Photographic Assessment of Change in Trichotillomania: Psychometric Properties and Variables Influencing Interpretation

David C. Houghton[1] · **Maddison R. Franklin**[1] · **Michael P. Twohig**[2] ·
Martin E. Franklin[3] · **Scott N. Compton**[4] · **Angela M. Neal-Barnett**[5] ·
Stephen M. Saunders[6] · **Douglas W. Woods**[1]

Abstract Although photographic assessment has been found to be reliable in assessing hair loss in Trichotillomania, the validity of this method is unclear, particularly for gauging progress in treatment. The current study evaluated the psychometric properties of photographic assessment of change in Trichotillomania. Photographs showing hair loss of adults with Trichotillomania were taken before and after participating in a clinical trial for the condition. Undergraduate college students ($N = 211$) rated treatment response according to the photos, and additional archival data on hair pulling severity and psychosocial health were retrieved from the clinical trial. Photographic assessment of change was found to possess fair reliability (ICC = 0.53), acceptable criterion validity ($r = 0.51$), good concurrent validity ($r = 0.30$–0.36), and excellent incremental validity ($\Delta R^2 = 8.67$, $p < 0.01$). In addition, photographic measures were significantly correlated with change in quality of life ($r = 0.42$), and thus could be considered an index of the social validity of Trichotillomania treatment. Gender of the photo rater and pulling topography affected the criterion validity of photographic assessment (partial $\eta^2 = 0.05$–0.11). Recommendations for improving photographic assessment and future directions for hair pulling research are discussed.

Keywords Hair pulling · Psychometrics · Assessment · Behavioral observation

Introduction

Trichotillomania (TTM) is a psychiatric disorder characterized by chronic hair pulling, resulting in significant hair loss (American Psychiatric Association 2013). Several methods of assessing the disorder have been developed, including self-report questionnaires, clinician-rated scales, diagnostic interviews, self-monitoring, and product measures (e.g., saving pulled hairs, photographing affected areas); all of which tend to focus on different facets of the disorder (Franklin and Tolin 2007). For instance, self- and clinician-rated assessments can measure cognitive and affective aspects of TTM, diagnostic interviews capture core aspects of TTM symptomology, and self-monitoring methods (e.g., recording instances of hair pulling in a diary) help establish the frequency of the behavior. However, assessments that depend on patient reports and clinical judgment might be limited by their subjective nature. As such, measures that are thought to be more objective, such as photographic assessment, might be a valuable addition to a multi-modal assessment strategy (Diefenbach et al. 2005).

TTM researchers and clinicians have long used photographic methods for assessing the extent of hair loss (Friman et al. 1984; Rothbaum and Ninan 1994). A single observation

✉ David C. Houghton
davidhoughton@tamu.edu

[1] Department of Psychology, Texas A & M University, 4235 TAMU, College Station, TX 77843-4235, USA

[2] Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, UT 84322-2810, USA

[3] Department of Psychiatry, University of Pennsylvania School of Medicine, 3535 Market Street, 6th Floor, Philadelphia, PA 19104, USA

[4] Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, DUMC 3527, Durham, NC 27710, USA

[5] Department of Psychology, Kent State University, 203 Kent Hall Addition, Kent, OH 44242, USA

[6] Department of Psychology, Marquette University, 328E Cramer Hall, Milwaukee, WI 53233, USA

can be used to convey the current severity of hair loss in TTM, and photos can also be obtained pre- and post-treatment, thus allowing hair loss to be recorded before and after intervention. Photographic assessments can be quantified through several methods, such as measuring the circumference of bald spots, counting hairs within the bald spot, or rating the damage on a Likert-type scale.

In addition to quantifying the severity of hair loss, photographic assessment can also be used to measure the social validity of TTM intervention. Social validity refers to the degree to which consumers of an intervention (e.g., client, family members, friends) endorse the therapeutic processes, therapist characteristics, goals, and overall effects on targeted symptoms (Baer et al. 1987). Indeed, one important aspect of an intervention's effect on hair pulling symptoms is the degree to which subjects regrow hair. Support for this notion comes from evidence indicating that persons with TTM are viewed negatively (Boudjouk et al. 2008; Long et al. 1999; Marcks et al. 2005; Ricketts et al. 2012; Woods et al. 1999) and that psychosocial impairment associated with TTM is largely due to embarrassment and shame caused by hair loss (Casati et al. 2000; Soriano et al. 1996; Stemberger et al. 2000; Wetterneck et al. 2006). Photographic assessment could theoretically serve several purposes in documenting changes in hair loss during treatment. First, a series of photographs could be shown to TTM clients in order to provide an easily interpretable visual representation of their change throughout treatment. In turn, this could increase the client's motivation and validate their therapeutic efforts. Second, therapists could consult photographs taken throughout treatment when making global assessments of change and in other aspects of clinical decision making, such that they do not have to rely on memory when trying to determine if hair loss has improved over treatment. Finally, photographic assessments could be used as secondary outcome measures in TTM treatment trials, as they might provide a more socially valid index of change that is not captured by standard hair pulling severity instruments.

