Stöber, J. (1998). Reliability and validity of two widely-used worry questionnaires: Self-report and self-peer convergence. *Personality and Individual Differences, 24,* 887-890.

Reliability and Validity of Two Widely-Used Worry Questionnaires: Self-Report and Self-Peer Convergence

Joachim Stöber*
Department of Psychology, Free University of Berlin, Germany

Summary—The reliability and validity of the Penn State Worry Questionnaire (PSWQ) and the Worry Domains Questionnaire (WDQ) were examined with self-ratings from a non-clinical sample of 148 students in a test-retest design across four weeks. Ratings from three well-acquainted peers were also obtained. With internal consistencies and test-retest correlations of at least 0.85, the present study confirmed the high reliability of the questionnaires. Moreover, both measures demonstrated substantial convergent validity: Average agreement among peers was 0.42 (PSWQ) and 0.47 (WDQ), and aggregated self-peer agreement was 0.55 (PSWQ) and 0.49 (WDQ). Self-peer agreement was not biased by social desirability. These findings challenge views that worry is an unreliable and unobservable phenomenon.

Keywords
anxiety, anxiety neurosis, measurement, questionnaires, reliability, validity, observers

* Correspondence should be addressed to Dr. Joachim Stöber who is now at the Pennsylvania State University, Department of Psychology, 417 Bruce V. Moore Building, University Park, PA 16802-3104, USA. Electronic mail may be sent via the Internet to jstoeber@psu.edu.

Reliability and Validity of Two Widely-Used Worry Questionnaires: Self-Report and Self-Peer Convergence

With the advent of the DSM-III-R (American Psychiatric Association, 1987), worry was established as a diagnostic criterion of generalized anxiety disorder. This was followed by increased research on both nonpathological and pathological worry (cf. Davey & Tallis, 1994, for a review) from which two classes of worry measures emerged: "content-free" and "content-based" measures.

From research on pathological worry came mostly content-free measures that assess the excessiveness, duration, and uncontrollability of worry and associated stress. Here, the most widely-used measure is the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger & Borkovec, 1990) listing 16 dysfunctional characteristics of worry (e.g. "I am always worrying about something"). Respondents indicate how typical these characteristics are on a five-point scale from Not at all typical of me (1) to Very typical of me (5).

From research on nonpathological worry came mostly content-based measures that present a list of potential worries and ask the respondents for intensity or frequency ratings. Here, the most widely-used measure is the Worry Domains Questionnaire (WDQ; Tallis, Eysenck & Mathews, 1992). The prefix "I worry..." is followed by a list of 25 worries (e.g. "that I will lose close friends") that cover five worry domains. With this, the WDQ has five subscales: Relationships, Lack of Confidence, Aimless Future, Work Incompetence, and Financial. For each item, respondents indicate how much they worry on a five-point scale from Not at all (0) to Extremely (4).

With respect to reliability, the PSWQ has shown an average Cronbach's alpha of 0.91 and, across intervals from two to ten weeks, an average test-retest correlation of 0.84 (Molina & Borkovec, 1994; Stöber, 1995). With an average Cronbach's alpha of 0.91, the WDQ has also demonstrated high internal consistency (Davey, 1993; Joormann & Stöber, in press; Stöber, 1995). Regarding WDQ stability, there is only one study with a small N of 16 (Tallis, Davey & Bond, 1994) showing a test-retest correlation of 0.79 across four weeks (0.46 to 0.86 for the WDQ subscales). With N = 16, however, these figures are highly unreliable: The 95% confidence interval for 0.79 ranges from 0.48 to 0.92, and the 95% confidence intervals for the subscales' test-retest correlations range from –0.04 to 0.95.

With respect to convergent validity, both questionnaires have shown substantial correlations with other self-report worry measures (Davey, 1993; Molina & Borkovec, 1994; Stöber, 1995; Stöber & Joormann, 1997): With the single-item measure "percentage of time spent worrying on a typical day" (Borkovec, Robinson, Pruzinsky & DePree, 1983), the PSWQ has shown an average correlation of 0.62 and the WDQ a single correlation of 0.44. With the Student Worry Scale (Davey, Hampton, Farrell & Davidson, 1992), average correlations were 0.55 (PSWQ) and 0.65 (WDQ). Finally, PSWQ and WDQ themselves have shown an average correlation of 0.63.

