Research Article

Psychological Science
1–17

DOI: 10.1177/0956797619831612
www.psychologicalscience.org/PS

How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project

Christopher J. Soto[TC]
Department of Psychology, Colby College

Abstract

The Big Five personality traits have been linked to dozens of life outcomes. However, metascientific research has raised questions about the replicability of behavioral science. The Life Outcomes of Personality Replication (LOOPR) Project was therefore conducted to estimate the replicability of the personality-outcome literature. Specifically, I conducted preregistered, high-powered (median N = 1,504) replications of 78 previously published trait–outcome associations. Overall, 87% of the replication attempts were statistically significant in the expected direction. The replication effects were typically 77% as strong as the corresponding original effects, which represents a significant decline in effect size. The replicability of individual effects was predicted by the effect size and design of the original study, as well as the sample size and statistical power of the replication. These results indicate that the personality-outcome literature provides a reasonably accurate map of trait–outcome associations but also that it stands to benefit from efforts to improve replicability.

Keywords

Big Five, life outcomes, metascience, personality traits, replication, open data, open materials, preregistered

Received 7/6/18; Revision accepted 11/29/18

Do personality characteristics reliably predict consequential life outcomes? A sizable research literature has identified links between the Big Five personality traits and dozens of outcomes (Ozer & Benet-Martinez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). On the basis of this personality-outcome literature, economists, educators, and policymakers have proposed initiatives to promote well-being through positive personality development (Chernyshenko, Kankaraš, & Drasgow, 2018; Kautz, Heckman, Diris, ter Weel, & Borghans, 2014; Organisation for Economic Co-operation and Development, 2015; Primi, Santos, John, & De Fruyt, 2016). However, recent metascientific research has raised questions about the replicability of behavioral science (Button et al., 2013; Camerer et al., 2016; Cova et al., 2018; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009). We therefore conducted the Life Outcomes of Personality Replication (LOOPR) Project, an effort to estimate the replicability of the personality-outcome literature. Specifically, we attempted preregistered, high-powered replications of 78 previously published associations between the Big Five traits and a diverse set of consequential life outcomes.

Personality Traits and Consequential Life Outcomes

A personality trait is a characteristic pattern of thinking, feeling, or behaving that tends to be consistent over time and across relevant situations (Allport, 1961). The world's languages include thousands of adjectives for describing personality, many of which can be organized

Corresponding Author:
Christopher J. Soto, Colby College, Department of Psychology, 5550 Mayflower Hill, Waterville, ME 04901
E-mail: christopher.soto@colby.edu

in terms of the Big Five trait dimensions: Extraversion (e.g., sociable, assertive, and energetic vs. quiet and reserved), Agreeableness (e.g., compassionate, respectful, and trusting vs. rude and suspicious), Conscientiousness (e.g., orderly, hardworking, and responsible vs. disorganized and unreliable), Negative Emotionality (or Neuroticism; e.g., worrying, pessimistic, and temperamental vs. calm and stable), and Open-Mindedness (or Openness to Experience; e.g., intellectual, artistic, and imaginative vs. incurious and uncreative; De Raad, Perugini, Hrebícková, & Szarota, 1998; Goldberg, 1993; John, Naumann, & Soto, 2008).

The Big Five constitute the most widely used framework for conceptualizing and measuring personality traits (Almlund, Duckworth, Heckman, & Kautz, 2011; John et al., 2008). This scientific consensus reflects their usefulness for organizing personality-descriptive language, as well as a substantial research literature linking the Big Five with life outcomes. The most comprehensive literature review conducted to date summarized associations between the Big Five and dozens of individual, interpersonal, and social-institutional outcomes (Ozer & Benet-Martinez, 2006). For example, research has linked high Extraversion to social status and leadership capacity, Agreeableness to volunteerism and relationship satisfaction, Conscientiousness to job performance and health, Negative Emotionality to relationship conflict and psychopathology, and Open-Mindedness to spirituality and political liberalism.

The Replicability of Behavioral Science

Drawing on both conceptual and empirical evidence, recent metascientific research (i.e., the scientific study of science itself) has raised questions about the replicability of behavioral science—the likelihood that independent researchers conducting similar studies will obtain similar results. Conceptually, this work has focused on researcher degrees of freedom, statistical power, and publication bias. Researcher degrees of freedom represent undisclosed flexibility in the design, analysis, and reporting of a scientific study (Simmons et al., 2011). Statistical power is the probability of obtaining a statistically significant result when the effect being tested truly exists in the population (Cohen, 1988). Publication bias occurs when journals selectively publish studies with statistically significant results, thereby producing a literature that underrepresents null results (Sterling, Rosenbaum, & Weinkam, 1995). Multiple observers have expressed concern that much behavioral science is characterized by many researcher degrees of freedom, modest statistical power, and strong publication bias, leading to the publication of numerous false-positive results—that is, statistical flukes that are unlikely to be replicable (Fraley & Vazire, 2014; Franco, Malhotra, & Simonovits, 2014; Rossi, 1990; Simmons et al., 2011; Sterling et al., 1995; Tversky & Kahneman, 1971).

Recently, large-scale replication projects have begun to empirically test these concerns. For example, the Reproducibility Project: Psychology (RPP) attempted to replicate 100 studies published in high-impact psychology journals. Despite high statistical power, the RPP observed a replication success rate of only 36% (when success was defined as a statistically significant result in the expected direction) and found that the replication effects were, on average, only half as strong as the original effects (Open Science Collaboration, 2015). Similar projects in economics and experimental philosophy have also obtained replicability estimates considerably lower than would be expected in the absence of published false positives, although results have varied somewhat across projects (Camerer et al., 2016; Cova et al., 2018). These findings reinforce concerns about the replicability of behavioral science and suggest that replicability may vary both between and within disciplines. For example, replicability appears to be higher for original studies that (a) examined main effects rather than interactions, (b) reported intuitive rather than surprising results, and (c) obtained a greater effect size, sample size, and strength of evidence (Camerer et al., 2016; Cova et al., 2018; Open Science Collaboration, 2015).

The LOOPR Project

In sum, previous research suggests that the Big Five personality traits relate to many consequential life outcomes, but it also raises questions about the replicability of behavioral science. We therefore conducted the LOOPR Project to estimate the replicability of the personality-outcome literature. Specifically, we attempted to replicate 78 previously published trait–outcome associations and then used the replication results to test two descriptive hypotheses. First, we hypothesized that trait–outcome associations would be less than perfectly replicable because of the likelihood of published false positives and biased reporting of effect sizes. Second, we hypothesized that the replicability of the personality-outcome literature may be greater than the estimates obtained by previous large-scale replication projects in psychology because of normative practices in personality research (e.g., using relatively large samples to examine the main effects of personality traits). We also conducted exploratory analyses to search for predictors of replicability, tentatively hypothesizing that original studies with greater effect size, sample size, and strength of evidence, as well as replication attempts with greater

sample size and statistical power, may yield greater replicability.

Method

The LOOPR Project was conducted in six phases, which are briefly described below. An extended description is available in the Supplemental Material available online. Additional materials—including coded lists of the selected trait–outcome associations, original sources, and measures, as well as the final survey materials, preregistration protocol (and revisions), data, and analysis code—are available at https://osf.io/d3xb7. This research was approved by the Colby College Institutional Review Board.

