Reliability and Internal Consistency of a Domain-General Adaptation of the Downs-Black (1998) Tool

Anthony Cruz

12/8/2021

# Abstract

Here, we assess the reliability and internal consistency of a domain-general modification to the Downs-Black tool, a checklist for the assessment of methodological quality in studies of healthcare interventions (Downs & Black, 1998). The tool is valid and reliable but is not well-suited for studies unrelated to healthcare interventions. Our modification of this tool is intended for use in a wider variety of research areas, such as the social and behavioral sciences, which do not often include risk of bias assessments in their literature reviews. We asked two independent Raters to use our modified Downs-Black tool to assess the quality of a set of 12 research articles and used their scores to assess the inter-rater reliability and internal consistency of the tool. Our results suggest that this tool has sufficient reliability and internal consistency for assessing the risk of bias in systematic reviews outside the healthcare domain. Future research should assess the quality of this checklist for a wider variety of review subjects and using Raters with differing levels of expertise.

# Reliability and Internal Consistency of a Domain-General Adaptation of the Downs-Black (1998) Tool

Literature review is paramount to the development of scientific ideas, but all literature reviews are not created equal. Reviews can easily become subject to bias, systematic error in results leading to overestimation or underestimation of effects (Boutron et al., 2021). A systematic review is a type of evidence synthesis that minimizes risk of bias through the identification, appraisal (or risk of bias assessment), and synthesis of all relevant research on a given topic (Uman, 2011). Bias can arise from the actions of primary investigators or from the actions of reviewers, whether intentional or not. Assessing the risk of bias (rather than bias itself, as it is impossible to truly assess bias) allows for scrutiny of the results. This feature is typically absent in narrative reviews , which are commonplace in the social and behavioral sciences, such as psychology.

There is currently no common, domain-general tool for the appraisal of non-clinical studies. There exist many tools for appraising clinical studies, however (Page et al., 2018). One such tool is the Downs-Black (1998) Tool, which assesses the external validity, internal validity, and open reporting in a given study . This tool is valid and reliable, but very domain-specific; it is not well-suited for studies outside of the healthcare domain. From 2012 to the time of writing, Web of Science reports that this paper has been cited 3870 times (Web of Science). Of these citations, the vast majority come from categories related to health or medicine; only 72 come from Psychology and 1 from Sociology. This domain specificity is not unique to the Downs-Black Tool. Page et al. (2018) conducted a systematic review of risk of bias assessment tools and found that the vast majority of available tools are domain-specific. Despite obvious differences between clinical trials and empirical psychological studies, there are key similarities that make the risk of bias assessment tools used in clinical trials adaptable to the non-clinical literature. In other words, given the clear need for a reliable quality assessment tool for conducting systematic reviews in non-clinical domains and given the wide-spread adoption of tools like Downs-Black, we addressed this with a modification of the Downs-Black, rather than starting from scratch to build our own domain-specific tool. By adopting this approach, we retain many of the advantages of Downs-Black but extend its use and generalizability.

The present study is a reliability and internal consistency assessment for a domain-general modification of the Downs-Black Tool. If valid and reliable, this tool may make risk of bias assessment in the social and behavioral sciences more commonplace. Given their similarities, we are expecting this tool and the original Downs & Black (1998) Tool to have similar reliability and internal consistency.

# Methods

## Pre-Registration

The design and pre-analytic plan for this study were pre-specified in an open-ended registration uploaded to Open Science Framework in October 2021 (<https://osf.io/8ycj9>).

## Sample

The sample size for this study (i.e. the number of papers that were analyzed) was chosen using an a priori power analysis in G\*Power 3.1.9.7 (Faul et al., 2007). With power=.95, false positive rate α=.05, and effect size |ρ|=.77 (from Downs & Black, 1998), the projected sample size needed for this study was N=12. Dropouts and incomplete data were non-issues due to the nature of this study. The set of 12 empirical research articles (published between 2006 and 2018) to be assessed were chosen arbitrarily from a set of articles that advanced past title & abstract screening for a related systematic review (pre-registered at <https://osf.io/5dmau>). The goal of this related review is to catalog the extant research examining neural evidence of changes in stimulus representation that result from artificial category learning. At the time of the present Downs-Black modification project, the papers we included here had not yet been fully screened and may not be included in the final review, but it can be assumed that their contents are related to the topic of the review.

