**Supplement**

*Communicating evidence in icons and summary formats for policy makers: What works? (2020)*

**Methods (Supplemental)**

*Table S1.* Partner Centres

|  |  |  |
| --- | --- | --- |
| *Short name* | *Full name* |  *Website (2019)* |
| Children's Social Care (CSC) | What Works for Children's Social Care | [whatworks-csc.org.uk](https://whatworks-csc.org.uk/)  |
| Conservation Evidence | What Works in Conservation | [conservationevidence.com](http://www.conservationevidence.com)  |
| Crime Reduction | College of Policing / What Works Centre for Crime Reduction | [whatworks.college.police.uk](https://whatworks.college.police.uk)  |
| Early Intervention | Early Intervention Foundation | [eif.org.uk](http://www.eif.org.uk)  |
| Educational Endowment | Education Endowment Foundation | [educationendowmentfoundation.org.uk](https://educationendowmentfoundation.org.uk/) |
| Homelessness Impact | Centre for Homelessness Impact | [homelessnessimpact.org](https://www.homelessnessimpact.org/) |
| Local Economic Growth | What Works Centre for Local Economic Growth | [whatworksgrowth.org](https://whatworksgrowth.org/) |
| NICE | National Institute for Health and Care Excellence | [nice.org.uk](https://www.nice.org.uk/) |

***Public sample quality and representativeness***

Prolific provides an accessible source for survey responses from a relatively diverse population (in this case, all across the UK). Online participants provide comparable responses to traditionally used samples and generate reliable results that replicate previous decision-making research [(Buhrmester et al., 2011; Goodman et al., 2013)](https://paperpile.com/c/FeJuGm/mDZNR%2BIuyJU). Prolific and Amazon MTurk have inherent demographic biases. Our respondents were on average more politically progressive, younger, and better educated than the national averages, as is often seen in these online populations. The other major concern for this research when using MTurk is that respondents do not always pay attention [(Buhrmester et al., 2011; Paolacci and Chandler, 2014)](https://paperpile.com/c/FeJuGm/3x5F%2BIuyJU). However, this design did not depend on attention at any critical juncture (there was no central manipulation or single key outcome), so the consequence of lower attention would just be increased random noise.

Given these details, we suggest the Prolific sample was sufficiently valid for inclusion and more diverse than university sample pools, and therefore likely more generalizable than some traditional sampling methods such as surveying university campuses or other public spaces [(Buhrmester et al., 2011)](https://paperpile.com/c/FeJuGm/IuyJU). The limits to generalisability are explicitly mentioned in the Discussion.

***Expert sample recruitment and response rate***

 There was a wide diversity in recruitment techniques due to the decentralised recruitment design. The centres would not have given over their mailing lists, nor would they send a recruitment message by itself, and they had different requirements about newsletter wording and content. Most centres advertised twice by email to their mailing lists. The overwhelming majority of invited users were either from NICE (three mailing lists, total *n* = 42,154) or the EEF (*n* = 25,500); these two centres accounted for 89 per cent of all approached users. CSC approached a few potential respondents directly with email invitations. EEF and Homelessness both posted invitations on their websites; e.g., EEF reported a 41 per cent bounce rate on their homepage with 2270 unique pageviews during the recruitment. Conservation, EIF, and Homelessness also tweeted the survey link. Direct email or Twitter recruitment was contrary to the authors' intentions, but in the end this was a complex multi-site study and the authors did not control the recruitment process. The Twitter respondents (about 10 per cent of survey clicks) appeared comparable in quality, respondent role, and data integrity to the other respondents, and the Twitter entries were included.

**Results (Supplemental)**

 ***Evidence priorities***

Priorities are shown for each Centre (Table S2) and between practitioners and policy makers in the expert sample (Table S3).. Effectiveness and evidence quality were ranked highest priority, consistent with the other subgroups shown in the main results.