The first phase of the project was to select a set of trait–outcome associations for replication. We selected these from a published review of the personality-outcome literature (Ozer & Benet-Martinez, 2006). Table 1 in Ozer and Benet-Martinez's review summarizes 86 associations between the Big Five traits and 49 life outcomes. A research assistant and I examined the summary table, main text, and citations of this review to identify the empirical evidence supporting each trait–outcome association. We then selected 78 associations, spanning all of the Big Five traits and 48 life outcomes, that could be feasibly replicated. These 78 hypothesized trait–outcome associations served as the LOOPR Project's primary units of analysis for estimating replicability.[1]

The second phase was to code the empirical sources supporting each association so our replication attempts could follow the original studies as closely as was feasible. We therefore coded information about the sample, measures, analytic method, and results of one empirical study or meta-analysis for each of the 78 trait–outcome associations, which resulted in the coding of 38 original sources. Some sources assessed multiple traits, outcomes, suboutcomes, or subsamples; when results differed across these components, we coded each one separately. Appendix A in the Supplemental Material lists citations for the 38 original sources. Detailed coding of the original studies, including information about their samples, measures, and design, is available at https://osf.io/mc3z7.

The third phase was to develop a survey procedure for assessing the Big Five traits and 48 selected life outcomes. We assessed personality using a brief consensus measure of the Big Five: the Big Five Inventory–2 (BFI-2; Soto & John, 2017). This 60-item questionnaire uses short phrases to assess the prototypical facets of each Big Five trait domain. The 48 target life outcomes were assessed using a battery of measures selected to follow the original studies as closely as possible. For most outcomes, this involved administering the same outcome measure used in the original study or a subset of the original measures. For some outcomes, it involved adapting interview items to a questionnaire format or constructing items on the basis of the information available in the original source. To conserve assessment time, we abbreviated lengthy outcome measures to approximately six items per outcome, sampling equally across subscales or content domains to preserve content validity. After developing this assessment battery, we used the Qualtrics platform (https://www.qualtrics.com) to construct two online surveys; each survey included the BFI-2 and approximately half of the outcome measures. Table S1 in the Supplemental Material lists the outcome measures used in the original studies and replications, and Appendix B in the Supplemental Material lists citations for these measures. Detailed coding of the original and replication outcome measures is available at https://osf.io/mc3z7, and the final LOOPR surveys can be viewed at https://osf.io/9nzxa (Survey 1) and https://osf.io/vdb6w (Survey 2).

The fourth phase was data collection. We used the Qualtrics Online Sample service to administer our surveys to four groups of adults (ages 18 and older; used to replicate studies that analyzed adult community samples) and young adults (ages 18–25; used to replicate studies that analyzed student or young-adult samples). This yielded samples of 1,559 adults and 1,550 young adults who completed Survey 1 and samples of 1,512 adults and 1,505 young adults who completed Survey 2. Quota sampling was used to ensure that each sample would be approximately representative of the United States population in terms of sex, race, and ethnicity and that the adult samples would also be representative in terms of age, educational attainment, and household income. Participants were compensated approximately $3 per 25-min survey. A minimum sample size of 1,500 participants per sample was selected to maximize statistical power within our budgetary constraints; this sample size provides power of 97.3% to detect a small true correlation (.10) and greater than 99.9% power to detect a medium (.30) or large (.50) correlation, using two-tailed tests and a .05 significance level (Cohen, 1988).

The fifth phase was preregistration. We registered our hypotheses, design, materials, and planned analyses on the Open Science Framework (see https://osf.io/d3xb7). The preregistration protocol was submitted during data collection and prior to data analysis, thereby minimizing the influence of researcher degrees of freedom.

The final phase was data analysis. Descriptive statistics for all personality and outcome variables are presented in Table S2 in the Supplemental Material. We

Table 1. Summary of the Hypothesized Trait–Outcome Associations and Replication Results

Outcome and expected trait association	Asso-ciation	Number of tests	Replication sample size	Original sample size	Replication success rate	Replication effect size[a]	Original effect size[a]	Effect-size ratio
				Individual outcomes				
Subjective well-being								
Extraversion	+	4	1,559	9,131	100/100	.37/.39	.18	2.18/2.31
Negative Emotionality	–	4	1,559	7,869	100/100	.52/.54	.22	2.64/2.78
Religious beliefs and behavior								
Agreeableness	+	2	1,550	595	100/100	.18/.19	.28	0.63/0.69
Conscientiousness	+	2	1,550	595	100/100	.14/.15	.24	0.60/0.65
Existential or phenomenological concerns: Open-Mindedness	+	2	1,550	595	100/100	.18/.20	.35	0.50/0.56
Existential well-being								
Extraversion	+	1	1,550	595	100/100	.35/.37	.32	1.12/1.18
Negative Emotionality	–	1	1,550	595	100/100	.60/.63	.66	0.87/0.93
Gratitude								
Extraversion	+	1	1,559	1,228	100/100	.37/.37	.32	1.17/1.17
Agreeableness	+	1	1,559	1,228	100/100	.54/.54	.41	1.39/1.39
Forgiveness: Agreeableness	+	1	1,550	140	100/100	.48/.57	.58	0.79/0.97
Inspiration								
Extraversion	+	1	1,514	152	100/100	.39/.39	.20	2.04/2.04
Open-Mindedness	+	1	1,514	152	100/100	.35/.35	.43	0.80/0.80
Humor								
Agreeableness	+	1	1,550	169	100/100	.16/.16	—	—
Negative Emotionality	–	1	1,550	169	100/100	.13/.13	—	—
Heart disease: Agreeableness	–	1	1,235	1,108	0/0	.04/.04	.15	0.24/0.24
Risky behavior: Conscientiousness	–	15	1,336	826	72/72	.08/.08	.26	0.31/0.31
Coping								
Extraversion	+	2	1,505	672	50/100	.17/.19	.16	1.04/1.19
Negative Emotionality	–	2	1,505	672	100/100	.21/.24	.16	1.32/1.50
Resilience: Extraversion	+	1	1,505	138	100/100	0.18/0.18	0.19	0.96/0.96
Substance abuse								
Conscientiousness	–	1	1,505	468	100/100	.06/.06	.25	0.25/0.25
Open-Mindedness	+	1	1,505	468	0/0	.02/.02	.18	0.12/0.12
Anxiety: Negative Emotionality	+	1	1,505	468	100/100	.31/.31	.34	0.90/0.90
Depression								
Extraversion	–	1	1,505	468	100/100	.13/.13	.42	0.28/0.28
Negative Emotionality	+	1	1,505	468	100/100	.31/.31	.46	0.64/0.64
Personality disorders								
Extraversion	±	4	1,505	194	75/75	.30/.41	.43	0.66/0.93
Agreeableness	–	3	1,505	194	100/100	.42/.58	.44	0.95/1.40
Conscientiousness	±	5	1,505	194	100/100	.30/.42	.41	0.71/1.03
Negative Emotionality	±	4	1,505	194	100/100	.31/.41	.38	0.82/1.11
Identity achievement: Conscientiousness	+	1	1,550	198	100/100	.23/.25	.30	0.75/0.83
Identity foreclosure: Open-Mindedness	–	1	1,550	198	100/100	.33/.35	.50	0.63/0.66
Identity integration or consolidation								
Negative Emotionality	–	1	804	111	100/100	0.47/0.57	0.22	2.31/2.86
Open-Mindedness	+	1	804	111	100/100	0.21/0.25	0.27	0.77/0.92

(continued)