## Survey Design

Our modification of the Downs-Black Tool is homologous to the original Downs-Black Tool in terms of scoring and subscales. Each checklist item receives a score and the total score can fall between 0 and 31. The total score will be referred to as the Quality Index. This checklist includes the following subscales: Reporting (items 1-10), External Validity (items 11-13), Internal Validity - Bias (Items 14-20), Internal Validity - Confounding (Items 21-26), Validity (items 11-26), and Power (item 27). At the end of the checklist, each Rater is asked to provide a subjective rating, ranging from 0 to 10, of the overall quality of the study. This measure will be referred to as the Global Score. Several items were reworded to support use outside of the clinical domain. Of the 27 survey items, 19 were modified. The final survey was uploaded to Qualtrics. Responses were required for all items, barring an optional question asking if Raters have any questions or comments related to the assessment. Raters completed the assessment independently and remotely for each study in no particular order.

## Test Procedure

Two Raters assessed each study. These Rater s were familiar with the protocol and assisted in title & abstract screening for the related systematic review. However, they were not involved in the development of this checklist or the development of the protocol for the present research. They were also not explicitly instructed that they were piloting this tool for a study on reliability. Rather, they were instructed that we were assessing the feasibility of conducting risk of bias assessment on Qualtrics. Data collection ran from the 25th of June 2021 to the 5th of November 2021.

# Results

Rater A’s mean Quality Index was 21.25 (Inter-Quartile Range: [19, 23]). Rater B’s mean Quality Index was 18.25 (Inter-Quartile Range: [16, 21]). Overall, the Quality Index had a mean score of 19.75 (Inter-Quartile Range: [17.75, 22]). See Table 1 for a summary of each subscale.

## Table 1

### Subscale Summary Data

|  | | Inter-Quartile Range | |
| --- | --- | --- | --- |
| Subscale | Mean | 25th Percentile | 75th Percentile |
| Reporting | 8.250 | 8.00 | 10 |
| External Validity | 0.042 | 0.00 | 0 |
| Internal Validity - Bias | 5.250 | 5.00 | 6 |
| Internal Validity - Confounding | 3.417 | 3.00 | 4 |
| Power | 2.792 | 1.75 | 4 |

*Note.* The values in this table are taken across both Raters.

## Inter-Rater Reliability

### Quality Index and Subscales

We assessed inter-rater reliability at the Quality Index and subscale levels using Spearman correlations. Confidence intervals were obtained using 20000 random bootstrapped samples identical in size to those used in the study. The inter-rater Quality Index correlation was significant, r = 0.556 (90% CI: [0.073, 0.917]). Due to the lack of variability within the External Validity subscale, it was impossible to compute the Spearman correlation for this subscale. Correlations for the other subscales are reported in Table 2.

## Table 2

### Inter-Rater Spearman Correlations

|  | | Bootstrapped 90% Confidence Interval | |
| --- | --- | --- | --- |
| Subscale | r | Lower Limit | Upper Limit |
| Reporting | 0.467 | -0.020 | 0.870 |
| Internal Validity - Bias | -0.333 | -0.577 | -0.135 |
| Internal Validity - Confounding | 0.524 | 0.147 | 0.777 |

### Item-by-Item

Cohen’s Kappa and percent disagreement were computed for each item in the scale. These values are reported in Table 3. On average, Raters only disagreed 17.593% of the time. 12 of the 27 items (44%) had poor agreement, similar to the 11 of 26 (42%) reported by Downs & Black (1998). 7 of the 27 items (26%) had no inter-rater disagreement, higher than the 2 of 26 (8%) reported by Downs & Black (1998).

## Table 3

### Item-by-Item Inter-Rater Reliability

| Subscale | Item | Percent Disagreement | Cohen's Kappa |
| --- | --- | --- | --- |
| Reporting | Q1 | 0.000 |  |
| Q2 | 8.333 | 0.000 |
| Q3 | 8.333 | 0.750 |
| Q4 | 25.000 | 0.000 |
| Q5 | 66.667 | -0.043 |
| Q6 | 0.000 |  |
| Q7 | 0.000 |  |
| Q8 | 8.333 | 0.000 |
| Q9 | 16.667 | 0.429 |
| Q10 | 41.667 | 0.250 |
| External Validity | Q11 | 8.333 | 0.000 |
| Q12 | 0.000 |  |
| Q13 | 0.000 |  |
| Internal Validity - Bias | Q14 | 16.667 | 0.000 |
| Q15 | 0.000 |  |
| Q16 | 66.667 | 0.000 |
| Q17 | 33.333 | -0.143 |
| Q18 | 8.333 | 0.000 |
| Q19 | 16.667 | 0.000 |
| Q20 | 8.333 | 0.000 |
| Internal Validity - Confounding | Q21 | 33.333 | -0.200 |
| Q22 | 0.000 |  |
| Q23 | 16.667 | 0.556 |
| Q24 | 16.667 | 0.556 |
| Q25 | 33.333 | 0.250 |
| Q26 | 16.667 | 0.556 |
| Power | Q27 | 25.000 | 0.690 |