Although photographic measures are theoretically useful, there is a paucity of data on the psychometric properties of this method. Preliminary studies have shown that photographic assessment possesses strong inter-rater reliability ($r = 0.87$; Tolin et al. 2002), and that in-person visual assessment of hair loss severity boasts excellent inter-observer agreement ($r = 0.93$; Diefenbach et al. 2005). More recently, Haaga et al. (2015) conducted a cross-sectional psychometric analysis of photographic ratings of hair loss in TTM and found evidence of high test-retest reliability and convergent validity (with in-person ratings of hair loss severity). However, photographic hair loss ratings were not significantly associated with measures of hair pulling severity or impairment. This result might be due to the fact that common measures of TTM severity address factors not directly related to hair loss, such as frequency and intensity of urges to pull. As such, because

photographic assessments are somewhat insensitive to some clinical features of TTM, hair loss as rated from photographs may not be a good direct measure of TTM severity at a single time point.

Several other factors may limit the clinical utility of photographic assessment of TTM. Not only are some patients uncomfortable being photographed (Franklin and Tolin 2007; Haaga et al. 2015), but factors such as camera lens type, lighting, angle of the photograph, hair style, and facial expression are thought to significantly affect the interpretation of photographs of hair loss (Schosser and Kendrick 1987; Slue 1989). Several demographic characteristics (e.g., gender, ethnicity, age) of individuals with TTM, as well as those rating the photos, might bias interpretation of photographic measures. For example, research has shown that women's feelings of identity as they relate to femininity, sexuality, attractiveness, and personality are more strongly linked to hair than for men (Wolf 1991), suggesting that women may be more perceptive of subtle hair-related changes than men. Research has also shown that when people assess physical attractiveness, their ratings are often moderated by the age and skin tone of the rated subject (Hunter 2007; McKelvie 1993), implying that perception of hair-related changes might be biased by physical characteristics related to age and ethnicity. There is also reason to believe that differences in TTM symptomatology might affect perception of photographic measures. Pulling most often occurs on the scalp, but also occurs frequently on the eyebrows and eyelashes (Woods et al. 2006). Research on facial perception has found that hair loss on the scalp is associated with both positive and negative attributions (Helleström and Teckle 1994), but the absence of eyebrows actually interferes with facial processing (Sadrô et al. 2003). Thus, people might be more familiar with encountering varying degrees of scalp baldness, especially in men, but less familiar with the absence of eyebrows. The eyebrows and eyelashes also contain fewer hairs and grow more slowly than the scalp (Cohen 2010; Myers and Hamilton 1951), meaning that changes in hair growth on the eyebrows or eyelashes during acute treatment might be more difficult to detect than on the scalp. Indeed, Haaga et al. (2015) argued that hair-regrowth during treatment occurs at different rates due to pulling site, sex, and age of the individual, meaning that many factors other than hair pulling affect the appearance of hair over the course of treatment.

Nevertheless, there are no empirical data on whether photographic assessment is sensitive to changes in pulling across treatment. This question is important because photographic assessment during TTM treatment could be used as a highly illustrative and socially valid measurement tool. For photographic measures to be considered a valid measure of change in TTM, it must be demonstrated that change in hair loss, as measured from photographs, is associated with change in other well-validated measures of TTM severity and treatment

improvement. Given the limited amount of information available about the reliability and validity of photographic measures of change in TTM, the current study sought to perform a psychometric assessment of this method using photographs of clinical patients. We evaluated the inter-rater reliability of photographic measures of change, as well as their criterion, concurrent, and incremental validity. We also examined whether criterion validity of photographic assessments varied (i.e., were moderated) by key demographic features of TTM patients.

Methods

Participants Human subjects review board approval for the current project was obtained at Texas A & M University (IRB2014-0282D). All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 2013 Declaration of Helsinki and comparable ethical standards. Informed consent was obtained from all participants included in the study. Participants were undergraduate students enrolled in introductory psychology courses (N = 211) who received course credit for participation. All participants were college-aged (Range = 18–27, mean age = 19.05) and 62.1 % were female. Ethnicity varied between participants, with 63.5 % Caucasian, 19 % Hispanic, 8.1 % Asian, 6.2 % African American, and 3.3 % of mixed ethnicity. Inclusion criteria were age ≥ 18, enrollment in an introductory psychology class at Texas A & M, and fluency in English. Two participants who provided informed consent were later excluded from the study for violating inclusion criteria (1 non-English speaking and 1 under age 18).

Materials Photo sets of TTM patients, as well as data from corresponding primary and secondary outcome measures, were taken from the data set of a recently completed randomized controlled trial of psychotherapy for TTM. The study was funded by the National Institutes of Mental Health (Grant #R01MH080966; D. Woods, PI), is publicly registered on ClinicalTrials.gov (#NCT00872742), and was approved for archival analysis at Texas A & M University (IRB2013-3025 M). In the clinical trial, participants were randomized to receive either Acceptance-Enhanced Behavior Therapy (Twohig and Woods 2004) or psychoeducation and supportive psychotherapy. Sixty-nine TTM participants completed the 12 weeks treatment and consented to have their photos used for educational and research purposes. Participants were photographed by blinded evaluators in their most frequent pulling area prior to the first treatment session and again after the last session. Photographic assessment was only directed at the facial and scalp region(s), from which pulling most frequently occurred.