*Table S2.* Evidence priority rank by Centre (Expert sample) compared to the Prolific sample (lower numbers are higher priority)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Mean rank* | CSC | Conserv.Evidence | CrimeReduct. | Early Interven. | Edu.Endow. | Homeless.Impact | NICE | Public |
| **Effectiveness** | **2.31** | **2.29** | **3.15** | **3.13** | **2.34** | **2.90** | **3.14** | ***3.02*** |
| **Evidence quality** | **3.78** | **2.71** | **2.85** | **3.04** | **2.87** | **3.15** | **3.01** | ***3.44*** |
| Number of studies | 6.59 | 5.24 | 5.15 | 5.63 | 6.21 | 7.30 | 6.22 | *5.67* |
| Type of studies | 7.28 | 7.74 | 4.00 | 5.88 | 7.79 | 6.60 | 4.51 | *7.34* |
| Effectiveness by sub-group | 6.59 | 7.26 | 7.46 | 6.33 | 6.62 | 6.50 | 7.14 | *7.14* |
| Harms to intervention target | 7.06 | 6.84 | 7.54 | 8.17 | 8.64 | 6.00 | 6.51 | *6.82* |
| Cost (financial) | 7.34 | 7.68 | 7.92 | 6.92 | 4.79 | 7.35 | 8.44 | *6.52* |
| Time course of effect | 6.53 | 9.18 | 8.00 | 7.71 | 7.13 | 8.25 | 8.49 | *8.01* |
| Implementation advice | 8.06 | 8.95 | 9.54 | 7.88 | 7.43 | 7.90 | 8.59 | *8.94* |
| Statistical significance (*p*-values) | 9.69 | 9.29 | 9.54 | 9.50 | 8.09 | 8.55 | 7.59 | *7.94* |
| Uncertainty around effectiveness estimate | 8.91 | 7.11 | 8.85 | 9.38 | 9.51 | 9.10 | 8.33 | *8.44* |
| Impacts to external groups or people | 8.03 | 7.84 | 9.54 | 9.58 | 9.51 | 8.95 | 9.10 | *7.66* |
| Location where intervention studies were conducted | 8.81 | 8.87 | 7.46 | 7.88 | 10.1 | 8.45 | 9.92 | *10.0* |
| *n* | *32* | *38* | *13* | *24* | *47* | *20* | *91* | *245* |

*Table S3.*Evidence priority rank (highest priority = 1; expert sample only)

|  |  |  |  |
| --- | --- | --- | --- |
| *Mean rank* | all | policy makers | practitioners |
| **Effectiveness** | **2.76** | **2.12** | **2.93** |
| **Evidence quality** | **3.04** | **3.94** | **2.84** |
| Number of studies  | 6.11 | 7.42 | 6.04 |
| Type of studies | 6.15 | 7.30 | 5.85 |
| Effectiveness by sub-group | 6.91 | 6.03 | 6.72 |
| Harms to intervention target | 7.16 | 5.85 | 7.79 |
| Cost (financial) | 7.28 | 5.79 | 7.45 |
| Time course of effect | 7.99 | 6.88 | 7.82 |
| Implementation advice | 8.31 | 8.76 | 7.61 |
| Statistical significance (*p*-values) | 8.53 | 8.94 | 8.12 |
| Uncertainty around effectiveness estimate | 8.60 | 9.64 | 8.86 |
| Impacts to external groups or people | 8.90 | 8.73 | 9.25 |
| Location where intervention studies were conducted | 9.26 | 9.61 | 9.73 |
| *n* | *266* | *33* | *109* |

*Note.* Effectiveness and evidence quality were ranked highest priority and are shown in bold.

*Subgroups of interest.* Both samples optionally indicated which population subgroups were of greatest interest to them when learning about intervention evidence (see Table S4). The most common response was related to age (e.g., adult vs. child, elderly vs. not elderly, child school age, etc.), followed by gender or sex. Note that these results are likely skewed by the over-representation of some What Works Centres, particularly NICE.