## Internal Consistency

Internal consistency was assessed for individual items and for overall subscales using Cronbach’s alpha, alpha-if-item-removed, and the KR-20. Note that the KR-20 is designed for use with dichotomous outcomes. Item 5 in this checklist is nondichotomous, but we include the KR-20, regardless, for the sake of comparison with Downs & Black (1998). Item 5 was dichotmoized by treated “Partially” and “Yes” responses identically. Table 4 describes the results. Inter-item correlations were also computed and are shown in Figure 1. Note that the items excluded from this figure had no variability, making it impossible to compute correlations.

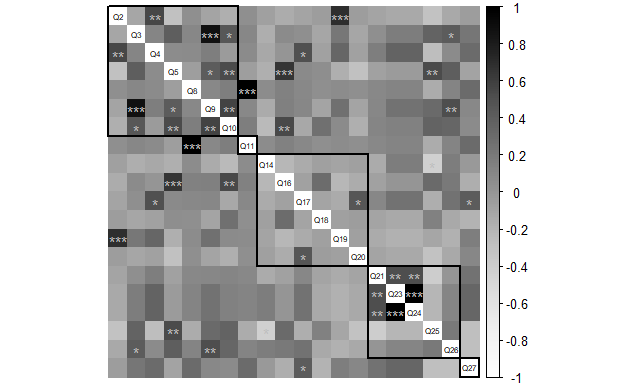
## Table 4

### Internal Consistency of Subscales and Individual Items

| Subscale | KR-20 | Alpha | Item | Alpha if Item Removed | Change in Alpha if Item Removed |
| --- | --- | --- | --- | --- | --- |
| Reporting | 0.383 | 0.659 | Q2 | 0.699 | 0.040 |
| Q3 | 0.541 | -0.117 |
| Q4 | 0.670 | 0.011 |
| Q5 | 0.628 | -0.031 |
| Q8 | 0.675 | 0.016 |
| Q9 | 0.522 | -0.136 |
| Q10 | 0.553 | -0.106 |
| Internal Validity - Bias | -0.429 | -0.429 | Q14 | -0.119 | 0.309 |
| Q16 | -0.173 | 0.256 |
| Q17 | -0.492 | -0.063 |
| Q18 | -0.605 | -0.176 |
| Q19 | -0.119 | 0.309 |
| Q20 | -0.605 | -0.176 |
| Internal Validity - Confounding | 0.477 | 0.477 | Q21 | 0.431 | -0.046 |
| Q23 | 0.115 | -0.362 |
| Q24 | 0.115 | -0.362 |
| Q25 | 0.707 | 0.230 |
| Q26 | 0.472 | -0.005 |

## Figure 1

### Inter-Item Correlation Matrix

 *Note*. \*\*\* p<.001, \*\* p<.01, \* p<.05.

## Noncompliance Check

A Spearman correlation was calculated to examine the relationship between Quality Index and Global Score for each Rater . This correlation indicates whether the Raters’ subjective assessments align with the objective assessments provided by the checklist. Confidence intervals were obtained using the aforementioned bootstrapping procedure. For Rater A, r = -0.023 (90% CI: [-0.64, 0.688]). For Rater B, r = 0.562 (90% CI: [-0.124, 0.734]). The overall correlation was r = 0.411 (90% CI: [-0.124, 0.734]).

# Discussion

Overall, the results seem similar to those reported by Downs & Black (1998), with inter-rater reliability and internal consistency appearing at similar levels. This suggests that our modification of this tool maintains many of the key properties that have resulted in the widespread adoption of this tool. Of note is the low variability in the External Validity subscale. Between both Raters, only one non-zero response was given to any question in this subscale. The studies that were assessed all utilized artificial category learning paradigms which are known to be very low in ecological validity. This subscale should receive further assessment in a different area of research.

