al. (2005), but to our knowledge, there have been no previous comparative studies of these interviewer ratings as markers of treatment response.

Another valuable research direction would be a meta-analysis of dysfunctional-population norms on TTM symptom measures. As noted in the Method section, recovery standards on the MGH-HPS vary across studies, detracting from the potential of clinical significance on this measure to function as a uniform metric of success in TTM treatment. A comprehensive analysis of clinical samples in the literature could be used to justify selection of one norm for the cutoff score reflecting 2 standard deviations less than the dysfunctional population mean.

Conclusion

We recommend that TTM researchers routinely report, along with whatever other analyses make sense in light of their hypotheses, the proportion of participants in each treatment condition who achieve a clinically significant response on the MGH-HPS, as per the Jacobson analytic framework. Complete abstinence from hair pulling could usefully be reported as a supplementary metric. Adherence to these recommendations would promote comparability across studies and ease interpretation of the burgeoning literature on treating TTM.

References

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- Arbuckle, R., Frye, M. A., Brecher, M., Paulsson, B., Rajagopalan, K., Palmer, S., & Innocenti, A. D. (2009). The psychometric validation of the Sheehan Disability Scale (SDS) in patients with bipolar disorder. *Psychiatry Research*, *165*, 163–174. doi:10.1016/j.psychres.2007.11.018
- Bjartveit, K., & Tverdal, A. (2005). Health consequences of smoking 1–4 cigarettes per day. *Tobacco Control*, *14*, 315–320. doi:10.1136/tc.2005.011932
- Bloch, M. H., Weisenberger, A. L., Dombrowski, P., Nudel, J., & Coric, V. (2007). Systematic review: Pharmacological and behavioral treatment for trichotillomania. *Biological Psychiatry*, *62*, 839–846. doi:10.1016/j.biopsych.2007.05.019
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Diefenbach, G. J., Tolin, D. F., Crocetto, J., Maltby, N., & Hannan, S. (2005). Assessment of trichotillomania: A psychometric evaluation of hair-pulling scales. *Journal of Psychopathology and Behavioral Assessment*, *27*, 169–178. doi:10.1007/s10862-005-0633-7
- Diefenbach, G. J., Tolin, D. F., Hannan, S., Crocetto, J., & Worhunsky, P. (2005). Trichotillomania: Impact on psychosocial functioning and quality of life. *Behaviour Research and Therapy*, *43*, 869–884. doi:10.1016/j.brat.2004.06.010
- Diefenbach, G. J., Tolin, D. F., Hannan, S., Maltby, N., & Crocetto, J. (2006). Group treatment for trichotillomania: Behavior therapy versus