*Table S4.*Subgroups of interest (counts coded from open response)

|  |  |  |
| --- | --- | --- |
| *count* | *Expert* | *Public* |
| Age | 109 | 71 |
| Gender/sex | 37 | 32 |
| Ethnicity | 11 | 15 |
| Species/habitat | 14 | 2 |
| Medical status | 28 | 9 |
| Region | 29 | 17 |
| SES/benefits | 24 | 18 |
| Household status | 16 | 3 |
| Other/uncodable | 33 | 36 |

*Table S5.* Preference for separate or aggregated outcomes

|  |  |  |
| --- | --- | --- |
| % | *Expert**n* = 295 | *Public**n* = 257 |
| Strongly prefer separate outcomes | 45.4 | 36.2 |
| Slightly prefer separate outcome | 18.0 | 22.6 |
| No preference | 13.6 | 13.6 |
| Slightly prefer a single outcome | 12.2 | 17.5 |
| Strongly prefer a single outcome |  4.8 | 4.3 |
| Don't know |  6.1 | 5.8 |

*Table S6.* Preferences for separate or aggregated intervention labels

|  |  |  |
| --- | --- | --- |
| % | *Expert**n* = 286 | *Public**n* = 255 |
| Strongly prefer separate labels | 45.5 | 42.4 |
| Slightly prefer separate labels | 31.8 | 30.2 |
| No preference |  7.0 |  7.8 |
| Slightly prefer a single label |  9.8 | 11.4 |
| Strongly prefer a single label |  2.5 |  4.7 |
| Don't know |  3.5 |  3.5 |

*Mean = 1.9/5 and median = "Slightly prefer SEPARATE labels" excluding don't know.*

*Table S7*. Preferences for communicating subgroup results

|  |  |  |
| --- | --- | --- |
| % | *Expert**n* = 292 | *Public**n* = 255 |
| Strongly prefer separate lines with subgroups | 70.9 | 52.2 |
| Slightly prefer separate lines with subgroups | 14.7 | 22.8 |
| No preference |  5.5 |  8.6 |
| Slightly prefer a single summary |  3.8 |  9.4 |
| Strongly prefer a single summary |  2.7 |  3.9 |
| Don't know |  2.4 |  3.1 |

*Mean = 1.5/5 and median = "Strongly prefer separate lines with subgroups" excluding don't know.*

*Table S8.*Policy decisions influenced (expert sample only)

|  |  |
| --- | --- |
| *n* = 205 | % |
| None | 15.6 |
| 1-2 | 22.4 |
| 3-4 |  8.3 |
| 5+ | 19.0 |
| Don't know | 22.4 |
| Not applicable | 12.2 |

## *Note.* Median = "1-2" decisions, excluding *don't know* (*n* = 46) and *not applicable* (*n* = 25).

##

##

## Effectiveness: All icons in both samples

The 'most common wrong answers' all refer to the Expert results. The response options were the same as the main results (see Methods and Table 3) or the unique response options are given for each item.

Items included in the main results table are marked "Main icon", and secondary items also included in the comprehension composite are marked "Secondary icon". The microscope and GRADE icons were not included because they were not in use at that time by any What Works Centre.

***Educational Endowment***

**

Main icon

**Expert: 52.4% correct** (*effectiveness or impact*). Most common wrong answers: *quality of evidence; how long it works for.*

**Public: 45.8% correct**

***Children's Social Care***

**

Main icon

**Expert: 63.2% correct** (*effectiveness or impact*). Most common wrong answer: *quality of evidence.*

**Public: 49.3% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *very low or no impact; unknown effectiveness (limited evidence); insufficient evidence available; mixed or insignificant impact; low-quality evidence; don't know.*

**Expert: 53.0% correct** (*very low or no impact; mixed or insignificant impact* were both counted as correct). Most common wrong answers: *unknown effectiveness (limited evidence), insufficient evidence available.*

**Public: 60.2% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *negative effect; very limited evidence; overall, evidence suggests an increase in bad outcomes; at least one high quality study shows negative impact; likely to be ineffective or harmful; don't know.*

**Expert: 66.8% correct** (*negative effect; likely to be ineffective or harmful* both marked correct). Most common wrong answer: none.