Table 1. *(continued)*

Outcome and expected trait association	Asso-ciation	Number of tests	Replication sample size	Original sample size	Replication success rate	Replication effect size[a]	Original effect size[a]	Effect-size ratio
Ethnic-culture identification (for minorities): Conscientiousness	+	1	181	164	100/100	0.18/0.18	0.20	0.91/0.91
Majority-culture identification (for minorities)								
Extraversion	+	1	181	164	0/0	0.10/0.10	0.35	0.28/0.28
Open-Mindedness	+	1	181	164	0/0	0.12/0.12	0.28	0.41/0.41
Interpersonal outcomes								
Family satisfaction								
Conscientiousness	+	2	1,466	980	0/0	−0.07/−0.08	0.11	−0.69/−0.77
Negative Emotionality	−	1	1,489	980	100/100	0.17/0.19	0.10	1.74/1.89
Peers' acceptance and friendship: Extraversion	+	1	1,549	418	100/100	.35/.35	.41	0.84/0.84
Dating variety: Extraversion	+	1	1,284	418	100/100	.12/.12	.17	0.73/0.73
Attractiveness: Extraversion	+	1	1,550	418	100/100	.33/.33	.24	1.39/1.39
Peer status: Extraversion	+	2	775	37	100/100	.39/.39	.41	0.93/0.93
Peer status (men): Negative Emotionality	−	1	749	42	100/100	.31/.31	.43	0.69/0.69
Romantic satisfaction								
Extraversion	+	2	795	210	100/100	.15/.18	.28	0.53/0.63
Negative Emotionality	−	2	795	210	100/100	.20/.23	.32	0.62/0.73
Romantic satisfaction (dating couples)								
Agreeableness	+	1	757	272	100/100	.18/.22	.35	0.51/0.63
Conscientiousness	+	1	757	272	100/100	.16/.19	.35	0.44/0.53
Romantic conflict: Negative Emotionality	+	1	1,154	712	0/0	.01/.01	.32	0.02/0.02
Romantic abuse: Negative Emotionality	+	1	1,154	712	100/100	.09/.09	.25	0.35/0.37
Romantic dissolution: Negative Emotionality	+	1	1,098		100/100	.10/.10	.21	0.45/0.45
Social-institutional outcomes								
Investigative occupational interests: Open-Mindedness	+	1	1,503	725	100/100	.15/.16	.25	0.58/0.63
Artistic occupational interests: Open-Mindedness	+	1	1,503	725	100/100	.41/.43	.30	1.39/1.51
Social occupational interests: Extraversion	+	1	1,503	725	100/100	.15/.17	.16	0.96/1.05
Enterprising occupational interests								
Agreeableness	+	1	1,503	725	100/100	.08/.09	.11	0.77/0.84
Extraversion	+	1	1,503	725	100/100	.18/.20	.16	1.14/1.23
Occupational performance: Conscientiousness	−	3	829	2,058	33/33	.03/.03	.11	0.31/0.31
Occupational satisfaction								
Extraversion	+	1	747	12,023	100/100	.19/.21	.18	1.09/1.17
Negative Emotionality	−	1	747	13,500	100/100	.17/.18	.23	0.72/0.77
Occupational commitment								
Extraversion	+	1	748	492	100/100	.32/.32	.17	1.96/1.96
Negative Emotionality	−	1	748	713	100/100	.26/.26	.19	1.38/1.38

(continued)

Table 1. *(continued)*

Outcome and expected trait association	Association	Number of tests	Replication sample size	Original sample size	Replication success rate	Replication effect size[a]	Original effect size[a]	Effect-size ratio
Extrinsic success								
Agreeableness	–	1	481	194	100/100	0.15/0.15	0.24	0.63/0.63
Conscientiousness	+	1	481	194	0/0	–.07/–.07	.50	–0.13/–0.13
Negative Emotionality	–	1	481	194	100/100	.10/.10	.34	0.28/0.28
Intrinsic success								
Conscientiousness	+	1	512	194	100/100	.24/.25	.20	1.22/1.25
Negative Emotionality	–	1	512	194	100/100	.31/.32	.26	1.20/1.24
Job attainment: Agreeableness	+	1	838	859	0/0	–.02/–.02	.19	-0.09/-0.09
Occupational involvement: Extraversion	+	1	944	859	100/100	.17/.17	.18	0.93/0.95
Financial security: Negative Emotionality	–	1	944	859	100/100	.33/.33	.22	1.52/1.52
Right-wing authoritarianism: Open-Mindedness	–	1	1,549	424	100/100	.29/.32	.35	0.80/0.92
Conservatism								
Conscientiousness	+	1	1,559	93	100/100	.14/.18	.24	0.56/0.75
Open-Mindedness	–	1	1,550	1,648	100/100	.17/.25	.34	0.49/0.74
Volunteerism								
Extraversion	+	1	1,504	796	100/100	.20/.20	.14	1.41/1.41
Agreeableness	+	1	1,504	796	100/100	.17/.17	.23	0.74/0.74
Leadership								
Extraversion	+	1	747	169	100/100	.45/.47	.22	2.16/2.28
Agreeableness	+	1	747	169	100/100	.27/.28	.27	1.00/1.05
Antisocial behavior								
Conscientiousness	–	1	1,550	187	100/100	.26/.29	.28	0.92/1.04
Negative Emotionality	+	1	1,550	187	100/100	.06/.07	.28	0.20/0.23
Criminal behavior								
Agreeableness	–	1	1,550	197	100/100	.23/.23	.20	1.14/1.17
Conscientiousness	–	1	1,550	197	100/100	.18/.19	.31	0.58/0.59

Note: Symbols in the association column indicate whether the hypothesized association was positive (+), negative (–), or both (±). In the columns showing replication success rate, replication effect size, and effect-size ratio, values to the left of the slash represent the observed trait–outcome associations, and values to the right of the slash represent the corrected associations. All effect sizes are oriented so that positive values represent effects in the hypothesized direction. For outcomes that include multiple suboutcomes or subsamples, results are aggregated within each outcome. Mean effect sizes and effect-size ratios were computed using Fisher's *r*-to-*z* transformation.

[a]All effect sizes in these two columns are expressed either as standardized regression coefficients (values with leading zeroes) or correlations (values without leading zeroes). The effect sizes for risky behavior are *r*-transformed averages of *z*-transformed standardized regression coefficients and correlations.

conducted two key sets of planned analyses and one set of exploratory analyses. The first set attempted to replicate each of the 78 hypothesized trait–outcome associations. The second set aggregated the results of these 78 replication attempts to estimate the overall replicability of the personality-outcome literature. We examined replicability in terms of both statistical significance and effect size, using Pearson's *r* (or standardized regression coefficients when the original results could not be converted to *r*s) as our common effect-size metric and using Fisher's *r*-to-*z* transformation to aggregate effects. In the final set of analyses, we searched for predictors of replicability by correlating indicators of replication success with characteristics of the original study and replication attempt.

Results

Testing the hypothesized trait–outcome associations

Did the trait–outcome associations replicate? Our first set of planned analyses attempted to replicate each of the 78 hypothesized associations. For each association, we conducted a preregistered analysis specified to parallel the original study. For outcomes that included multiple suboutcomes or subsamples, we conducted a separate analysis for each component then aggregated these results (e.g., effect size, number of statistically significant results) to the outcome level. For analyses involving outcome measures that had been abbreviated

to conserve assessment time, we computed the observed trait–outcome associations and also estimated the associations that would be expected if the outcome measure had not been abbreviated. Specifically, we used the Spearman-Brown prediction formula and Spearman disattenuation formula to estimate the trait–outcome associations that would be expected if our outcome measure had used the same number of items or indicators as in the original study (Lord & Novick, 1968). These corrected associations address the possibility that some failures to replicate could simply reflect the attenuated reliability and validity of the abbreviated measures.

Table 1 presents the basic results of these analyses, including the number of significance tests conducted for each hypothesized association, the mean sample size, the proportion of tests that were statistically significant (i.e., two-tailed $p < .05$) in the hypothesized direction, the mean original effect size, the mean replication effect size, and the ratio of the replication effect size to the original effect size. To check the robustness of these results to variations in sample size, we calculated the replication success rates that would be expected using different sample sizes (see Table 2): the sample size used in the original study, a sample size 2.5 times as large as in the original study (as recommended by Simonsohn, 2015), and a sample size with 80% power to detect the original effect size (a heuristic that is often used to plan follow-up studies). More detailed information about all of these analyses, including complete results by suboutcome and subsample, is available at https://osf.io/mc3z7.