Using a *Canon PowerShot A470* digital camera with 7.1-megapixel resolution, participants were photographed while sitting down in the clinic where treatment was conducted. Most photographs were taken at a distance that captured the entire face or head and omitted additional body parts, but some photos were closely focused on the pulling site and the whole head was not visible. Each TTM participant contributed only one photo set to the present study, and if multiple sets existed (for the same or different pulling sites), the photo set that had the clearest resolution and best captured the pulling area, as determined by the first author (D.C.H.), was used. Forty-one of individuals who completed treatment (N = 69) had "usable" photo sets for the current study. "Usable" photo sets had to meet several criteria: (1) have clearly visible hair pulling effects (i.e., those with blurry or unfocused photographs were not used), (2) contain at least two photos (e.g., baseline and post-treatment) of the same area of hair pulling as well as similar distance and angle from the head, and (3) the participant consented to the use of their photos. The self-report measures include the Massachusetts General Hospital Hairpulling Scale, Beck Depression Inventory, Beck Anxiety Inventory, and the Quality of Life Inventory. The clinician-rated measures included the NIMH Trichotillomania Symptom Severity Scale and Clinical Global Impressions Scale. These measures are detailed below.

The Massachusetts General Hospital Hairpulling Scale (MGH-HPS; Keuthen et al. 1995) is a 7-item self-report questionnaire that measures frequency, resistance, and control of hair-pulling urges and behaviors as well as distress associated with hair pulling. Each item is rated on a 5-point Likert scale ranging from 0 (lower severity) to 4 (higher severity). The total score is acquired by summing the responses for all 7 items. The MGH-HPS has consistently demonstrated strong internal consistency ($\alpha = 0.89$) and test-retest reliability ($r = 0.97$; Keuthen et al. 1995; O'Sullivan et al. 1995), as well as acceptable convergent and divergent validity (O'Sullivan et al. 1995). In the current sample, internal consistency of the MGH-HPS was good ($\alpha = 0.83$).

The National Institute of Mental Health Trichotillomania Symptom Severity Scale (NIMH-TSS; Swedo et al. 1989) is a clinical interview that consists of 5 items assessing time spent pulling the past week, time spent pulling the previous day, resistance to pulling, distress from pulling, and interference with daily life. Items are rated on a scale ranging from 0 to 5. The total score is acquired by summing the responses for all 5 items. Higher scores indicate greater symptom severity. The NIMH total score has demonstrated adequate inter-rater reliability and acceptable internal consistency, but mixed concurrent validity (Diefenbach et al. 2005). In the current sample, the reliability of the NIMH-TSS was very low ($\alpha = 0.41$). The low internal constancy of the NIMH-TSS suggests that

results derived from this measure should be interpreted with caution.

The Clinical Global Impressions Scale (CGI; Guy 1976) consists of two clinician-rated single-item measures that assess the client's global improvement (CGI-I) and severity (CGI-S). The CGI-S is rated on a scale of 1 (normal, not at all ill) to 7 (extremely ill), and the CGI-I is rated on a scale of 1 (very much improved) to 8 (very much worse). The CGI has good to strong psychometric properties (Dahlke et al. 1992; Leon et al. 1993) and has been used to measure outcome in adult TTM (Diefenbach et al. 2006; Grant et al. 2009; Ninan et al. 2000). In the current sample, the CGI-S showed excellent test-retest reliability between the screening and baseline assessment periods ($r_s = 0.78$).

The Beck Depression Inventory II (BDI; Beck 1972) is a 21-item self-report measure of depression severity in the previous 2 weeks. Items are rated on a scale of 0 (no depressive symptom) to 3 (severe depressive symptom). Higher scores indicate greater depression severity. The BDI has demonstrated strong internal consistency, high test-retest reliability and good convergent validity (Sprinkle et al. 2002). In the current sample, internal consistency of the BDI was questionable ($\alpha = 0.60$). The low internal constancy of the BDI suggests that results derived from this measure should be interpreted with caution.

The Beck Anxiety Inventory (BAI; Beck et al. 1988) is a 21-item self-report measure of anxiety severity in the past month. Items are rated on a scale of 0 (not at all) to 3 (severely bothered). Higher scores indicate greater anxiety severity. The BAI has demonstrated high internal consistency and test-retest reliability over 1 week, along with adequate concurrent validity (Beck et al. 1988). In the current sample, internal consistency of the BAI was very low ($\alpha = 0.42$). The low internal constancy of the BAI suggests that results derived from this measure should be interpreted with caution.

The Quality of Life Inventory (QOLI; Frisch et al. 1992) is a 32-item self-report measure, separated into 16 domains assessing life satisfaction, well-being, and positive psychology and mental health. Items are rated in terms of importance on a scale from 0 (not at all important) to 2 (extremely important) before being rated in terms of satisfaction on a scale from −3 (very dissatisfied) to 3 (very satisfied). Total scores are calculated by multiplying the importance and satisfaction scores for each domain, omitting domains with a score of zero importance, and averaging the scores from the remaining 16 domains. These raw scores are then converted to t-scores based on normative distributions (Frisch et al. 1992). The QOLI has variable test-retest reliability and internal consistency, and adequate convergent and divergent validity (Frisch

et al. 1992). In the current sample, internal consistency was good ($\alpha = 0.82$).