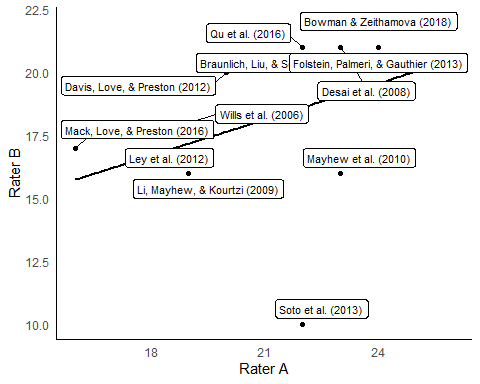
Downs & Black (1998) reported a mean Quality Index of 11.7 (SD = 4.64) for non-randomized studies, which is substantially lower than our mean of 19.75. This difference may be partially driven by the Reporting subscale, with our mean score (8.25) and minimum score (4) being higher than those reported by Downs & Black. Additionally, some of the studies we assessed received non-zero scores on the Power subscale, unlike those assessed by Downs & Black. Crude summary statistics for the other subscales look comparable. It is unclear if these differences reflect a true difference in methodological quality between the two samples, however. It is possible that our modifications resulted in some checklist items becoming easier to achieve. Additionally, Downs & Black used raters who were experts in the field, which may have resulted in more conservative assessments.

## Inter-Rater Reliability

A weak negative inter-rater correlation was found for the Internal Validity - Bias subscale. This suggests that both Raters had a tendency to give opposite scores within this subscale, indicating that one or more subscale items may have been unclear. Despite this, moderate positive inter-rater correlations in Quality Index and Internal Validity - Confounding were observed, suggesting high agreement between Raters. This high correlation for the Quality Index suggests that, despite variability at the subscale level, Raters tended to have high agreement in their risk of bias assessments. Figure 2 shows the agreement between both Raters. Of note is the study by Soto et al. (2013); this study received an unusually low score from Reviewer B. It is unclear at a glance what sets this study apart from the others, but removing this study from the analysis results in a much stronger inter-rater correlation (r = 0.674). Given the overall pattern, we expect that Rater A’s score is closer to the true score.

## Figure 2

### Inter-Rater Agreement in Quality Indices



Of the 27 checklist items, 7 (25.926%) had no inter-rater disagreement. It is difficult to interpret this outcome given the low variability in these measures. Still, this suggests that these test items were clear and unambiguous. Items 16 and 5 had the highest rate of disagreement at 66.667%. Cohen’s Kappa, a measure of agreement that considers the chance of random agreement, was negative for items 5, 17, and 21; these values are all small, however, and do not represent a notable amount of disagreement. Minimal agreement is seen in items 10 and 25,; weak agreement is seen in items 9, 23, 24, and 26; and moderate agreement is seen in items 3 and 27 (McHugh, 2012).

## Internal Consistency

The Bias subscale had a negative Cronbach’s Alpha. In many cases, this indicates an issue with data, such as failure to reverse score some (but not all) data points. This is not the case here. Instead, it seems to be an issue with the subscale itself. Despite this, positive Alpha values for the other subscales indicate acceptable levels of internal consistency. Figure 1 shows an inter-item correlation matrix. Items within the same subscale should be correlated with each other. While this was not always the case, the figure shows a small amount of clustering within the subscales, also suggesting a good amount of internal consistency.

## Noncompliance Check

The potential for non-compliance in this study was low, but present. Reviewers had an interest in adequately completing the checklist due to their involvement in the associated systematic review. However, raters completed the assessments remotely, which made it difficult to directly monitor compliance. This was addressed by conducting a Spearman correlation between Quality Index and Global Score for each reviewer. We expected these values to be positively correlated. Counter to our prediction, we did not find a significant correlation between these values. Global Scores tended to be high and had low variability. This may be due to sampling bias; the papers included in this study were all peer-reviewed journal articles that passed through title & abstract screening for a systematic review, making it possible that reviewers assigned high Global Scores to all studies despite variability in their Quality Indices. Future work should be sure to include pre-prints, dissertations, theses, and gray literature of varying quality to avoid this issue. Additionally, the reviewers were not experts in the field. Experts in the field may be more consistent in their responses.

## Conclusion

This study was limited by its reviewers and included studies. Despite limitations, the results suggest that a domain-general adaptation of the Downs-Black Tool (1998) is feasible for use in systematic literature reviews.