Raw data was imported using STATA, first excel row was imported as variable names, and all variable names were converted to lowercase letters. Survey data was then declared using the following code:

```
svyset identifier [pweight=weighttercile], strata(field) fpc(totalinfield) vce(linearized) singleunit(missing)
```

- pweight: weighttercile
- VCE: linearized
- Single unit: missing
- Strata 1: field
- SU 1: identifier
- FPC 1: totalinfield

**Codes used for calculating percentages for (Supplementary) Table 1 were:**

**Basic – percentages rounded to 0 decimal places**

```
svy linearized: tabulate variable area, column percent format(%3.0f)
```

**Percentages with confidence intervals rounded to 0 decimal places**

```
svy linearized: tabulate variable area, column ci percent format(%3.0f)
```

**Percentages with confidence intervals rounded to 1 decimal places (for percentages lower than 1)**

```
svy linearized: tabulate variable area, column ci percent format(%3.1f)
```

**Codes used for regression analyses (Supplementary) Table 2 were:**

```
svy linearized: logistic variable snip2016 ib(first).doaj ib(first).publishercode ib(last).areacode, nopvalues cformat(%9.1f)
```

**Code used for average number of topics addressed per journal across disciplines:**

```
svy linearized: mean summedscore, over(areacode)
```

**Note:** variable was changed to each of the variable names we used in the study.