Procedure After providing informed consent, the participants (called photo raters in this study) completed a demographic questionnaire. Trained research assistants provided instructions and ensured that the photo raters understood the procedure. Photo raters were presented with one of three possible slide sets, with each set containing forty-one PowerPoint slides. A desktop computer with a 15″ Dell 1909 W digital display using 1360 × 768 resolution was used to display photographs, and participants were seated one to two feet away from the screen. Each slide contained a single baseline photo (prior to randomization) and a single post-treatment photo (the last day of the 12 weeks treatment) for one TTM participant, whereby the baseline photo was presented on the left and the post-treatment photo on the right. Photo raters were informed of this order. To minimize sequence effects, the order of photos within each slide set was randomly ordered and participants were randomly assigned to one of three possible slide sets. There was no effect of photo set order on photo rating accuracy ($F(2, 96) = 0.20$, $p = 0.82$).

Photo raters were asked to provide a verbal rating of changes in amount and density of hair. Participants were free to provide a verbal rating at their own pace using a 7-point Likert scale (derived from Kaufman et al. 1998). Scores on this scale ranged from −3 to 3 and used the following anchors: −3 (*greatly decreased*), −2 (*moderately decreased*), −1 (*slightly decreased*), 0 (*no change*), 1 (*slightly increased*), 2 (*moderately increased*), and 3 (*greatly increased).* A research assistant, seated next to the photo rater, recorded each response. "Treatment response" from the perspective of the photo raters was operationalized as a photo set receiving a rating from 1 to 3 on the photo rating measure. "Treatment non-response" was operationalized as a photo set receiving a rating from −3 to 0 on the photo rating measure. CGI-I scores obtained from blinded clinical evaluators at post-treatment during the clinical trial were used to assess criterion treatment response. Scores on the CGI-I of 1 (*greatly improved)* or 2 (*very much improved)* were operationalized as treatment response, while all other scores were operationalized as treatment non-response. Treatment response, as measured by the photo rating measure and the CGI-I, were both dichotomized. Criterion validity was then determined by comparing the agreement of the photo rating measure and the CGI-I.

Results

Reliability Because multiple photo raters and multiple photo sets were used in the current study, we employed the two-way random model intraclass correlation coefficient (ICC) to measure inter-rater reliability (Shrout and Fleiss 1979). According

to Cicchetti (1994), ICC values less than 0.40 reflect poor agreement, values between 0.40 and 0.59 reflect fair agreement, values between 0.60 and 0.74 reflect good agreement, and values above 0.75 reflect excellent agreement. Results produced a single measures ICC value of 0.53 ($F(40, 8400) = 257.62$, $p < 0.000$), reflecting fair reliability.

Validity We found that photo raters were able to correctly predict criterion treatment response (i.e., CGI-I status) an average of 65.55 % of the time (SD = 0.07; Range = 37–80 %). However, in order to determine whether the agreement between photo ratings and CGI-I scores might be inflated by chance agreement, Cohen's Kappa statistics were calculated between each photo rater and TTM individual's CGI-I scores. The mean Cohen's Kappa was 0.35 (SD = 0.14), reflecting fair to poor agreement (Fleiss 1981; Landis and Koch 1977). This suggests that some of the agreement between photo ratings and CGI-I scores might be due to chance, but Kappa statistical analysis collapsed photo ratings from a 7-point scale into a dichotomous scale and reduced variance in participant responses. Thus, we correlated the full 7-point photo rating scale (averaged across raters for each TTM participant) and 7-point CGI-I scale, and found a large association ($r = -0.51$, $p < 0.001$) between high photograph ratings and low scores on the CGI-I (which correspond to better treatment response). This suggests that photographic assessment has at least acceptable criterion validity.

To measure convergent and incremental validity, we computed change scores between baseline and post-treatment for the TTM patients on the primary outcome measures (i.e., MGH-HPS, NIMH-TSS, CGI-S). Results showed that averaged photo ratings and change in these outcome variables had small amounts of shared variance, as evidenced by small, but significant (or as in one case, marginally significant) positive correlations between photo ratings and improvements in MGH-HPS ($r = 0.30$, $p = 0.054$), NIMH-TSS ($r = 0.35$, $p = 0.03$), and CGI-S ($r = 0.36$, $p = 0.02$) scores. This suggests that photographic measures and other indices of TTM severity are sensitive to the same construct, but are not redundant. In order to test whether photo ratings contribute significant unique variance to treatment outcome, hierarchical binary regression analyses were performed using self-report and clinician-rated measures. The self-report measure, the MGH-HPS, was tested separately from the clinician-report measures, to minimize overlapping variance. In the first hierarchical analyses (see Table 1), the MGH-HPS was entered into the model in Step 1 and was an individually significant predictor of treatment response. The photo ratings were entered at Step 2 and resulted in a significant increase in the variance accounted for in the overall model. The second model (see Table 2), using clinician-rated measures of hair pulling severity, produced similar results. In Step 1, the overall model was