- supportive therapy. *Behavior Therapy*, 37, 353–363. doi:10.1016/j.beth.2006.01.006
- Fiore, M. C., Jaén, C. R., Baker, T. B., Bailey, W. C., Benowitz, N. L., Curry, S. J., . . . Wewers, M. E. (2008). *Treating tobacco use and dependence: 2008 update* (Clinical practice guideline). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service.
- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders—Research Version, Patient Edition With Psychotic Screen*. New York, NY: New York State Psychiatric Institute, Biometrics Research.
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, 15, 336–352. doi:10.1016/S0005-7894(84)80002-7
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19. doi:10.1037/0022-006X.59.1.12
- Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1, 427–452. doi:10.1177/014544557714001
- Keijsers, G. P. J., van Minnen, A., Hoogduin, C., Klaassen, B., Hendriks, M., & Tanis-Jacobs, J. (2006). Behavioural treatment of trichotillomania: Two-year follow-up results. *Behaviour Research and Therapy*, 44, 359–370. doi:10.1016/j.brat.2005.03.004
- Keuthen, N. J., O’Sullivan, R. L., Ricciardi, J. N., Shera, D., Savage, C. R., Borgman, A. S., . . . Baer, L. (1995). The Massachusetts General Hospital (MGH) Hairpulling Scale: I. Development and factor analysis. *Psychotherapy and Psychosomatics*, 64, 141–145. doi:10.1159/000289003
- Lambert, M. J., Hansen, N. B., & Bauer, S. (2008). Assessing the clinical significance of outcome results. In A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 359–378). New York, NY: Oxford University Press.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. doi:10.2307/2529310
- Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., & Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Evaluation and Program Planning*, 2, 197–207. doi:10.1016/0149-7189(79)90094-6
- Leon, A. C., Olfson, M., Portera, L., Farber, L., & Sheehan, D. V. (1997). Assessing psychiatric impairment in primary care with the Sheehan Disability Scale. *International Journal of Psychiatry in Medicine*, 27, 93–105. doi:10.2190/T8EM-C8YH-373N-1UWD
- McDonald, K. E. (2012). Trichotillomania: Identification and treatment. *Journal of Counseling & Development*, 90, 421–426. doi:10.1002/j.1556-6676.2012.00053.x
- McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance methods: Which one to use and how useful are they? *Behavior Therapy*, 33, 529–550. doi:10.1016/S0005-7894(02)80015-6
- Miller, W. R., & Manuel, J. K. (2008). How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. *Drug and Alcohol Review*, 27, 524–528. doi:10.1080/09595230801956165
- Mouton-Odum, S., Keuthen, N. J., Wagener, P. D., Stanley, M. A., & Debaek, M. E. (2006). *StopPulling.com*: An interactive, self-help program for trichotillomania. *Cognitive and Behavioral Practice*, 13, 215–226. doi:10.1016/j.cbpra.2005.05.004
- Odlaug, B. L., Kim, S. W., & Grant, J. E. (2010). Quality of life and clinical severity in pathological skin picking and trichotillomania. *Journal of Anxiety Disorders*, 24, 823–829. doi:10.1016/j.janxdis.2010.06.004
- O’Sullivan, R. L., Keuthen, N. J., Hayday, C. F., Ricciardi, J. N., Buttolph, M. L., Jenike, M. A., & Baer, L. (1995). The Massachusetts General Hospital (MGH) Hairpulling Scale: 2. Reliability and validity. *Psychotherapy and Psychosomatics*, 64, 146–148. doi:10.1159/000289004
- Rogers, K., Banis, M., Falkenstein, M. J., Malloy, E. J., McDonough, L., Nelson, S. O., . . . Haaga, D. A. F. (in press). Stepped care in the treatment of trichotillomania. *Journal of Consulting and Clinical Psychology*.
- Rothbaum, B. O., & Ninan, P. T. (1994). The assessment of trichotillomania. *Behaviour Research and Therapy*, 32, 651–662. doi:10.1016/0005-7967(94)90022-1
- Sheehan, D. V. (1983). *The anxiety disease*. New York, NY: Scribner’s.
- Stanley, M. A., & Mouton, S. G. (1996). Trichotillomania treatment manual. In V. B. Van Hasselt & M. Hersen (Eds.), *Sourcebook of psychological treatment manuals for adult disorders* (pp. 657–687). New York, NY: Plenum Press. doi:10.1007/978-1-4899-1528-3_17
- Swedo, S. E., Leonard, H. L., Rapoport, J. L., Lenane, M. C., Goldberger, B. A., & Cheslow, B. A. (1989). A double-blind comparison of clomipramine and desipramine in the treatment of trichotillomania (hair pulling). *New England Journal of Medicine*, 321, 497–501. doi:10.1056/NEJM198908243210803
- van Minnen, A., Hoogduin, A. L., Keijsers, G. P. J., Hellenbrand, I., & Hendriks, G. (2003). Treatment of trichotillomania with behavioral therapy or fluoxetine: A randomized, waiting-list controlled study. *Archives of General Psychiatry*, 60, 517–522. doi:10.1001/archpsyc.60.5.517
- Wetterneck, C. T., Woods, D. W., Norberg, M. M., & Begotka, A. M. (2006). The social and economic impact of trichotillomania: Results from two nonreferred samples. *Behavioral Interventions*, 21, 97–109. doi:10.1002/bin.211
- World Health Organization Quality of Life Group (WHOQOL Group). (1998). Development of the World Health Organization WHOQOL-BREF Quality of Life Assessment. *Psychological Medicine*, 28, 551–558. doi:10.1017/S0033291798006667
- Winchel, R. M., Jones, J. S., Molcho, A., Parsons, B., Stanley, B., & Stanley, M. (1992). The Psychiatric Institute Trichotillomania Scale (PITS). *Psychopharmacology Bulletin*, 28, 463–476.
- Woods, D. W., Flessner, C. A., Franklin, M. E., Keuthen, N. J., Goodwin, R. D., Stein, D. J., . . . Trichotillomania Learning Center-Scientific Advisory Board. (2006). The Trichotillomania Impact Project (TIP): Exploring phenomenology, functional impairment, and treatment utilization. *Journal of Clinical Psychiatry*, 67, 1877–1888. doi:10.4088/JCP.v67n1207
- Woods, D. W., Wetterneck, C. T., & Flessner, C. A. (2006). A controlled evaluation of acceptance and commitment therapy plus habit reversal for trichotillomania. *Behaviour Research and Therapy*, 44, 639–656. doi:10.1016/j.brat.2005.05.006

Received August 8, 2013

Revision received December 9, 2013

Accepted February 12, 2014 ■