Inter-Coder Reliability Check and Sensitivity Analysis

1 Procedure

In order to test my application of the coding instrument, I hired two graduate students to code select chapters from Taylor's *Struggle for Mastery of Europe*. Specifically, I had the students code chapters 10, 11, and 12 (approximately 80 pages), as these three chapters contained perhaps the largest concentration of failed negotiations.

I instructed the students to read the chapters looking for evidence of (1) a **meeting** (correspondance of letters, physical meeting) at the diplomatic level (between ambassadors, heads of state, foreign ministers) in which (2) a **proposal** of a formal (i.e. written) alliance (defensive, offensive, neutrality, consultative, or non-aggression) is made and there is evidence of (3) a **rejection** (one side must decline forming the alliance).

2 Intercoder Reliability Results

All three coders (myself and the two graduate students) individually identified the same cases of failed negotations with economic linkage. With respect to the total number of failed alliance negotiations, I coded 15 failures, student 1 coded 19 failures and student 2 coded 17 failures. Though the two graduate students coded more cases of alliance failure than me, my failures are a subset of the failures identified by the students. This generates an intercoder reliability rate of between $\frac{15}{19} = 0.79$ and $\frac{15}{17} = 0.89$, which is well above the 0.75 rate of acceptability.

3 Sensitivity Analysis

The inter-coder reliability check suggests that another individual could obtain a slightly different set of failed negotiations following my criteria. That descrepancies in coding could arise is no surprising. Even when coding successful negotiations (for which actual treaty text exists), scholars disagree. For example, the Correlates of War (COW) project and ATOP both provide datasets of the alliances that have existed since 1815. However, whereas COW identifies just under 500 alliances, ATOP identifies nearly 650. Coding failed negotiations requires a larger degree of judgement than coding successful negotiations (due to lack of a treaty), so the real question is not "will descrepancies arise?", but "are the inferences I draw from the data sensitive to such descrepancies?"

To determine if this is the case, I compare two estimates of the ATE for economic linkage offers during the 1870 to 1881 time period (the period covered by the chapters coded by the graduate student coders): the ATE from using my coding of failed negotiations and the ATE from using graduate student 1's coding of failed negotiations. I chose the coding of graduate student 1 as the

number of failed negotiations identifeid by this graduate student (19) serves as an upper bound on the number of failed negotiations a coder could have identified for this time period. The ATE using my coding is -0.25, while the ATE using graduate student 1's coding is -0.275. Thus, my coding of failed negotiations could be 10 percent larger ($\frac{0.275-0.25}{0.25}$. IN fact, if one assume that the bias could be in either direction, then the estimated positive ATE for the post-1880 time period could be as large as $0.24 * (1 + .10) = 26.4$ or as small as $0.24 * (1 - 0.10) = 21.6$, while the negative effect identified for the pre-1880 time period (using the sample that removes k-ads with states that share borders) could be as large as $-0.38 * (1 + .10) = -41.8$ or as small as $-0.38 * (1 - 0.10) = -34.2$. In short, such coding descrepancies are unlikely to substantively influence my results.