Study 1 coder agreement

Our coding approach was designed so as to involve no judgment on the part of the coders. As a result, coding involved transcription. Once transcription was complete, coders used the transcribed values to compute within-study gender variance. For example, if a primary study was comprised of 52% men, then the coders computed gender variance for that study as .52\*.48=.2496. Finally, these values were used to compute *average* within-study gender variance as well as between-study variance in gender proportion.

Although our coding involved no judgment, we wanted to ensure that computed values were computed properly. We, therefore, made an effort to ensure intercoder agreement. To examine agreement, two authors independently coded approximately 15% of the moderation tests (39 tests from 14 meta-analyses). To do so, from each meta-analysis, each author identified the test(s) of gender moderation and then extracted relevant primary study-level information from the meta-analytic table or appendix, including each primary study’s sample size, gender composition, and information about which primary studies were included in each test of gender moderation. Using this information, each coder computed the mean within-study variance in gender proportion and the between-study variance in gender proportion. Then, the percent agreement across was computed for the within- and between-study variance computations for these 39 tests of gender as a moderator (i.e., across 78 computed values).

For example, both authors coded Cheng et al. (2013), which reports on the relationships between external locus of control and depression and between external locus of control and anxiety and includes tests of gender as a moderator of these two relationships (neither test reached significance). Cheng et al.’s (2013) Table 1 includes information about the 152 samples included in the meta-analysis. From this table, each author independently extracted the study label, *N,* % female, mean sample age, depression measure, anxiety measure, and effect sizes. The depression and anxiety measures reported in this table allowed the coders determine which primaries were included in each moderation test. Additionally, the mean sample age was extracted because the authors indicated that, in their moderation tests, they only included studies that reported both gender composition and age (see Cheng et al., 2013, Table 2, note a). Using this information, the coders computed the average within-study variance and the between-study variance in gender composition for each gender moderation test.

Once the independent coding was complete, the authors compared their computations, and there were four values that did not match (the within- and between-study variance from two tests). In each of these cases, one author included one additional sample in their computations that should not have been included. For example, in Cheng et al.’s (2013) test of gender as a moderator of the relationship between external locus of control and anxiety, Sammon et al. (1985) was not included because this article reported age range of participants rather than mean age. One coder mistakenly included this article in her computations, so the computed values did not match (in this case, between-variance was incorrectly computed as 0.106, but should have been 0.104, a trivial difference in any case). Agreement was 97.3%. We judged this to be sufficient, and the remainder of the coding was conducted by the coder who had not included the two tests that should have been omitted.