Supplemental Materials

Ong, D. C., Zaki, J., & Gruber, J. (2017). Increased cooperative behavior across remitted bipolar I disorder and major depression: Insights utilizing a behavioral economic trust game. *Journal of Abnormal Psychology, 126(1)*, 1-7. doi: 10.1037/xge0000252.

All code and data can be found at our public Github repository at:

https://github.com/desmond-ong/CooperationInMoodDisorders

Additional Details for Interrater Reliability Procedure in Ong et al. (2017)

We provide some additional details for interested readers beyond the typical description of inter-rater reliability (IRR) procedures common in the field for mood-disordered studies (e.g., Gruber & Weinstock, 2018). First, the raters were two clinical psychology graduate students and one licensed clinical psychology faculty member. The interviewer was a full-time post-baccalaureate lab manager. Raters received training in diagnostic criteria and clinical rating measures, practiced with mock cases with the rating system, and engaged role-playing. Once raters achieved a minimum threshold for clinical competency by completing this training they conducted IRR ratings for the study.

Second, raters were assigned non-overlapping participants to code and performed ratings independently (i.e., did not look at the interviewer's scores beforehand). 75 of the 90 participants (90%) were coded by an independent rater (i.e., one rater coded approximately $n = 45$, a second rater coded approximately $n = 30$, and third rater (i.e., study P.I.) coded a small number (approximately $n = 2$) at the beginning of the study).

Third, raters met to discuss discrepancies, correct errors that arose, and solidify their training during informal consensus meetings. The meetings further ensured quality

control, maintain fidelity, and prevent rater drift. Raters were instructed not to change scores due to honest differences of opinion. If a member was not present then that individual consulted with the PI after the meeting. The meetings occurred primarily at the end of data collection (given the P.I. switched institutions during data collection, and soon after data collection was completed raters at the new institution coded interviews).

Fourth, IRR ratings reported in Ong et al. (2017) reflect corrected or post-consensus scores. This is within normal standards of accepted practice in bipolar research (e.g., Weinstock & Gruber, 2018) and has been used previously by this research group (e.g., Cohen et al., 2017; Dutra et al., 2014; 2015; 2017; Ford et al., 2014; Ford, Mauss & Gruber, 2015; Gilbert & Gruber, 2014; Kang & Gruber, 2013; Gruber et al., 2013a; 2013b; 2016; Purcell, Phillips, & Gruber, 2013; though see Hay et al. (2015) that did not hold informal consensus meetings given a more advanced independent rater performed ratings independently and IRR values reflect pre-consensus meeting scores). Both IRR approaches from this research group typically yield "almost perfect" (i.e., above .90) levels of agreement (e.g., McHugh, 2012), suggesting good diagnostic accuracy regardless. We also note that the IRR ratings in Ong et al. were conducted after the main behavioral outcome analyses as a descriptive measure of reliability for the clinical measures (i.e., and were independent from the main analyses).

Fifth, we additionally provide pre-consensus scores for interested readers. Pre-consensus scores matched the reported post-consensus scores in Ong et al. (2017) for primary diagnoses (K =1.00) as well as symptoms of mania (YMRS; ICC= 0.90) and depression (IDS-C; ICC= 0.99). Consensus meetings hence had little impact on the degree of reliability on these measures, perhaps not surprisingly given the "skip out" nature of the SCID known to increase reliability (e.g., Gotlib et al., 2004; Talbot et al

2012) and restricted symptom variability in the remitted bipolar and unipolar samples recruited for this study. Pre-consensus scores for GAF were lower (ICC=0.36) than post-consensus scores for the GAF (ICC=0.65). This is consistent with previously documented low to moderate reliability for GAF scores (e.g., Rey et al., 2005; Vatnaland, Vatnaland, Friis, & Opjordsmoen, 2006) and suggests that consensus meetings provided divergent utility for our clinical diagnosis and symptom measures versus more historically unreliable measures such as the GAF.

Related, we note that the consensus meetings were useful in detecting scoring issues for the interviewer as well (e.g., misapplication of GAF scoring to lifetime instead of using the correct timeframe leading to somewhat lower scores). As the clinical characteristics reported in Ong et al. (2017) reflected the original interviewer's scores, we include Table 1b below as well that reflects the interviewer's scores after consensus meetings for interested readers (note that these do not affect any of the main results reported).

Table 1b

Clinical Characteristics (Reported using Interviewer Scores After Consensus Meeting)

	rBD (*n*=28)	rMDD (*n*=30)	CTL (*n*=27)	Statistic
YMRS	1.32 (1.49)	1.07 (1.20)	0.56 (0.97)	F= 2.70 [a]
IDS-C	3.43 (2.33)	4.77 (3.04)	2.30 (2.05)	F= 6.83 [bc]
GAF	70.46 (7.22)	76.97 (7.11)	85.85 (6.41)	F= 34.08 [abc]

Additional References

Cohen, J. N., Dryman, T., Morrison, A. S., Gilbert, K. E., Heimberg, R. G., & Gruber, J. (2017). Positive and negative affective links between social anxiety and depression: Predicting concurrent and prospective mood symptoms in mood disorders. *Behavior Therapy, 48*, 820-883. doi: 10.1016/j.beth.2017.07.003.