The results shown in Tables 1 and 2 indicate that many of the 78 replication attempts obtained statistically significant support for the hypothesized associations, with effect sizes comparable to the original results. However, these tables also suggest substantial variability in the results of the replication attempts, in terms of both statistical significance and effect size.

Testing overall replicability

How replicable is the personality-outcome literature overall? Our second set of planned analyses addressed this question by aggregating the results of the 78 replication attempts summarized in Table 1. These analyses compared the results of the LOOPR Project with two benchmarks: (a) the results that would be expected if all of the original findings represented true effects (i.e., if the personality-outcome literature did not include any false-positive results) and (b) the results of the RPP, a previous large-scale replication project conducted to estimate the overall replicability of psychological science (Open Science Collaboration, 2015).[2]

We began by examining the rate of successful replication, defined simply as the proportion of replication attempts that yielded statistically significant results in the hypothesized direction. The results of this analysis are presented in Figure 1. Across the 76 trait–outcome associations with an original effect size available for power analysis, the present research obtained successful replication rates of 87.2% (66.3 successes; 95% confidence interval, or CI = [79.7%, 94.7%]) in tests of the observed associations and 87.9% (66.8 successes; 95% CI = [80.6%, 95.2%]) after partially correcting for the unreliability of abbreviated outcome measures. These success rates were significantly lower than the rate of 99.3% (75.5 successes; 95% CI = [97.4%, 100.0%]) expected from power analyses of the original effect sizes and replication sample sizes—for observed associations, $\chi^2(1, N = 152) = 8.79$, $p = .003$; for corrected associations, $\chi^2(1, N = 152) = 8.23$, $p = .004$. However, they were significantly higher than the success rate of 36.1% (35 successes in 97 attempts; 95% CI = [26.5%, 45.6%]) obtained in the RPP—for observed associations, $\chi^2(1, N = 173) = 45.96$, $p < .001$; for corrected associations, $\chi^2(1, N = 173) = 47.25$, $p < .001$. These significant differences from the RPP also held for the complete set of 78 trait–outcome associations, with success rates of 87.6% (68.3 successes; 95% CI = [80.2%, 94.9%]) for the observed associations and 88.2% (68.8 successes; 95% CI = [81.1%, 95.4%]) for the corrected associations—for observed associations, $\chi^2(1, N = 175) = 47.39$, $p < .001$; for corrected associations, $\chi^2(1, N = 175) = 48.69$, $p < .001$.

The results presented in Table 2 indicate that these findings were also fairly robust to variations in sample size. Specifically, the expected replication success rates would be 80.9% (60.7 successes in 75 attempts; 95% CI = [72.0%, 89.8%]) when using the same sample size as in the original study,[3] 89.1% (66.8 successes in 75 attempts; 95% CI = [82.0%, 96.1%]) when using a sample size 2.5 times as large as the original study, and 59.9% (45.5 successes in 76 attempts; 95% CI = [48.9%, 70.9%]) when using a sample size that provides 80% statistical power to detect the original effect. After we partially corrected them for unreliability, these expected success rates were 80.9% (60.7 successes; 95% CI = [72.0%, 89.8%]), 89.7% (67.3 successes; 95% CI = [82.9%, 96.6%]), and 64.1% (48.7 successes; 95% CI = [53.3%, 74.9%]), respectively. All of these success rates were significantly lower than would be expected from power analyses—all $\chi^2(1, N\text{s} = 150\text{–}174\text{s}) \geq 4.77$, $p \leq .029$—but significantly higher than those obtained in the RPP—all $\chi^2(1, N\text{s} = 150\text{–}174\text{s}) \geq 9.71$, $p \leq .002$.

Next, we examined the frequency with which the replication attempts obtained a trait–outcome association weaker than the corresponding original effect or not in the expected direction. Across the 76 trait–outcome associations with an original effect size available for comparison, the observed replication effect

Table 2. Obtained and Expected Replication Success Rates for Varying Sample Sizes

Outcome and expected trait association	Asso-ciation	Number of tests	Replication success rate			
			Replication sample size	Original sample size	Original sample size × 2.5	Sample size with 80% power
		Individual outcomes				
Subjective well-being						
Extraversion	+	4	100/100	100/100	100/100	100/100
Negative Emotionality	–	4	100/100	100/100	100/100	100/100
Religious beliefs and behavior						
Agreeableness	+	2	100/100	100/100	100/100	50/50
Conscientiousness	+	2	100/100	100/100	100/100	50/50
Existential or phenomenological concerns: Open-Mindedness	+	2	100/100	100/100	100/100	0/0
Existential well-being						
Extraversion	+	1	100/100	100/100	100/100	100/100
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Gratitude						
Extraversion	+	1	100/100	100/100	100/100	100/100
Agreeableness	+	1	100/100	100/100	100/100	100/100
Forgiveness: Agreeableness	+	1	100/100	100/100	100/100	100/100
Inspiration						
Extraversion	+	1	100/100	100/100	100/100	100/100
Open-Mindedness	+	1	100/100	100/100	100/100	100/100
Humor						
Agreeableness	+	1	100/100	100/100	100/100	
Negative Emotionality	–	1	100/100	0/0	100/100	
Heart disease: Agreeableness	–	1	0/0	0/0	0/0	0/0
Risky behavior: Conscientiousness	–	15	72/72	61/61	89/89	33/33
Coping						
Extraversion	+	2	50/100	50/50	50/100	50/50
Negative Emotionality	–	2	100/100	100/100	100/100	100/100
Resilience: Extraversion	+	1	100/100	100/100	100/100	100/100
Substance abuse						
Conscientiousness	–	1	100/100	0/0	100/100	0/0
Open-Mindedness	+	1	0/0	0/0	0/0	0/0
Anxiety: Negative Emotionality	+	1	100/100	100/100	100/100	100/100
Depression						
Extraversion	–	1	100/100	100/100	100/100	0/0
Negative Emotionality	+	1	100/100	100/100	100/100	0/0
Personality disorders						
Extraversion	±	4	75/75	75/75	75/75	25/75
Agreeableness	–	3	100/100	100/100	100/100	100/100
Conscientiousness	±	5	100/100	100/100	100/100	60/80
Negative Emotionality	±	4	100/100	100/100	100/100	50/100
Identity achievement: Conscientiousness	+	1	100/100	100/100	100/100	100/100
Identity foreclosure: Open-Mindedness	–	1	100/100	100/100	100/100	0/0
Identity integration or consolidation						
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Open-Mindedness	+	1	100/100	100/100	100/100	100/100
Ethnic-culture identification (for minorities): Conscientiousness	+	1	100/100	100/100	100/100	100/100
Majority-culture identification (for minorities)						
Extraversion	+	1	0/0	0/0	100/100	0/0
Open-Mindedness	+	1	0/0	0/0	100/100	0/0

(continued)