However, in establishing the validity of a psychological construct, it clearly is important to consider non-self-report data as well (Cronbach & Meehl, 1955). To

substantiate the validity of self-report trait measures, peer ratings have become a common means (McCrae, 1994). Because both worry questionnaires assess trait-like worry, peer ratings may also be appropriate for a further validation of PSWQ and WDQ.

However, self-peer agreement of worry could be moderated by social desirability (SD). Previous research has shown small to moderate negative correlations between measures of SD and worry (McCann, Stewin & Short, 1991; Stöber, 1995) suggesting that high-SD participants may possibly underreport worry—and thus provide less valid self-ratings. For self-reported neuroticism, this was demonstrated by Borkenau and Ostendorf (1992): Low-SD participants showed significant self-peer agreement on neuroticism (average correlation 0.30) whereas high-SD participants did not (average correlation 0.17). Because worry and neuroticism are highly correlated (e.g. Wells, 1994), social desirability may also moderate self-peer agreement on worry.

In sum, the aim of the present study was to provide additional information on the psychometric properties of the PSWQ and WDQ, particularly with respect to stability (WDQ) and to convergent validity using peer ratings (PSWQ and WDQ) while controlling for social desirability.

Method

Participants

A complete set of test, retest, and peer ratings was obtained from a sample of 148 participants (87 female), the majority of whom (80.3%) were psychology students at the Free University of Berlin. Mean age was 26.9 years (SD = 7.0). All participants volunteered for the experiment in exchange of three hours of extra course credit.

Peer Raters

Each participant was asked to collect three peer ratings. This resulted in a sample of 444 peer raters (247 female). Mean age was 28.9 years (SD = 9.3). Overall, the peers appeared to be well-acquainted with the participants: 78% reported knowing the participant "for some years" and 19% for "some months"; 86% talked with the participant "daily" or "once a week"; 23% reported that the participant asked them for help/advice "often" and 45% "sometimes"; 92% reported that the participant told them confidential/private things (20% "very often", 43% "often", and 29% "sometimes"); 41% talked with the participant about problems/worries "often" and 36% "sometimes."

Measures

At the test session and at the retest session, first the PSWQ and then the WDQ were administered (German versions by Stöber, 1995). Additionally, the Social Desirability Scale (Crowne & Marlowe, 1960; German version by Lück & Timaeus, 1969) was administered at the test session. All sessions were held individually.

For the peer ratings, the PSWQ and the WDQ were adapted. For the PSWQ, all items were modified (e.g. "He/she is always worrying about something") and the answer scale was changed to range from Not at all typical of him/her (1) to Very typical of him/her (5). For the WDQ, the prefix was changed to "The target person

worries..." and all items were modified accordingly (e.g. "that he/she will lose close friends"), whereas the answer scale remained unchanged.

Procedure

At the first session, participants were informed that the aim of the study was to investigate "how assessable private thoughts—like, for example, worries—are for others." Therefore, they were required not only to provide self-ratings, but also to ask three peers (good friends/acquaintances) for ratings. These peers had to answer the same questionnaires from an observer perspective, put them in an envelope, seal it, and return it to the participants. To achieve unbiased peer ratings, both written instructions and additional verbal instructions stressed (a) that the peer ratings had to be made without any assistance from or discussion with the peers and (b) that all data would be treated highly confidential. Participants were asked to make an appointment for the second session (the retest session) as soon as they received the envelopes from the three peers. On average, the retest session was held four weeks later.

Results

Self-Ratings

Means and standard deviations of PSWQ, WDQ, and WDQ subscale scores (Table 1) were comparable to previous figures from non-clinical samples (Molina & Borkovec, 1994; Stöber, 1995; Tallis et al., 1994). Also the 0.68 correlation ($p < 0.001$)* between PSWQ and WDQ scores was in line with previous findings.

Internal consistency (Cronbach's alpha) was excellent for PSWQ and WDQ scores and was acceptable to good for WDQ subscale scores (Table 1). Test-retest reliability across four weeks was well above 0.80 for PSWQ, WDQ, and WDQ subscale scores (except for Work Incompetence).

