Online Appendix For:

**When does garbage start to stink? The impact of imperfect human annotations as a gold standard on the validation of automated content analysis**

**1. Variables coded in Study 1, detailed coding instructions, and reliability estimates**

Following variables utilized in Study 1 were coded by 5 trained coders.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition & Coding instructions** | **Reliability** |
| Relevance | Whether empirical text analysis is conducted and reported (Yes = 1, No = 0) | Alpha = 1 |
| Method Used | 1 = Search string based / Dictionary Approach  2 = Machine Learning  3 = Topic Modeling (excluded from further analysis)  4 = Other (excluded from further analysis) | Alpha = 1 |
| Refer to gold standard | 1 = Yes, a “gold standard” is used, and info is reported  0 = No is not used reported | Alpha = 1 |
| Report reliability | Whether intercoder-reliability of human-coded materials are reported?  (1 = Yes, reported, 0 = Not reported) | Alpha = 1 |
| Refer to validation / Report validation measures | Whether validation of automated procedures are mentioned, and if so, whether either one of validation metrics (e.g., Recall, Sensitivity, Precision, Accuracy, F1, or other measures) is reported?  (1 = Yes, mentioned, 0 = Not mentioned) | Alpha = .750 |

A total of ﬁve highly qualiﬁed coders tested the initial coding scheme by independently coding 10 sample articles (approximately 5% of the total retrieved sample) and collectively discussed any coding problems and disagreement. Coding instructions were iteratively revised until the coding schemes would produce reliable results.



Figure A1. Overall classification accuracy against true value across MC simulation conditions, Study 2 (reference line is the overall mean).