Table 1 Hierarchical binary logistic regression predicting treatment response from change in a self-report measure of hair pulling severity and averaged photographic assessments of change

	B	S.E.	X^2	p-value	Odds Ratio
Step 1					
MGH-HPS	0.22	0.08	8.10**	0.004	1.25
Model			11.62**	0.001	
Step 2					
MGH-HPS	0.25	0.10	6.00*	0.01	1.28
Photo Rating	1.88	0.73	6.68*	0.03	6.55
Step			11.82**	0.001	
Model			23.44***	0.000	

*. $p < 0.05$; **. $p < 0.01$; ***. $p < 0.001$

significant, but the NIMH-TSS and CGI-S were only marginally significant individual predictors (possibly because they share significant overlapping variance). When the photo ratings were introduced at Step 2, there was a significant increase in the variance accounted for in the overall model. In both models, the photo ratings were individually significant predictors and had comparatively large odds ratios. There was no evidence of multicollinearity in either model, as tested by the procedure of Mansfield and Helms (1981). Although results using the NIMH-TSS should be interpreted with caution due to the measure's low internal consistency, the fact that the MGH-HPS and CGI-S scales show similar associations to photo ratings bolsters these results.

Because photographic measurement could be viewed as an assessment of the social validity of TTM treatment, we tested whether photo ratings were correlated with measures of depression, anxiety, and quality of life. Improvements in hair growth might be associated with lower TTM-related embarrassment and increased behavioral activation, potentially leading to better overall psychosocial functioning. When

Table 2 Hierarchical binary logistic regression predicting treatment response from change in clinician-rated measures of hair pulling severity and averaged photographic assessments of change

	B	S.E.	X^2	p-value	Odds Ratio
Step 1					
NIMH-TSS	0.26	0.13	3.84	0.05	1.30
CGI-S	0.85	0.44	3.66	0.06	2.33
Model			21.36***	0.000	
Step 2					
NIMH-TSS	0.35	0.18	3.91*	0.048	1.42
CGI-S	0.78	0.51	2.34	0.13	2.18
Photo Rating	2.15	1.02	4.47*	0.04	8.55
Step			8.65**	0.003	
Model			30.01***	0.000	

*. $p < 0.05$; **. $p < 0.01$; ***. $p < 0.001$

comparing photo ratings to change in these variables from baseline to post-treatment, no significant correlations were seen with change in BDI ($r = 0.09$, $p = 0.60$) or BAI ($r = -0.23$, $p = 0.16$), but there was a positive correlation between photo ratings and change in QOLI ($r = 0.42$, $p = 0.01$). These results suggest that changes in hair loss during TTM treatment are associated with improvements in general quality of life, offering some preliminary support that photographic assessment is a useful measure of the social validity of TTM treatment. However, the low internal consistency of the BDI and BAI could have diluted their construct validity, meaning that the lack of association between photo ratings and these measures should be viewed with caution. Further analysis revealed that none of our self-report or clinician-administered TTM severity instruments were significantly correlated with QOLI scores (all p-values ≥ 0.08), and hierarchical linear regression analyses (shown in Tables 3 and 4) showed that photo ratings significantly predicted change in QOLI scores while controlling for measures of TTM and explained significant additional variance in change in quality of life.

Moderating Variables In order to examine whether any characteristics of photo raters or persons with TTM influenced criterion validity (i.e., correct prediction of treatment outcome), data were arranged in two separate formats. To examine characteristics of photo raters, the percentage of "correct" ratings was calculated for each photo rater. Then, one-way ANOVAs were performed using rater characteristics as grouping variables and "percent correct" as the dependent variable. To examine characteristics of TTM participants, the percentage of "correct" ratings was tabulated for each photo set. Thus, linear regressions and one-way ANOVAs were performed using TTM participant characteristics as grouping or predictor variables and "percent correct" as the dependent variable. Because 211 raters and 41 photo sets were used, tests will show varying degrees of freedom.

A significant gender effect for photo rater was found, such that female raters correctly predicted treatment response more often (66.69 %) than male raters (62.69 %) ($F(1, 209) = 9.97$, $p = 0.002$; partial $\eta^2 = 0.05$). No analysis was conducted for the gender of photographed individuals, because only 9.76 % of our photo sample was male, and cell sizes were highly unequal (38 vs. 3). Rating accuracy was not associated with the age of raters (B = 0.003, $p = 0.98$) nor TTM participants (B = 0.002, $p = 0.54$), and there were no differences among the various ethnicities of raters ($F(4, 206) = 0.34$, $p = 0.85$, partial $\eta^2 = 0.006$) or TTM participants ($F(1, 39) = 0.71$, $p = 0.40$, partial $\eta^2 = 0.02$). Results did show a significant difference in rating accuracy among different pulling regions ($F(1, 39) = 4.86$, $p = 0.03$, partial $\eta^2 = 0.11$), with photos of scalp pulling showing more accurate ratings (72.63 %) than eyebrow and eyelash pulling photos (51.90 %).