Peer Ratings

First, agreement across peer raters was examined by calculating intraclass correlations, ICCs (Shrout & Fleiss, 1979, Case 1; equivalent to the average correlation between all possible pairs of peers). Agreement among raters was substantial for PSWQ, WDQ and WDQ subscales scores (Table 1). With pair-wise correlations in the 0.40s, it was comparable to agreement for personality-trait ratings gathered from well-acquainted peers (cf. McCrae & Costa, 1987). Only Work Incompetence and Financial showed ICCs below 0.40.

Second, self-peer agreement was examined. The correlations between self-ratings and aggregated peer ratings were 0.55 (PSWQ) and 0.49 (WDQ) indicating substantial self-peer agreement for the worry measures. The same held also for the WDQ subscales, again with the exception of Work Incompetence.

Social Desirability

As in previous studies (McCann et al., 1991; Stöber, 1995), PSWQ and WDQ showed significant negative correlations with social desirability (SD), namely –0.22, $p < 0.01$ (PSWQ) and –0.35, $p < 0.001$ (WDQ). In line with Borkenau and Ostendorf (1992), the sample was split at the median of SD scores. Contrary to expectations, the self-peer agreement for low-SD participants was not different from that for high-

*Throughout this article, p values are from one-tailed tests.

SD participants (PSWQ: 0.52 vs. 0.58; WDQ: 0.47 vs. 0.52; both difference-Zs ≤ 0.53, ns). While high SD was related to lower reported worry, it did not attenuate the validity of the self-reports.

Conclusions

With a test-retest correlation of 0.85, the WDQ displayed high stability. Moreover, this figure is highly reliable because, with a sample size of N = 148, the 95% confidence interval for 0.85 ranges only from 0.80 to 0.89. Also the WDQ subscales showed test-retest correlations of at least 0.80 (except for Work Incompetence). Thus, Tallis et al.'s (1994) figures underestimated stability. The present figures show that the WDQ subscales are not only internally consistent measures (cf. Joormann & Stöber, in press) but also stable measures of the facets of nonpathological worry.

Furthermore, both worry measures showed substantial agreement both among peer ratings and between self- and peer ratings. Self-peer agreement for the PSWQ (0.55) and the WDQ (0.49) matched, or even surpassed, validity coefficients for widely-used personality measures (McCrae & Costa, 1987) or trait-like emotion ratings (Watson & Clark, 1991). Moreover, self-peer agreement was not moderated by social desirability. Validity coefficients were high also for the WDQ subscales, except for Work Incompetence. The low reliability/validity of Work Incompetence may have been related to variance restriction: Of all subscales, this subscale had the highest mean and the lowest standard deviation (Table 1). Most students seem to worry about their work, with little individual variance.

When worry research began to establish, O'Neill (1985a, 1985b) questioned that worry was a valuable concept because he rendered worry unobservable and poorly-defined. The present findings, however, stand in marked contrast to O'Neill's judgment. Worry is unlikely to be unobservable and poorly-defined given that there is substantial agreement across persons about the degree to which worry is present in themselves and in others. Consequently, there must be some "observables" in the experience of worry. The precise (verbal or nonverbal) indicators of worry and how they are evaluated and weighted to form an observer's judgment remain interesting topics for future research.

References

American Psychiatric Association (Ed.). (1987). Diagnostic and statistical manual of mental disorders. Third edition - revised (DSM-III-R). Washington, DC: Author.

Borkenau, P. & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. European Journal of Personality, 6, 199-214.

Borkovec, T. D., Robinson, E., Pruzinsky, T. & DePree, J. A. (1983). Preliminary exploration of worry: Some characteristics and processes. Behaviour Research and Therapy, 21, 9-16.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354.

Davey, G. C. L. (1993). A comparison of three worry questionnaires. Behaviour Research and Therapy, 31, 51-56.

Davey, G. C. L., Hampton, J., Farrell, J. & Davidson, S. (1992). Some characteristics of worrying: Evidence for worrying and anxiety as separate constructs. Personality and Individual Differences, 13, 133-147.

Davey, G. C. L. & Tallis, F. (Eds.). (1994). Worrying. Perspectives on theory, assessment, and treatment. New York: Wiley.