Table 2. *(continued)*

Outcome and expected trait association	Asso-ciation	Number of tests	Replication success rate: Replication sample size	Original sample size	Original sample size × 2.5	Sample size with 80% power
Interpersonal outcomes						
Family satisfaction						
Conscientiousness	+	2	0/0	0/0	0/0	0/0
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Peers' acceptance and friendship: Extraversion	+	1	100/100	100/100	100/100	100/100
Dating variety: Extraversion	+	1	100/100	100/100	100/100	100/100
Attractiveness: Extraversion	+	1	100/100	100/100	100/100	100/100
Peer status: Extraversion	+	2	100/100	100/100	100/100	100/100
Peer status (men): Negative Emotionality	–	1	100/100	100/100	100/100	0/0
Romantic satisfaction						
Extraversion	+	2	100/100	100/100	100/100	50/50
Negative Emotionality	–	2	100/100	50/50	100/100	50/50
Romantic satisfaction (dating couples)						
Agreeableness	+	1	100/100	100/100	100/100	0/0
Conscientiousness	+	1	100/100	100/100	100/100	0/0
Romantic conflict: Negative Emotionality	+	1	0/0	0/0	0/0	0/0
Romantic abuse: Negative Emotionality	+	1	100/100	100/100	100/100	0/0
Romantic dissolution: Negative Emotionality	+	1	100/100			0/0
Social-institutional outcomes						
Investigative occupational interests: Open-Mindedness	+	1	100/100	100/100	100/100	0/0
Artistic occupational interests: Open-Mindedness	+	1	100/100	100/100	100/100	100/100
Social occupational interests						
Extraversion	+	1	100/100	100/100	100/100	100/100
Agreeableness	+	1	100/100	100/100	100/100	100/100
Enterprising occupational interests: Extraversion	+	1	100/100	100/100	100/100	100/100
Occupational performance: Conscientiousness	–	3	33/33	33/33	67/67	33/33
Occupational satisfaction						
Extraversion	+	1	100/100	100/100	100/100	100/100
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Occupational commitment						
Extraversion	+	1	100/100	100/100	100/100	100/100
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Extrinsic success						
Agreeableness	–	1	100/100	100/100	100/100	0/0
Conscientiousness	+	1	0/0	0/0	0/0	0/0
Negative Emotionality	–	1	100/100	0/0	100/100	0/0
Intrinsic success						
Conscientiousness	+	1	100/100	100/100	100/100	100/100
Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Job attainment: Agreeableness	+	1	0/0	0/0	0/0	0/0
Occupational involvement: Extraversion	+	1	100/100	100/100	100/100	100/100
Financial security: Negative Emotionality	–	1	100/100	100/100	100/100	100/100
Right-wing authoritarianism: Open-Mindedness	–	1	100/100	100/100	100/100	100/100
Conservatism						
Conscientiousness	+	1	100/100	0/0	100/100	0/100
Open-Mindedness	–	1	100/100	100/100	100/100	0/100
Volunteerism						
Extraversion	+	1	100/100	100/100	100/100	100/100
Agreeableness	+	1	100/100	100/100	100/100	100/100

(continued)

Table 2. *(continued)*

Outcome and expected trait association	Association	Number of tests	Replication success rate: Replication sample size	Original sample size	Original sample size × 2.5	Sample size with 80% power
Leadership						
Extraversion	+	1	100/100	100/100	100/100	100/100
Agreeableness	+	1	100/100	100/100	100/100	100/100
Antisocial behavior						
Conscientiousness	–	1	100/100	100/100	100/100	100/100
Negative Emotionality	+	1	100/100	0/0	0/0	0/0
Criminal behavior						
Agreeableness	–	1	100/100	100/100	100/100	100/100
Conscientiousness	–	1	100/100	100/100	100/100	0/0

Note: Symbols in the association column indicate whether the hypothesized association was positive (+), negative (–), or both (±). For replication success rates, separate columns show (from left to right) the sample size in the replication study (see Table 1), the sample size in the original study, a sample size 2.5 times as large as in the original study (Simonsohn, 2015), and the sample size required to provide 80% statistical power to detect the original effect size. Cells in which required information was not available from the original study have been left blank. Values to the left of the slash represent the observed trait–outcome associations, and values to the right of the slash represent the corrected associations. For outcomes that include multiple suboutcomes or subsamples, results are aggregated within each outcome.

was weaker than the original effect 71.1% of the time (54 cases; 95% CI = [60.9%, 81.2%]); after we ran analyses partially correcting for the unreliability of abbreviated outcome measures, the rate was 63.2% (48 cases; 95% CI = [52.3%, 74.0%]). Binomial tests indicated that both of these rates were significantly higher than the 50% rate that would be expected if all of the original effect sizes represented true effects (for observed associations, $p < .001$; for corrected associations, $p = .029$). However, Fisher's exact tests indicated that the rate of weaker replication effects obtained in the present research was less than the corresponding rate of 82.8% (82 of 99 cases; 95% CI = [75.4%, 90.3%]) obtained in the RPP and that this difference was significant after

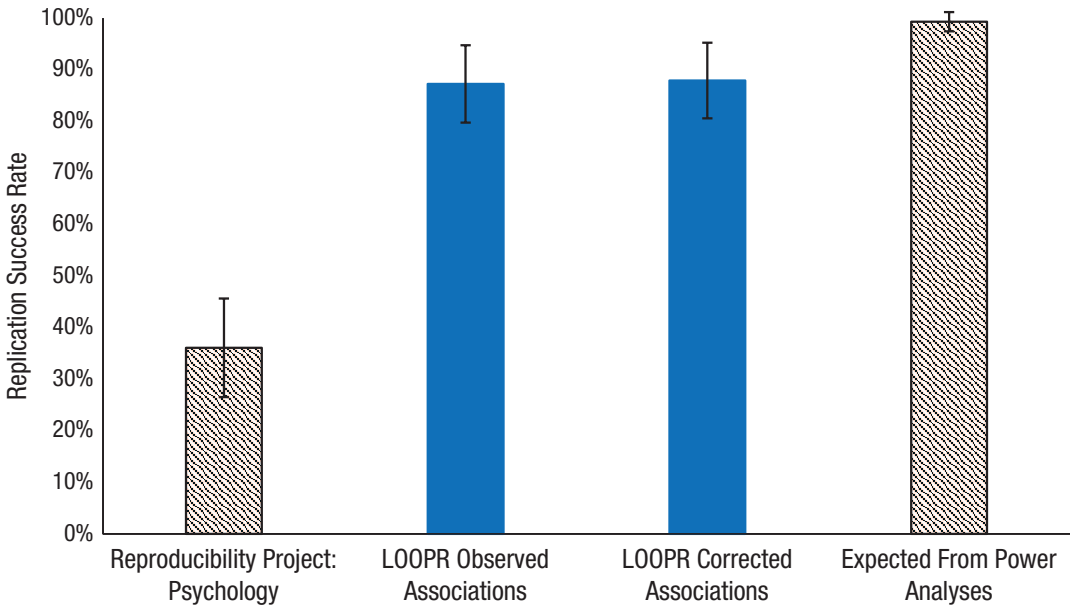

Fig. 1. Replication success rates obtained in the Life Outcomes of Personality Replication (LOOPR) Project, compared with the rate expected from power analyses of the original effect size and replication sample size and with the rate obtained in the Reproducibility Project: Psychology. A successful replication was defined as a statistically significant effect (i.e., two-tailed $p < .05$) in the hypothesized direction. Corrected associations were partially disattenuated to correct for the unreliability of abbreviated outcome measures. Error bars represent 95% confidence intervals.

correcting for unreliability (for observed associations, $p = .070$; for corrected associations, $p = .005$).

Focusing on cases in which the observed replication effect was either not in the expected direction or was substantially weaker than the original effect (i.e., the z-transformed replication effect was at least .10 less than the transformed original effect; Cohen, 1988) yielded a similar pattern of results. In the present research, the observed replication effect was substantially weaker than the original effect 42.1% of the time (32 of 76 cases; 95% CI = [31.0%, 53.2%]); after we ran analyses correcting for unreliability, the rate was 30.3% (23 of 76 cases; 95% CI = [19.9%, 40.6%]). Fisher's exact tests indicated that both of these rates were significantly lower than the corresponding rate of 69.1% (67 of 97 cases; 95% CI = [59.9%, 78.3%]) obtained in the RPP (for observed associations, $p = .001$; for corrected associations, $p < .001$).