Table 3 Hierarchical linear regression predicting change in quality of life from change in a self-report TTM severity instrument and averaged photographic assessments

	β	t	p-value	F-change
Step 1				
MGH-HPS	0.25	1.49	0.15	
R^2		0.06		2.23
Step 2				
MGH-HPS	0.25	0.89	0.38	
Photo Ratings	0.38	2.32*	0.03	
R^2		0.20		5.37*

*. $p < 0.05$

Discussion

The current study investigated the psychometric properties of photographic measurement of change in TTM. Results indicate that photographic assessment of change has fair inter-rater reliability, acceptable criterion validity, good concurrent validity, and excellent incremental validity. However, the problematic internal consistencies of the NIMH-TSS, BDI, and BAI could have hindered the validity of some results. The positive correlation between photo ratings and improvements in quality of life suggests that photographic assessment might be a socially valid measure of TTM treatment outcome. Gender of the photo rater and the site of pulling being evaluated were found to significantly moderate rating accuracy.

Our results suggest that photographic assessment of change has adequate psychometric properties and would be a valuable component of a multi-modal TTM assessment strategy. Indeed, it seems that while changes in existing self- and clinician-rated TTM severity scales do predict treatment outcome, photographic assessment added significant incremental information in predicting treatment outcome. This suggests

Table 4 Hierarchical linear regression predicting change in quality of life from change in clinician-rated TTM severity instruments and averaged photographic assessments

	β	t	p-value	F-change
Step 1				
NIMH-TSS	0.03	0.16	0.88	
CGI-S	0.29	1.33	0.19	
R^2		0.09		1.62
Step 2				
NIMH-TSS	-0.05	-0.25	0.81	
CGI-S	0.19	0.91	0.37	
Photo Ratings	0.38	2.15*	0.04	
R^2		0.22		4.61*

*. $p < 0.05$

that photographs could be an important part of TTM assessment batteries.

Results of the current study seem to confirm the notion that persons with TTM who respond to treatment tend to show hair regrowth. While the degree of hair loss captured in a single photograph might not be associated with TTM severity at that time point, as was shown in the Haaga et al. (2015) study, it does appear that change in TTM severity can be detected by a series of photographs. This lends credibility to clinicians who use "before" and "after" photos to document their clients' change in hair pulling over treatment. However, it should be noted that the strengths of correlations between photographic measures and change in traditional measures of TTM severity were in the moderate range, meaning that there is substantial non-overlapping variance between self- and clinician-report scales and photographic measures.

Despite these limitations, it does appear that photographic assessment of change in TTM offers additional information about hair pulling severity and is sensitive to additional psychosocial aspects of the disorder. Photographic measures significantly predicted additional variance in treatment response when entered alongside traditional measures of TTM severity, and hair re-growth was found to predict greater quality of life, whereas changes in TTM severity measures were not associated with changes in quality of life. As such, photographic assessment of change could be a very useful measure during TTM treatment.

Nevertheless, it might be too early to suggest that noticeable changes in hair loss necessarily create universally positive effects, and several important questions remain to be answered. Increases in hair regrowth lead to self-reported increases in quality of life, but future research should more closely investigate whether hair regrowth leads to more positive peer acceptability. Research indicates that hair-pulling is viewed negatively by peers (Boudjouk et al. 2008; Long et al. 1999; Marcks et al. 2005; Ricketts et al. 2012; Woods et al. 1999), and it remains to be seen if the stigma against TTM is ameliorated once one undergoes successful treatment. Additionally, future research should test whether increases in hair growth are truly unrelated to changes in comorbid psychiatric symptoms, or if improvements in such symptoms are simply delayed. Results of the current study showed that photo ratings did not correlate with 12 weeks improvements in depression and anxiety, but perhaps hair regrowth is only the beginning of the process that ultimately leads to a reduction of depressive and anxious symptoms. Perhaps hair-regrowth at 12 weeks relates better to long-term changes in anxiety and depression.

Although the overall validity of photographic assessment was shown to be good to excellent, the moderate reliability of photographic assessment might require some improvement. Several variables that might hinder the reliability of photographic assessment of TTM include characteristics of raters and stimuli, such as gender and pulling site, which were found to affect rating accuracy. Because gender of the rater was found to influence rating accuracy, it stands to reason that some raters are more attuned to subtle changes in hair regrowth than others. Thus, clinicians who are familiar with TTM might be much more valid and reliable raters of changes in hair loss that occur during treatment. This is a limitation to the current study, and future research should examine whether reliability and validity coefficients are higher for clinicians than laypersons. Furthermore, it is likely that additional individual differences in hair characteristics (e.g., color, thickness) and dermatological health might influence the detectability of meaningful differences in hair growth after brief interventions. Likewise, it stands to reason that other pulling sites might be more difficult to assess via photographic means. The scalp, eyebrows, and eyelashes are the most common pulling sites (Woods et al. 2006), but our finding that photographs of change in eyebrow and eyelash pulling sites were not rated accurately (e.g., 51.9 %) suggests that scalp photos might be the best suited pulling site for photographic assessment. As pulling can occur in any part of the anatomy where hair grows (e.g., pubic, axillary, and abdominal regions), this has negative implications for the broader use of photographs in TTM. Moreover, pulling also frequently occurs from multiple pulling sites (Franklin and Tolin 2007; Woods et al. 2006), and pulling from one site might improve while another worsens, meaning that a comprehensive photographic assessment would likely entail documenting all pulling areas over time. It is currently unknown if pulling in more private areas or multiple areas is similarly documentable via photographic means, and it stands to reason that many clients might be uncomfortable with this type of assessment.

