**Porter, Verdery and Gaddis. 2016. Study 3 Detailed Reporting**

**Substantive Goal:** Code availability of mental health information and services on university websites

**Methodological Goal:** Compare worker accuracy and time for more complex web search/coding tasks depending on structure of demands (longer or more diverse) and presence or absence of explicit warning of possible rejection.

**HIT Description:**

Workers were directed to an external site (SurveyMonkey) and asked to answer a series of questions about features related to mental health information and services of a specified university’s website. Half of workers were randomly assigned to answer a long series of questions related to a single topic, while half received less questions but covering three different types of information. A second experiment assigned 50% of each group to be exposed to an explicit warning that poor work could be rejected, while the remaining workers did not. For each of three universities, 10 workers were assigned to each of the four conditions. All questions were either true/false or open response.

Additional characteristics are described in Table A4.

**Worker Qualifications:**

Only workers with at least 5,000 HITs previously approved at greater than 95% approval rate were allowed to accept the HITs. Each worker was prevented from completing more than one HIT.

**Rejection Criteria:**

Sample checked for accuracy against consensus; lowest accuracy rates rejected.

**Additional Validation Checks:**

Three trained coders independently answered the full series of questions in advance of the HIT. In the event of disagreement, the coders reached a consensus answer for each True/False question. Worker accuracy was evaluated based on the consensus answer. Accuracy rates for each question by institution and experimental condition can be found online.

**Overall Results:**

Table A4 summarizes findings. Mechanical Turk, as implemented here, was not found to be an effective method of collecting the complexity of data required for the HIT, as overall accuracy rates for many questions were only a marginal improvement on chance.

**Table A4: Accuracy by question and experimental condition**

