which implies that the probability of rejecting a true null hypothesis equals the alpha level for any value of alpha. On the other hand, when the larger variance is associated with the larger sample size, the frequency of *p*-values less than 5 percent decreases to 0.028 (see **Figure 2c**), and when the larger variance is associated with the smaller sample size, the frequency of *p*-values less than 5 percent increases to 0.083 (see **Figure 2d**). Welch's *t*-test has a more stable Type 1 error rate (see Keselman et al., 1998; Keselman, Othman, Wilcox, & Fradette, 2004; Moser & Stevens, 1992; Zimmerman, 2004) Additional simulations, presented in the additional file, show that these scenarios are similar for several shapes of distributions (see tables A3.1 to A3.9 and table A4 in the additional file).

Moreover, as discussed previously, with very small SDRs, Welch's *t*-test still has a better control of Type 1 error rates than Student's *t*-test, even if neither of them give critical values (i.e., values under 0.025 or above 0.075, according to the definition of Bradley, 1975). With SDR = 1.1, when the larger variance is associated with the larger sample size, the frequency of Student's *p*-value being less than 5 percent decreases to 0.046, and when the larger variance is associated with the smaller sample size, the frequency of Student's *p*-value being less than 5 percent increases to 0.054. On the other side, the frequency of Welch's *p*-values being below 0.05 is exactly 5 percent in both cases.

Yuen's *t*-test is not a good unconditional alternative because we observe an unacceptable departure from the nominal alpha risk of 5 percent for several shapes of distributions (see tables A3.1, A3.4, A3.7, A3.8, and A3.9 in the additional file), particularly when we are studying asymmetric distributions of unequal shapes (see tables A3.8 and A3.9 in the additional file). Moreover, even when Yuen's Type 1 error does not show a critical departure from the nominal alpha risk (i.e., values above 0.075), Welch's *t*-test more accurately controls the Type 1 error rate (see tables A3.2, A3.3, A3.5, and A3.6 in the additional file). The Type 1 error rate of Welch's *t*-test remains closer to the nominal size (i.e., 5%) in all the previously discussed cases and also performs better with very extreme SDRs and unbalanced designs, as long as there are at least 10 subjects per groups (See table A4 in the additional file).

In **Figure 3**, *p*-values from Welch's *t*-test and Student's *t*-test tests, shown separately in **Figure 2** (through histograms), are now plotted against each other. **Figure 3a** shows Student's *p*-values plotted against Welch's *p*-values of Scenario 1, where the variance is the same in each group (SDR = 1) and sample sizes are unequal. **Figure 3b** displays Student's *p*-values plotted against Welch's *p*-values of Scenario 2, where the

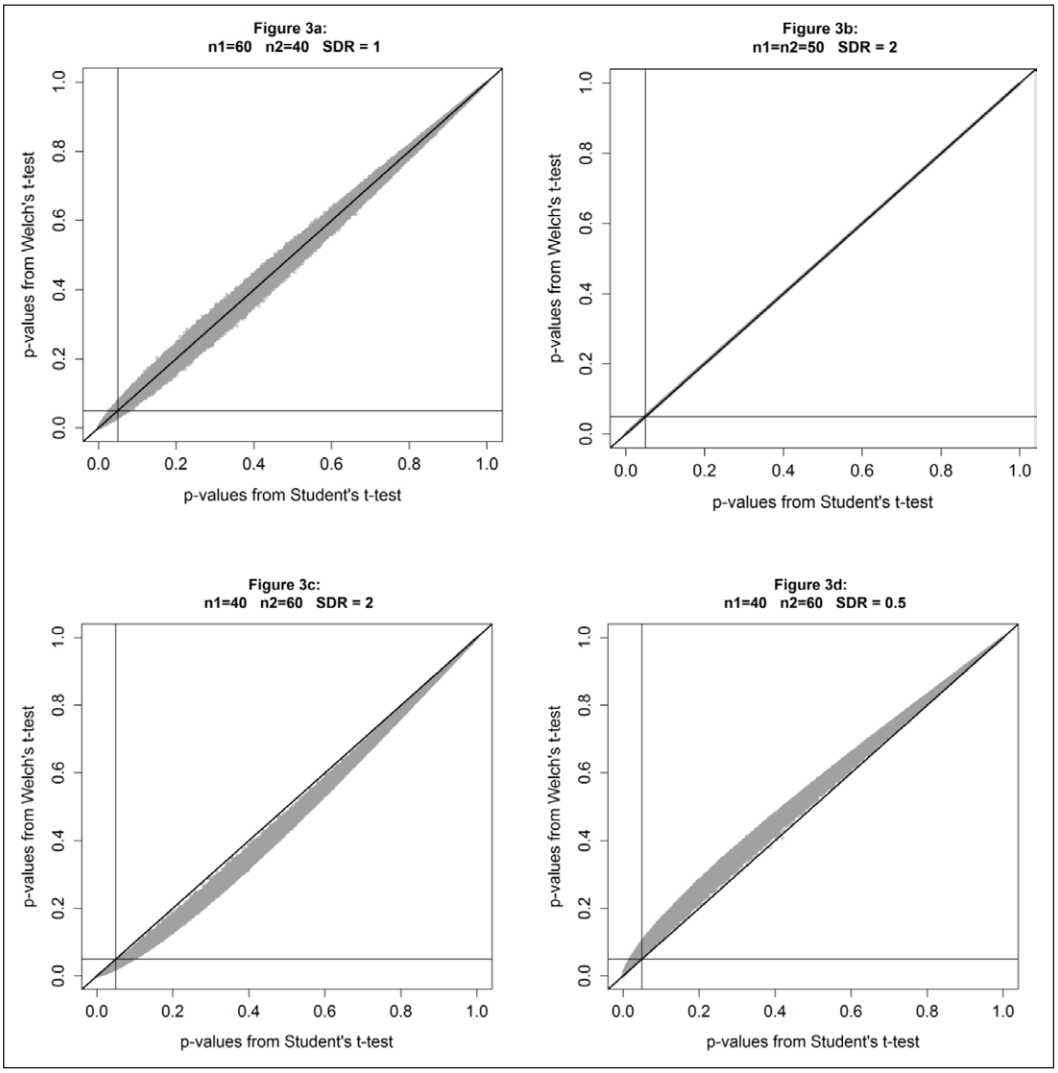

Figure 3: *P*-values from Student's *t*-test against *p*-values from Welch's *t*-test under the null.