Dutra, S. J., Cunningham, W. A., Kober, H., & Gruber, J. (2015). Elevated striatal reactivity across monetary and social rewards in bipolar I disorder. *Journal of Abnormal Psychology, 124*(4), 890-904. doi: 10.1037/abn0000092.

Dutra, S. J., Man, V., Kober, H., Cunningham, W. A., & Gruber, J. (2017). Disrupted cortico-limbic connectivity during reward processing in bipolar I disorder. *Bipolar Disorders, 19(8)*, 661-675. https://doi.org/10.1111/bdi.12560

Dutra, S. J., Reeves, E., J., Mauss, I. B., & Gruber, J. (2014). Boiling at a different degree: An investigation of trait and state anger in remitted bipolar I disorder. *Journal of Affective Disorders, 168,* 37-43. doi: 10.1016/j.jad.2014.06.044.

Ford, B., Mauss, I. B, & Gruber, J. (2015). Extreme valuing of happiness is associated with risk for and diagnosis of bipolar disorder. *Emotion, 15(2)*, 211-222. doi: 10.1037/emo0000048.

Ford, B., Shallcross, A. J., Mauss, I. B., Floerke, V. A., & Gruber, J. (2014). If you seek it, it won't come: Pursuing happiness is associated with depressive symptoms and diagnosis of depression. *Journal of Social and Clinical Psychology, 33*, 890-905. doi: 10.1521/jscp.2014.33.10.890.

Gilbert, K. E., & Gruber, J. (2014). Emotion regulation of goals in bipolar disorder and major depression: A comparison of rumination and mindfulness. *Cognitive Therapy and Research, 38*, 375-288. doi: 10.1007/s10608-014-9602-3.

Gotlib, I. H., Kasch, K. L., Traill, S., Joormann, J., Arnow, B. A., & Johnson, S. L. (2004). Coherence and specificity of information-processing biases in depression and social phobia. *Journal of Abnormal Psychology*, *113*(3), 386. doi: 10.1037/0021-843x.113.3.386

Gruber, J. & Weinstock, L. M. (2018). Interrater reliability in bipolar disorder research: Current practices and suggestions for enhancing best practices. *International Journal of Bipolar Disorders, 6(1)*, 1-3. https://doi.org/10.1186/s40345-017-0111-7.

Gruber, J., Siegel, E. H., Purcell, A. L., Earls, H. A., Cooper, G., & Feldman Barrett, L. (2016). Unseen positive and negative affective information influences social perception in bipolar I disorder and healthy adults. *Journal of Affective Disorders, 192*, 191-198. doi: 10.1016/j.jad.2015.12.037.

Gruber, J., Purcell, A. L., Perna, M., & Mikels, J. A. (2013a). Letting go of the bad: Deficits in maintaining negative, but not positive, emotion in bipolar disorder. *Emotion, 13(1)*, 168-175. doi: 10.1037/a0029381.

Gruber, J., Kogan, A., Mennin, D., & Murray, G. (2013b). Real-world emotion? An experience-sampling approach to emotion disturbance and regulation in bipolar disorder. *Journal of Abnormal Psychology, 122(4)*, 971-983. doi: 10.1037/a0030262.

Hay, A. C., Sheppes, G., Gross, J. J., & Gruber, J. (2015). Choosing how to feel: Emotion regulation choice in bipolar I disorder. *Emotion, 15(2)*, 139-145. doi: 10.1037/emo0000024.

Kang, Y. & Gruber, J. (2013). Harnessing happiness? Uncontrollable positive emotion in bipolar disorder, major depression, bipolar and healthy adults. *Emotion, 13(2)*, 290-301. doi: 10.1037/a0030780.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemica Medica, 22(3)*, 276-282.

Purcell, A., Phillips, M. L., & Gruber, J. (2013). In your eyes: Does theory of mind predict impaired life functioning in bipolar disorder? *Journal of Affective Disorders, 151*, 1113-1119. doi: 10.1016/j.jad.2013.06.051.

Rey, J. M., Starling, J., Wever, C., Dossetor, D. R., & Plapp, J. M. (1995). Inter-rater reliability of Global Assessment of Functioning in a clinical setting. *Journal of Child Psychology and Psychiatry*, *36*(5), 787-792. doi: 10.1111/j.1469-7610.1995.tb01329.x

Talbot, L. S., Stone, S., Gruber, J., Hairston, I. S., Eidelman, P., & Harvey, A. G. (2012). A test of the bidirectional association between sleep and mood in bipolar disorder and insomnia. *Journal of Abnormal Psychology, 121(1)*, 39-50. doi: 10.1037/a0024946.

Vatnaland, T., Vatnaland, J., Friis, S., & Opjordsmoen, S. (2007). Are GAF scores reliable in routine clinical use?. *Acta Psychiatrica Scandinavica*, *115*(4), 326-330. doi: 10.1111/j.1600-0447.2006.00925.x