Finally, we tested whether the mean and median of the z-transformed replication effect sizes differed from the transformed original effect sizes and whether the median effect-size ratio (i.e., the ratio of the replication effect size to the original effect size) differed between the present research and the RPP. Paired-samples t tests indicated that the mean original effect size of .29 (95% CI = [.26, .32]) was significantly stronger than both the mean observed replication effect of .23 (95% CI = [.20, .27], $t(75) = 3.46$, $p = .001$) and the mean corrected replication effect of .26 (95% CI = [.22, .29]), $t(75) = 2.06$, $p = .043$. Similarly, Wilcoxon signed-rank tests indicated that the median original effect of .27 (95% CI = [.23, .31]) was significantly stronger than the median observed replication effect of .19 (95% CI = [.17, .26], $z = 3.59$, $p < .001$) and the median corrected replication effect of .22 (95% CI = [.18, .27], $z = 2.40$, $p = .016$). However, Mann-Whitney U tests indicated that the median effect-size ratios of .77 (95% CI = [.63, .92]) for observed trait–outcome associations and .87 (95% CI = [.73, .97]) for corrected associations obtained in the present research were both significantly greater than the corresponding median ratio of .43 (95% CI = [.28, .62]) obtained in the RPP (for observed effects, $z = 4.22$, $p < .001$; for corrected effects, $z = 4.86$, $p < .001$). The results of this analysis, presented in Figure 2, indicate that the replication effects obtained in the LOOPR Project were typically about 80% as large as the corresponding original effects.

Taken together, these results support our hypothesis that the personality-outcome literature is less replicable than would be expected if it did not include any false-positive results but more replicable than the broader set of psychology studies examined by the RPP. This conclusion held whether replicability was assessed in terms of statistical significance or effect size.

Predictors of replicability

What factors might influence the replicability of a trait–outcome association? In our final, exploratory set of analyses, we searched for predictors of replicability.

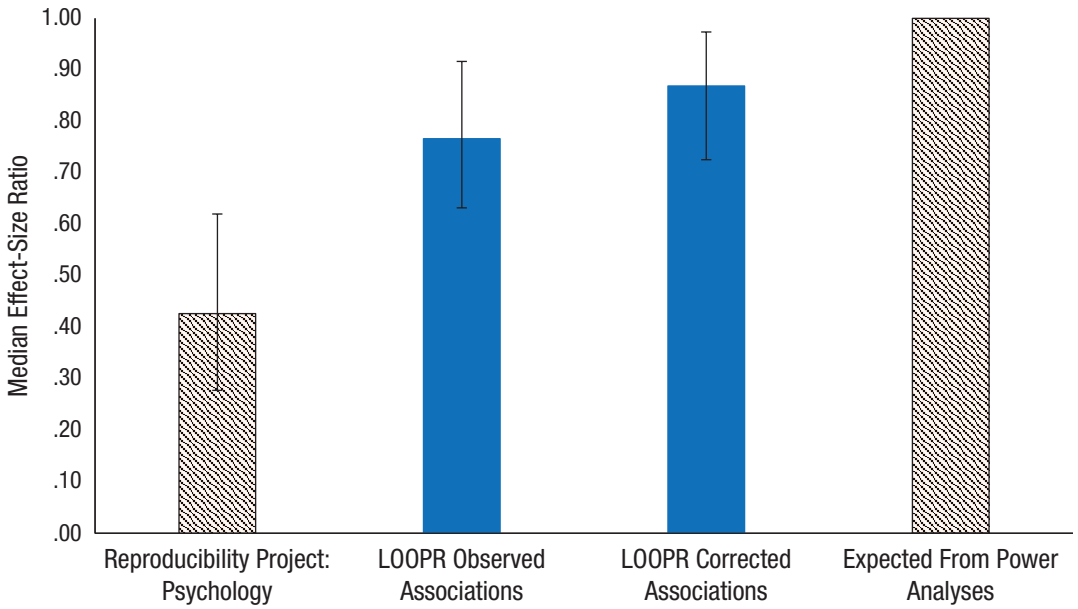

Fig. 2. Median effect-size ratios obtained in the Life Outcomes of Personality Replication (LOOPR) Project, compared with the ratio expected if all original effect sizes represented true effects and with the median ratio obtained in the Reproducibility Project: Psychology. Effect-size ratios were computed as the ratio of the z-transformed replication effect size to the transformed original effect size. Corrected associations were partially disattenuated to correct for the unreliability of abbreviated outcome measures. Error bars represent 95% confidence intervals.

Specifically, we computed Spearman's rank correlations (ρs) across the set of 78 hypothesized trait–outcome associations to correlate three characteristics of the original studies (effect size, sample size, and obtained *p* value[4]), two characteristics of the replication attempts (sample size and statistical power to detect the original effect), and three aspects of similarity between the original study and the replication attempt (whether the outcome was measured using the same indicators, the same data source, and the same assessment timeline) with five indicators of replicability (statistical significance of the replication effect, replication effect size, whether the replication effect was stronger than the original effect, whether the replication effect was not substantially weaker than the original effect, and ratio of the replication effect size to the original effect size).

These correlations, presented in Table 3, suggest three noteworthy patterns. First, the original effect size positively predicted the replication effect size—for observed effects, ρ(74) = .34, 95% CI = [.12, .53], p = .002; for corrected effects, ρ(74) = .39, 95% CI = [.18, .58], p < .001. The original effect size also negatively predicted the likelihood that the replication effect would be stronger than the original effect—for observed effects, ρ(74) = −.40, 95% CI = [−.58, −.18], p < .001; for corrected effects, ρ(74) = −.30, 95% CI = [−.49, −.07], p = .009—as well as the likelihood that the replication effect would not be substantially weaker than the original effect—for observed effects, ρ(74) = −.40, 95% CI = [−.58, −.18], p < .001; for corrected effects, ρ(74) = −.22, 95% CI = [−.43, .01], p = .053. This pattern, illustrated in Figure 3, indicates that strong original effects were more likely to yield strong replication effects but also provided more room for the replication effect to be weaker than the original effect.

The second noteworthy pattern was that the likelihood of successful replication (i.e., a statistically significant effect in the hypothesized direction) was positively predicted by the statistical power and sample size of the replication attempt—statistical power: for observed effects, ρ(74) = .37, 95% CI = [.15, .55], p = .001; for corrected effects, ρ(74) = .33, 95% CI = [.11, .52], p = .003; sample size: for observed effects, ρ(76) = .25, 95% CI = [.03, .45], p = .026; for corrected effects, ρ(76) = .27, 95% CI = [.05, .47], p = .017. This pattern likely reflects the influence of sample size on statistical significance, especially when one attempts to detect small effects.

The final pattern was that the replication effect size and the effect-size ratio were both positively predicted by whether the original study and the replication measured the target outcome using the same items or indicators, as well as the same data source and format (i.e., a self-report questionnaire; all ρs ≥ .19, all ps ≤ .107; see Table 3 for 95% CIs). This pattern, although weaker and less consistent than the previous two, indicates that replications using assessment methods more similar to those employed in the original studies tended to obtain trait–outcome associations that were somewhat stronger and more comparable with the original effects.

Taken together, the results presented in Table 3 and Figure 3 suggest that the predictors of replicability vary depending on how replicability is indexed: Original effect size was the best predictor of replication effect size, whereas replication power and sample size were the best predictors of statistical significance. However, the conclusions that can be drawn from these results should be tempered by the limited variability of some predictors (e.g., replication sample size and statistical power were generally quite high) and some replicability indicators (e.g., relatively few replication effects were not statistically significant).

Discussion

The LOOPR Project was conducted to estimate the replicability of the personality-outcome literature by attempting preregistered, high-powered replications of 78 previously published trait–outcome associations. When replicability was defined in terms of statistical significance, we successfully replicated 87% of the hypothesized effects, or 88% after partially correcting for the unreliability of abbreviated outcome measures. A replication effect was typically 77% as strong as the corresponding original effect, or 87% after we corrected for unreliability. Moreover, the statistical significance of a replication attempt was best predicted by the sample size and statistical power of the replication, whereas the strength of a replication effect was best predicted by the original effect size.