**Public: 71.7% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *beneficial; very positive effect; high-quality evidence; considerable reliable evidence; multiple high-quality studies show a positive impact; don't know.*

**Expert: 57.7% correct** (*very positive effect).* Most common wrong answer: *beneficial; multiple high-quality studies show a positive impact.*

**Public: 68.9% correct**

***Homelessness***

**

Main icon

**Expert: 58.0% correct** (*effectiveness or impact*). Most common wrong answer: *quality of evidence.*

**Public: 42.9% correct**

***Conservation***

**

Main icon

**Expert: 24.1% correct** (*effectiveness or impact*). Most common wrong answers: *quality of evidence; type of studies; ease of implementation.*

**Public: 20.8% correct**

Note that this graphic is poorly suited to colorblind users. This is how individuals with deuteranopia would view the colors from red to green:

**

*--*

**

Secondary icon

Different prompt and response scale: *Now, please look at this summary of an intervention from Conservation Evidence. This is the summary of an intervention from multiple studies. What do you think "Effectiveness: 67%" means in this summary?* Response options: *Out of 100 times, this intervention works 67 times; This intervention provides 67% of the target benefits; There is 67% certainty that the intervention will work; 67% refers to a combination of effectiveness, certainty, and harms; Don't know.*

**Expert: 6.0% correct** (*67% refers to a combination of effectiveness, certainty, and harms*). Most common wrong answer: *Out of 100 times, this intervention works 67 times.*

**Public: 4.7% correct**

***Crime Reduction***

**

Main icon

**Expert: 49.8% correct** *(effectiveness or impact)*. Most common wrong answer: *quality of evidence.*

**Public: 32.1% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *negative effect; very limited evidence; overall, evidence suggests an increase in bad outcomes; at least one high quality study shows negative impact; likely to be ineffective or harmful; don't know.*

**Expert: 50.6% correct** (*negative effect; overall, evidence suggests an increase in bad outcomes* both scored correct). Most common wrong answers: *likely to be ineffective or harmful; don't know.*

**Public: 48.4% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *overall, evidence suggests no impact (but some studies show either an increase or a decrease); overall, evidence suggests a decrease in bad outcomes (but some studies suggest an increase); low-quality evidence; mixed or insignificant impact; trade-off between benefit and harms; don't know.*

**Expert: 62.0% correct** [*overall, evidence suggests no impact (but some studies show either an increase or a decrease]; mixed or insignificant impact* both scored correct). Most common wrong answer: *trade-off between benefits and harms.*

**Public: 56.0% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *very low or no impact; unknown effectiveness (limited evidence); negative effect; mixed or insignificant impact; low-quality evidence; don't know.*

**Expert: 28.1% correct** [*unknown effectiveness (limited evidence); low-quality evidence* both scored correct]. Most common wrong answers: *negative effect; don't know.*

**Public: 25.6% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *overall, evidence suggests no impact (but some studies show either an increase or a decrease); overall, evidence suggests a decrease in bad outcomes (but some studies suggest an increase); low-quality evidence; mixed or insignificant impact; trade-off between benefit and harms; don't know.*

**Expert: 16.4% correct** *(overall, evidence suggests a decrease in bad outcomes (but some studies suggest an increase)*. Most common wrong answers: *trade-off between benefits and harms; don't know.* Note that this graphic is poorly suited to color-blind users: the red and green colors have similar hues.

**Public: 29.2% correct**

## Evidence Quality: All Icons

***Educational Endowment***

**

Main icon

**Expert: 12.0% correct** (*quality of evidence*). Most common wrong answer: *data security* (78.0%). This Centre describes these icons as 'security ratings', which may partially explain this result. See their [Toolkit manual](https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf). Practitioners only scored 10%, but policy makers scored 17% in the Expert sample.

**Public: 1.2% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this graphic*.

Responses: *extensive evidence; likely to be beneficial; multiple studies show a positive impact; moderate-quality evidence; some reliable evidence; don't know.*

**Expert: 46.1% correct** (*extensive evidence; multiple studies show a positive impact* both scored correct). Most common wrong answers: *likely to be beneficial; don't know.*

**Public: 41.6% correct**

***Children's Social Care***

**

Main icon

**Expert: 13.8% correct** (*quality of evidence*). Most common wrong answers: *effectiveness or impact; ease of implementation; public opinion*. Note that practitioners scored 12%, but policy makers only scored 5% in the Expert sample.