Joormann, J. & Stöber, J. (in press). Measuring facets of worry: A LISREL analysis of the Worry Domains Questionnaire. Personality and Individual Differences.

Lück, H. E. & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM) [Scales for the measurement of manifest anxiety (MAS) and social desirability (SDS-E and SDS-CM)]. Diagnostica, 15, 134-141.

McCann, S. J., Stewin, L. L. & Short, R. H. (1991). Sex differences, social desirability, masculinity, and the tendency to worry. Journal of Genetic Psychology, 152, 295-310.

McCrae, R. R. (1994). The counterpoint of personality assessment: Self-reports and observer ratings. Assessment, 1, 159-172.

McCrae, R. R. & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 52, 81-90.

Meyer, T. J., Miller, M. L., Metzger, R. L. & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. Behaviour Research and Therapy, 28, 487-495.

Molina, S. & Borkovec, T. D. (1994). The Penn State Worry Questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), Worrying. Perspectives on theory, assessment, and treatment (pp. 265-283). New York: Wiley.

O'Neill, G. W. (1985a). Is worry a valuable concept? Behaviour Research and Therapy, 23, 479-480.

O'Neill, G. W. (1985b). Response to Dr Borkovec. Behaviour Research and Therapy, 23, 483.

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

Stöber, J. (1995). Besorgnis: Ein Vergleich dreier Inventare zur Erfassung allgemeiner Sorgen [Worry: A comparison of three questionnaires for the measurement of general worries]. Zeitschrift für Differentielle und Diagnostische Psychologie, 16, 50-63.

Stöber, J. & Joormann, J. (1997). Differentiating worry from anxiety and dysphoria: Specific characteristics of high-worriers. Unpublished manuscript, Free University of Berlin, Germany.

Tallis, F., Davey, G. C. L. & Bond, A. (1994). The Worry Domains Questionnaire. In G. C. L. Davey & F. Tallis (Eds.), Worrying. Perspectives on theory, assessment, and treatment (pp. 285-297). New York: Wiley.

Tallis, F., Eysenck, M. W. & Mathews, A. (1992). A questionnaire for the measurement of nonpathological worry. Personality and Individual Differences, 13, 161-168.

Watson, D. & Clark, L. A. (1991). Self- versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. Journal of Personality and Social Psychology, 60, 927-940.

Wells, A. (1994). A multi-dimensional measure of worry: Development and preliminary validation of the Anxious Thoughts Inventory. Anxiety, Stress, and Coping, 6, 289-299.

Acknowledgments

Preparation of this article was supported in part by German Research Foundation (DFG) grant STO 350/1-1. I want to thank Iris Penner for help with the data collection as well as Tom Borkovec, Alexandra Freund, Pete Gianaros, Bärbel Knäuper, and two anonymous reviewers for helpful comments and suggestions on earlier versions of this article.

Table 1

Descriptive Statistics, Reliability, and Cross-Rater Correlations

Measure	M	(SD)	range	α	r_{tt}[a]	ICC[a]	r_{sp}[a]
PSWQ	43.78	(10.09)	24–71	0.89	0.87	0.42	0.55
WDQ	28.99	(15.13)	3–68	0.91	0.85	0.47	0.49
Rel	4.97	(3.98)	0–16	0.76	0.81	0.48	0.45
L of C	6.47	(4.66)	0–17	0.88	0.86	0.47	0.52
Aim Fut	5.82	(3.85)	0–16	0.72	0.80	0.42	0.49
Work Inc	6.83	(3.64)	0–18	0.73	0.71	0.34	0.32
Fin	4.91	(4.01)	0–17	0.82	0.81	0.38	0.53

Note. N = 148. PSWQ = Penn State Worry Questionnaire, total score; WDQ = Worry Domains Questionnaire, total score. WDQ subscales: Rel = Relationships, L of C = Lack of Confidence, Aim Fut = Aimless Future, Work Inc = Work Incompetence, Fin = Financial. α = Cronbach's alpha, r_{tt} = test-retest correlation, ICC = intraclass correlation (average correlation between pairs of peer ratings), r_{sp} = correlation between self-ratings and averaged peer ratings.

[a]All correlations are significant at $p < 0.001$.