These results can be interpreted either optimistically or pessimistically. An optimistic interpretation is that replicability estimates of 77% to 88% (across statistical-significance and effect-size criteria) are fairly high. These findings suggest that the extant personality-outcome literature provides a reasonably accurate map of how the Big Five traits relate with consequential life outcomes (Ozer & Benet-Martinez, 2006). In contrast, a pessimistic interpretation is that our replicability estimates are lower than would be expected if all the originally published findings were unbiased estimates of true effects. This suggests that the personality-outcome literature includes some false-positive results and that reported effect sizes may be inflated by researcher degrees of freedom and publication bias. Thus, personality psychology—like other areas of behavioral science—stands to benefit from efforts to improve replicability by

Table 3. Predictors of Replicability Across the 78 Hypothesized Trait–Outcome Associations

	Replication success		Replication effect size		Replication effect stronger		Replication effect not substantially weaker		Effect-size ratio	
Value	Observed	Corrected	Observed	Corrected	Observed	Corrected	Observed	Corrected	Observed	Corrected
Original-study characteristics										
Effect size	.12	.07	.34**	.39***	−.40***	−.30**	−.40***	−.22	−.26*	−.22
	[−.11, .33]	[−.15, .29]	[.12, .53]	[.18, .58]	[−.58, −.18]	[−.49, −.07]	[−.58, −.18]	[−.43, .01]	[−.46, −.03]	[−.43, .01]
Sample size	−.13	−.12	−.17	−.18	.26*	.14	.10	.05	.05	.04
	[−.35, .10]	[−.33, .11]	[−.38, .06]	[−.39, .05]	[.03, .46]	[−.09, .36]	[−.13, .32]	[−.18, .28]	[−.18, .27]	[−.19, .26]
p	.09	.08	.02	−.01	.07	−.02	.14	.09	.16	.13
	[−.14, .31]	[−.15, .30]	[−.21, .25]	[−.24, .22]	[−.16, .29]	[−.25, .21]	[−.10, .35]	[−.14, .31]	[−.07, .37]	[−.10, .35]
Replication characteristics										
Sample size	.25*	.27*	.29*	.31**	.07	.09	.02	.20	.14	.17
	[.03, .45]	[.05, .47]	[.06, .48]	[.08, .50]	[−.16, .29]	[−.14, .31]	[−.21, .24]	[−.03, .41]	[−.09, .35]	[−.06, .39]
Statistical power	.37**	.33**	.27*	.30**	−.21	−.09	−.12	.03	−.05	−.03
	[.15, .55]	[.11, .52]	[.05, .47]	[.08, .50]	[−.41, .02]	[−.31, .14]	[−.34, .11]	[−.20, .25]	[−.28, .17]	[−.26, .20]
Similarity between original study and replication										
Outcome indicators	−.04	−.05	.24*	.19	.14	.08	.23*	.17	.24*	.19
	[−.26, .19]	[−.27, .18]	[.01, .44]	[−.03, .40]	[−.09, .36]	[−.14, .30]	[.00, .43]	[−.06, .38]	[.02, .45]	[−.04, .40]
Outcome data source	.16	.19	.25*	.29*	.09	.11	.06	.27*	.19	.23*
	[−.07, .37]	[−.04, .40]	[.03, .45]	[.06, .48]	[−.13, .31]	[−.12, .33]	[−.17, .28]	[.05, .47]	[−.04, .40]	[.00, .44]
Assessment timeline	.03	.04	−.04	.00	−.16	−.06	−.20	−.07	−.19	−.16
	[−.20, .25]	[−.18, .26]	[−.26, .18]	[−.23, .22]	[−.37, .07]	[−.28, .17]	[−.41, .03]	[−.29, .16]	[−.40, .03]	[−.38, .07]

Note: *N*s = 73–78. The table shows Spearman's rank correlations; 95% confidence intervals are given in brackets. *Replication success* means that the replication effect was statistically significant in the hypothesized direction, *replication effect stronger* means that the replication effect was in the hypothesized direction and stronger than the original effect, and *replication effect not substantially weaker* means that the replication effect was in the hypothesized direction and not substantially weaker than the corresponding original effect (i.e., Cohen's $q > -.10$). *Effect-size ratio* is the ratio of the *z*-transformed replication effect to the transformed original effect. Results are shown separately for analyses of observed trait–outcome associations ("observed" columns) and analyses of trait–outcome associations partially corrected for the unreliability of abbreviated outcome measures ("corrected" columns). For the rows showing the similarity between the original study and the replication, *outcome indicators* refers to whether the original study and replication used the same items or indicators to measure the outcome (1 = both used the same indicators; .5 = replication used a subset of the original indicators; 0 = replication used different indicators), *outcome data source* refers to whether the original study and replication used the same data source and format to measure the outcome (1 = both used self-report questionnaire data; .5 = original study used either self-report or questionnaire data; 0 = original study used neither self-report nor questionnaire data), and *assessment timeline* refers to whether the original study and replication used the same timeline to assess the trait and outcome (1 = both used concurrent assessment of the trait and outcome; .5 = original study aggregated results from concurrent and nonconcurrent assessments; 0 = original study did not assess the trait and outcome concurrently).
*$p < .05$. **$p < .01$. ***$p < .001$.

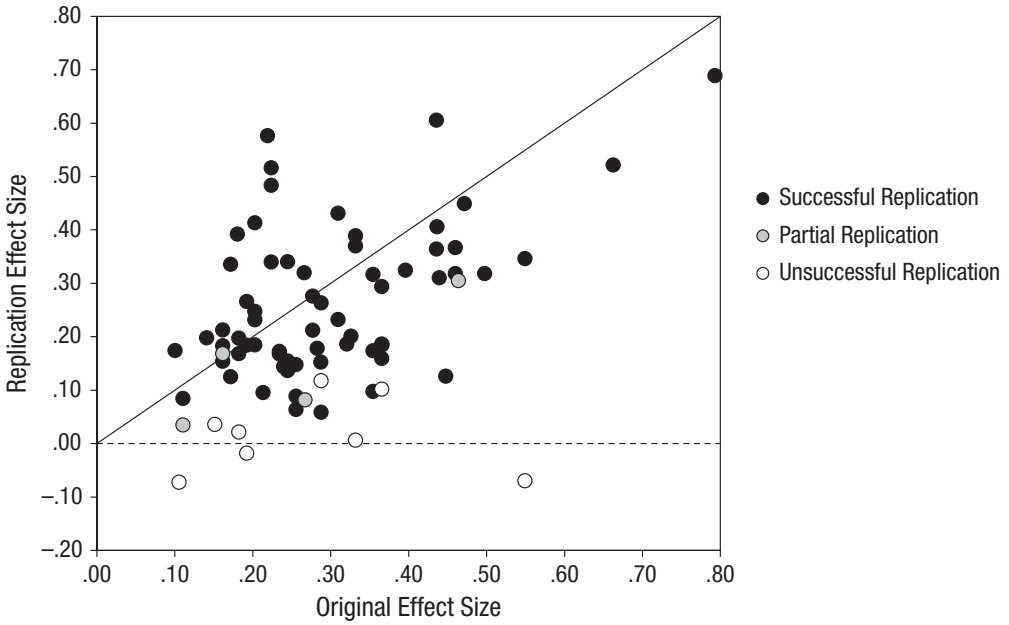

Fig. 3. Scatterplot showing the association between the *z*-transformed original and (observed) replication effect sizes, by success of the replication attempt. *Successful replication* means that the replication effect was statistically significant in the hypothesized direction, *unsuccessful replication* means that the replication effect was not statistically significant or was not in the hypothesized direction, and *partial replication* means that replication was successful for some suboutcomes or subsamples but not for others. The solid diagonal line represents replication effect sizes equal to the original effect sizes, the dashed horizontal line represents a replication effect size of 0, and points below the dashed line represent replication effects that were not in the hypothesized direction.

constraining researcher degrees of freedom, increasing statistical power, and reducing publication bias. Taken together, these interpretations leave us cautiously optimistic about the current state and future prospects of the personality-outcome literature (cf. Nelson, Simmons, & Simonsohn, 2018).