**Public: 8.3% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this graphic*.

Responses: *very low-quality evidence; likely to be ineffective or harmful; insufficient evidence available; very low-quality evidence; limited reliable evidence; overall, evidence suggests an increase in bad outcomes; don't know.*

**Expert: 20.7% correct** (*very low-quality evidence; insufficient evidence available* both scored correct. Note that the verbatim answer was not listed: "*Very low strength evidence* / *No acceptable quality studies"*). Most common wrong answers: *limited reliable evidence; very low-quality evidence; don't know.*

**Public: 23.6% correct**

***Homelessness***

**

Main icon

**Expert: 9.1% correct** (*quality of evidence*). Most common wrong answers: *data security* (74.8%). Note that practitioners only scored 5%, but policy makers scored 16% in the Expert sample.

**Public: 1.2% correct**

***Microscope*** (designed by [Luna 9](http://www.luna9design.com); CC-BY free-use license. Not in current use.)

**

Main icon

**Expert: 62.8% correct** (*quality of evidence*). Most common wrong answer: *type of studies*. Note that practitioners scored 67%, but policy makers only 46% in the Expert sample.

**Public: 58.1% correct**

***GRADE*** *(separate evidence quality framework)*

**

Different prompt and response scale: *Please guess the correct label for this graphic*.

Responses: *very extensive evidence; considerable reliable evidence; beneficial; multiple high quality studies show a positive impact; high-quality evidence; don't know.*

**Expert: 68.3% correct** (*multiple high quality studies show a positive impact; high-quality evidence* both scored correct). This item was included to compare with how the commonly used icons were interpreted. The most common answer was 38.2%: *multiple high quality studies show a positive impact*.

**Public: 73.7% correct**

***Crime Reduction***

**

Secondary icon

Different prompt and response scale: *The cross and tick figures here are each combined here with another icon below: the filled rectangles. What do you think it means that A has more filled rectangles?*

Responses: *A is more effective; A has higher-quality evidence; B is more effective; B has higher-quality evidence; A is more expensive; B is more expensive; don't know.*

**Expert: 54.5% correct** (*A has higher-quality evidence*). Most common wrong answer: *A is more effective.*

**Public: 47.1% correct**

***Early Interventions***

**

Secondary icon

Different prompt and response scale: *Which is the rating for the best evidence?*

Responses: *NL2; 4+; Don't know.*

**Expert: 97.5% correct**

**Public: 98.3% correct**


Secondary icon

Different prompt and response scale: *Please guess what this graphic means.*

Responses: *beneficial; very positive effect; high-quality evidence; considerable reliable evidence; multiple high-quality studies show a positive impact; don't know.*

**Expert: 66.4% correct** (*multiple high-quality studies show a positive impact; high-quality evidence* both marked correct). Most common wrong answers: *considerable reliable evidence.*

**Public: 55.6% correct**

**Location**

***Crime Reduction***

**

Secondary icon

**Expert: 81.1% correct**. Most common wrong answer: *effectiveness or impact*.

**Public: 84.5% correct**

**Mechanism**

***Crime Reduction***

**

Secondary icon

**Expert: 64.6% correct**. Most common wrong answer: *effectiveness or impact*.

**Public: 59.6% correct**

**Implementation**

***Crime Reduction***

**

Secondary icon

**Expert:** Most common answer: **56.3%** *how it works*. Note that the verbatim correct answer was accidentally not listed: "*How to do it"*. Most common wrong answer: *don't know*.

**Public:** Most common answer: **63.1%** *how it works*.

**Cost**

***Educational Endowment***

**

Secondary icon

**Expert: 98.5% correct**

**Public: 99.6% correct**

**

Secondary icon

Different prompt and response scale: *Please guess the correct label for this icon*.

Responses: *high cost; likely to be beneficial; multiple studies show a positive impact; moderate-quality evidence; some reliable evidence; don't know.*

**Expert: 97.4% correct**

**Public: 93.4% correct**