Variations in the method through which photographic assessments were taken (e.g., angle, lighting) is another limitation with the procedure, and may contribute to the inconsistency in ratings. There are currently no formal standardization procedures for photographing hair loss in TTM. Given the limitations to current unstandardized practices in TTM photography, perhaps TTM researchers should collaborate with dermatological assessment experts to develop a photographic assessment guideline. Once standardization procedures are in place, this approach should receive another psychometric evaluation.

In conclusion, the present study represents the first effort at elucidating the psychometric properties of photographic measures of change in TTM, as well as quantifying the relationship between visible hair regrowth and treatment outcome. Still, there is a paucity of data regarding the experience of recovering from TTM as well as other similar conditions. Future research should examine the relationship between changes in the visible effects of other body-focused repetitive behavior disorders and treatment outcomes. For conditions related to TTM, such as chronic skin picking and nail biting,

it is currently unclear whether photographic measures would serve as a viable assessment strategy. However, perhaps similar assessment methods, such as the measurement of pock marks or nail length might fulfill a similar role.

Compliance with Ethical Standards

Conflict of Interest   David C. Houghton, Maddison R. Franklin, Michael P. Twohig, Martin E. Franklin, Scott N. Compton, Angela M. Neal-Barnett, Stephen M. Saunders and Douglas W. Woods declare that they have no conflict of interest.

Experiment Participants   All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*, 313–327.

Beck, A. T. (1972). Measuring depression: the depression inventory. In T. A. Williams, M. M. Katz, & J. A. Shields (Eds.), *Recent advances in the psychobiology of the depressive illnesses* (pp. 299–302). Washington DC: U.S. Government Printing Office.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology, 56*, 893–897.

Boudjouk, P., Woods, D., Miltenberger, R., & Long, E. (2008). Negative peer evaluation in adolescents: effects of tic disorders and trichotillomania. *Child and Family Behavior Therapy, 22*(1), 17–28.

Casati, J., Toner, B., & Yu, B. (2000). Psychosocial issues for women with trichotillomania. *Comprehensive Psychiatry, 41*(5), 344–351.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cohen, J. L. (2010). Enhancing the growth of natural eyelashes: the mechanism of Bimatoprost-induced eyelash growth. *Dermatologic Surgery, 36*, 1361–1371.

Dahlke, F., Lohaus, A., & Gutzman, H. (1992). Reliability and clinical concepts underlying global judgments in dementia: implications for clinical research. *Psychopharmacology Bulletin, 28*(4), 425–432.

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*(3), 169–178.

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*(4), 353–363.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Franklin, M. E., & Tolin, D. F. (2007). *Treating trichotillomania: cognitive-behavioral therapy for hairpulling and related problems*. New York, NY: Springer.

Friman, P. C., Finney, J. W., & Christopherson, E. R. (1984). Behavioral treatment of trichotillomania: an evaluative review. *Behavior Therapy, 15*(3), 249–265.

Frisch, M. B., Cornell, J., Villanueva, M., & Retzlaff, P. J. (1992). Clinical validation of the quality of life inventory: a measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological Assessment, 4*(1), 92–101.

Grant, J. E., Odlaug, B. L., & Kim, S. W. (2009). N-acetylcysteine, a glutamate modulator, in the treatment of trichotillomania: a double-blind, placebo-controlled study. *Archives of General Psychiatry, 66*(7), 756–763.

Guy, W. (1976). Clinical global impression scale. *The ECDEU Assessment Manual for Psychopharmacology-Revised, 76*(338), 218–222. Volume DHEW Publ No ADM.

Haaga, D. A. F., Falkenstein, M. J., Banis, M., Malloy, E. J., McDonough, L., & Rogers, K. (2015). Reliability and validity of a photo rating measure of alopecia. *Journal of Psychopathology and Behavioral Assessment, 37*(4), 673–678.

Helleström, Å., & Teckle, J. (1994). Person perception through facial photographs: effects of glasses, hair, and beard on judgments of occupation and personal qualities. *European Journal of Social Psychology, 24*(6), 693–705.

Hunter, M. (2007). The persistent problem of colorism: skin tone, status, and inequality. *Sociology Compass, 1*(1), 237–254.

Kaufman, K. D., Olsen, E. A., Whiting, D., Savin, R., DeVillez, R., Bergfeld, W., et al. (1998). Finasteride in the treatment of men with androgenetic alopecia. *Journal of the American Academy of Dermatology, 39*(4), 578–589.