Compared with previous large-scale replication projects in the behavioral sciences, the LOOPR Project obtained relatively high replicability estimates. Why was this? One likely contributor to our high success rates when evaluating replicability in terms of statistical significance was the large sample size (median N = 1,504) and the correspondingly high statistical power (median > 99.9%) of the replication attempts. When evaluating replicability in terms of relative effect size, we speculate that the relatively high estimates obtained here may reflect methodological norms in personality-outcome research, which typically examines the main effects of traits using (a) samples of several hundred participants and (b) standardized measures (Fraley & Vazire, 2014; Open Science Collaboration, 2015; Simmons et al., 2011). However, we note that comparisons between replication projects should be tempered by the fact that different projects have used different approaches to select the original studies and design the replication attempts. Additional research is clearly needed to further investigate variation in replicability across scientific disciplines and research literatures.

The present findings also have implications for understanding why replication attempts in the behavioral sciences might generally succeed or fail. Failures to replicate are sometimes attributed to unmeasured moderators—subtle differences between the original study and the replication attempt that cause an effect to be observed in the former but not the latter (e.g., Stroebe & Strack, 2014). In the LOOPR Project, there were unavoidable differences between the original studies and the replication attempts in terms of historical context (original studies conducted from the 1980s to 2000s vs. replication in 2017), local context (many original research sites vs. national American samples), sampling method (mostly student or community samples vs. survey panels), administration method (mostly in-person surveys or interviews vs. online surveys), and personality measures (many original measures vs. the BFI-2). The relatively high replicability estimates obtained despite these differences converge with previous results suggesting that unmeasured moderators are not generally powerful enough to explain many failures to replicate (Ebersole et al., 2016; Klein et al., 2014).

Strengths, limitations, and future directions

The LOOPR Project had a number of important strengths, including its broad sample of life outcomes, representative samples, preregistered design, and high statistical power. However, it also had some noteworthy limitations that suggest promising directions for future research. Most notably, all of the present data come from cross-sectional, self-report surveys completed by online research panels, whereas some of the original studies used longitudinal designs or other data sources (e.g., interviews, informant reports, community samples). Indeed, our analyses of replicability predictors indicated that replication effect sizes tended to be somewhat stronger when the original study had also used a self-report survey to measure the target outcome. Thus, the present research is only a first step toward establishing the replicability of these trait–outcome associations, and future research using longitudinal designs, as well as alternative sampling and assessment methods, is clearly needed.

A broader issue is that large-scale replication projects can be conducted using different approaches (McShane, Tackett, Bockenholt, & Gelman, 2017). Any particular approach will have advantages and disadvantages, and the choice of an optimal approach will depend on the goals of a particular project. The main goal of the LOOPR Project was to estimate the overall replicability of the personality-outcome literature. We therefore adopted an approach that attempted to replicate a large number of original effects from many studies, with one replication attempt per effect and relatively brief outcome measures (Camerer et al., 2016; Cova et al., 2018; Open Science Collaboration, 2015). An alternative approach would be to replicate a smaller number of effects with lengthier measures or multiple replication attempts per effect (i.e., a many-labs approach; Ebersole et al., 2016; Hagger et al., 2016; Klein et al., 2014). Such an approach would be less well suited for estimating the overall replicability of a literature but better suited for achieving other goals. For example, future research can complement the LOOPR Project by testing individual trait–outcome associations more robustly and by directly investigating factors—such as location, sampling method, mode of administration, measures, and analytic method—that might moderate these associations.

Conclusion

The results of the LOOPR Project provide grounds for cautious optimism about the personality-outcome literature—optimism because we successfully replicated most of the hypothesized trait–outcome associations, with many replication effect sizes comparable with the original effects, and caution because these replicability estimates were lower than would be expected in the absence of published false positives. We therefore conclude that the extant literature provides a reasonably accurate map of how the Big Five personality traits relate to consequential life outcomes but that personality psychology still stands to gain from ongoing efforts to improve the replicability of behavioral science.

Action Editor

Brent W. Roberts served as action editor for this article.

Author Contributions

C. J. Soto is the sole author of this article and is responsible for its content.

ORCID iD

Christopher J. Soto https://orcid.org/0000-0002-2875-8919

Acknowledgments

I thank Alison Russell and Samantha Rizzo for their assistance with this research. The Big Five Inventory–2 is freely available for research use at http://www.colby.edu/psych/personality-lab.

Declaration of Conflicting Interests

C. J. Soto is a copyright holder for the Big Five Inventory–2, which was used in the present research.

Funding

This research was supported by a faculty research grant from Colby College to C. J. Soto.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619831612

Open Practices

TC

All data, materials, and analysis code for this study are publicly available on the Open Science Framework at https://osf.io/d3xb7. The preregistration protocol can be found at https://osf.io/py5n6/, and revisions to the preregistration are documented at https://osf.io/hz265/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619831612. This article has received badges for Open Data, Open Materials, and Preregistration. The "TC" notation on the preregistration badge refers to "transparent changes" to the preregistration protocol. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Previous large-scale replication projects have typically treated the individual study as the primary unit of analysis. Because personality-outcome studies often examine multiple trait–outcome associations, we selected the individual association as the most appropriate unit of analysis for estimating replicability in this literature.
2. Because some of our replication attempts were dependent (they had a shared Big Five trait or overlapping sample of participants) rather than independent, or were conducted using aggregated results across multiple suboutcomes or subsamples, the *p* values for these analyses should be considered approximate rather than exact.
3. The original sample size was not available for one outcome.
4. Because many original studies did not report exact *p* values, we estimated these from the reported effect size and degrees of freedom.

References

Allport, G. W. (1961). *Pattern and growth in personality*. Oxford, England: Holt, Rinehart & Winston.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 4, pp. 1–181). Amsterdam, The Netherlands: Elsevier.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Hang, W. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 1433–1436.

Chernyshenko, O. S., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and well-being*. Paris, France: OECD Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. Advance online publication. doi:10.1007/s13164-018-0400-9

De Raad, B., Perugini, M., Hrebícková, M., & Szarota, P. (1998). Lingua franca of personality: Taxonomies and structures based on the psycholexical approach. *Journal of Cross-Cultural Psychology*, *29*, 212–232.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Boucher, L. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE*, *9*(10), Article e109019. doi.org/10.1371/journal.pone.0109019

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26–34.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford Press.

Kautz, T., Heckman, J. J., Diris, R., ter Weel, B., & Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (National Bureau of Economic Research Working Paper 20749). doi:10.3386/w20749

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Brumbaugh, C. C. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

McShane, B. B., Tackett, J. L., Bockenholt, U., & Gelman, A. (2017). *Large scale replication projects in contemporary psychological research*. Retrieved from arXiv: https://arxiv.org/abs/1710.06031

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Organisation for Economic Co-operation and Development. (2015). *Skills for social progress: The power of social and emotional skills*. Paris, France: OECD Publishing.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421.

Primi, R., Santos, D., John, O. P., & De Fruyt, F. (2016). Development of an inventory assessing social and emotional skills in Brazilian youth. *European Journal of Psychological Assessment*, *32*, 5–16.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data

collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.