Keuthen, N. J., O'Sullivan, R. L., Ricciardi, J. N., Shera, D., Savage, C. R., Borgmann, A. S., et al. (1995). The Massachusetts general hospital (MGH) hairpulling scale: 1. Development and factor analysis. *Psychotherapy and Psychosomatics, 64*(3–4), 141–145.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Leon, A. C., Shear, M. K., Klerman, G. L., Portera, L., Rosenbaum, J. F., & Goldenberg, I. (1993). A comparison of symptom determinants of patient and clinician global ratings in patients with panic disorder and depression. *Journal of Clinical Psychopharmacology, 13*(5), 327–331.

Long, E., Woods, D., Miltenberger, R., Fuqua, W., & Boudjouk, P. (1999). Examining the social effects of habit behaviors exhibited by individuals with mental retardation. *Journal of Developmental and Physical Disabilities, 11*(4), 295–312.

Mansfield, E. R., & Helms, B. P. (1981). Detecting multicollinearity. *The American Statistician, 36*(3a), 158–160.

Marcks, B., Woods, D., & Ridosko, J. (2005). The effects of trichotillomania disclosure on peer perceptions and social acceptability. *Body Image, 2*(3), 299–306.

McKelvie, S. J. (1993). Stereotyping in perception of attractiveness, age, and gender in schematic faces. *Social Behavior and Personality: An International Journal, 21*(2), 121–128.

Myers, R. J., & Hamilton, J. B. (1951). Regeneration and rate of growth of hairs in man. *Annals of the New York Academy of Sciences, 53*(3), 562–568.

Ninan, P. T., Rothbaum, B. O., Marstellar, F. A., Knight, B. T., & Eccard, M. B. (2000). A placebo-controlled trial of cognitive-behavioral therapy and clomipramine in trichotillomania. *Journal of Clinical Psychiatry, 61*(1), 47–50.

O'Sullivan, R. L., Keuthen, N. J., Hayday, C. F., Ricciardi, J. N., Buttolph, M. L., Jenike, M. A., et al. (1995). The Massachusetts general hospital (MGH) hairpulling scale: 2. Reliability and validity. *Psychotherapy and Psychosomatics, 64*(3–4), 146–148.

Ricketts, E., Brandt, B., & Woods, D. (2012). The effects of severity and causal explanation on social perception of hair loss. *Journal of Obsessive-Compulsive and Related Disorders, 1*(4), 336–343.

Rothbaum, B. O., & Ninan, P. T. (1994). The assessment of trichotillomania. *Behaviour Research and Therapy, 32*(6), 651–662.

Sadrô, J., Jarudi, I., & Sinhaô, P. (2003). The role of eyebrows in recognition. *Perception, 32*(3), 285–293.

Schosser, R. H., & Kendrick, J. P. (1987). Dermatologic photography. *Dermatologic Clinics, 5*(2), 445–461.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Slue, W. E. (1989). Photographic cures for dermatologic disorders. *Archives of Dermatology, 125*(7), 960–962.

Soriano, J., O'Sullivan, R., Baer, L., Phillips, K., McNally, R., & Jenike, M. (1996). Trichotillomania and self-esteem: a survey of 62 female hair pullers. *Journal of Clinical Psychiatry, 57*(2), 77–82.

Sprinkle, S. D., Lurie, D., Insko, S. L., Atkinson, G., Jones, G. L., Logan, A. R., et al. (2002). Criterion validity, severity cut scores, and test-retest reliability of the beck depression inventory-II in a university counseling center sample. *Journal of Counseling Psychology, 49*(3), 381–385.

Stemberger, R., Thomas, A., Mansueto, C., & Carter, J. (2000). Personal toll of trichotillomania: behavioral and interpersonal sequelae. *Journal of Anxiety Disorders, 14*(1), 97–104.

Swedo, S. E., Leonard, H. L., Rapoport, J. L., Lenane, M. C., Goldberger, E. L., & Cheslow, D. L. (1989). A double-blind comparison of clomipramine and desipramine in the treatment of trichotillomania (hair pulling). *New England Journal of Medicine, 321*(8), 497–501.

Tolin, D. F., Franklin, M. E., Diefenbach, G. J., & Gross, A. (2002). *Cognitive-behavioral therapy for pediatric trichotilomania: an open trial. Paper presented at the association for advancement of behavior therapy.* NV: Reno.

Twohig, M. P., & Woods, D. W. (2004). A preliminary investigation of acceptance and commitment therapy and habit reversal as a treatment for trichotillomania. *Behavior Therapy, 35*(4), 803–820.

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*(2), 97–109.

Wolf, N. (1991). *The beauty myth*. New York: Anchor/Doubleday.

Woods, D. W., Fuqua, W., & Outman, R. (1999). Evaluating the social acceptability of persons with habit disorders: the effects of topography, frequency, and gender manipulation. *Journal of Psychopathology and Behavioral Assessment, 21*, 1–18.

Woods, D. W., Flessner, C. A., Franklin, M. E., Keuthen, N. J., Goodwin, R. D., Stein, D. J., et al. (2006). The trichotillomania impact project (TIP): exploring phenomenology, functional impact, and treatment utilization. *Journal of Clinical Psychiatry, 67*(12), 1877–1888.