Author's response to reviews

Title: The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions

Authors:

Susan Michie (s.michie@ucl.ac.uk)
Maartje van Stralen (MM.vanStralen@vumc.nl)
Robert West (robertwest100@googlemail.com)

Version: 5 Date: 7 February 2011

Author's response to reviews: see over
Referee 1

1. Many commentators may quibble about the conceptual analysis in the ‘COM-B system’ (e.g., Wood’s account of habitual control would hardly countenance an analysis in terms of ‘motivation’) and proponents of particular models are likely to question the accuracy or fairness of the characterisation of their favoured framework in Table 4. Such quibbles and questions risk missing the key point, however; the behaviour change wheel does more to integrate ideas in this field than any previous research.

The referee is not proposing any changes here but it is worth noting that the comment about motivation may reflect an interpretation of the term to mean purposeful or goal-directed behaviour (which is certainly one of the uses) but in standard psychological usage it refers to all brain processes that energise and direct behaviour (as in the book by Doug Mook called Motivation). We have added the following text: “Capability is defined as the individual's psychological and physical capacity to engage in the activity concerned. It includes having the necessary knowledge and skills. Motivation is defined as all those brain processes that energise and direct behaviour, not just goals and conscious decision-making; it includes habitual processes, emotional responding and analytical decision-making. Opportunity is defined as all the factors that lie outside the individual that make the behaviour possible or prompt it.”

We also take the point that proponents of existing frameworks will not necessarily agree with our characterisation of them; however, this is the first time that such a systematic evaluation has been conducted and so we are not merely basing criticisms on opinion.

2. The only discretionary revision that I would suggest for an otherwise excellent paper concerns the title – is ‘re-inventing the wheel’ a method? It might be better to have ‘The Behavior Change Wheel’ as the prefix (or better still, ‘The Behavior Change Circumplex’ which might more fully capture the levels of analysis involved and the inter-relations among levels. The affect circumplex and the values circumplex had revolutionary effects in organising research on emotion and values, respectively; the same could be true for the Behaviour Change Circumplex).

We have changed the title to read: “The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions”

Referee 2

1. The paper is really two parts: (i) a "systematic review" to identify existing frameworks, and (ii) a synthesis and extension to create the behaviour change wheel (BCW) framework. The latter is intuitively very appealing, and the systematic review suggests it is more comprehensive and coherent than other frameworks. The authors tested this new framework on two areas - obesity and tobacco control - but give us little of the details of this testing. The combined parts make for a rather heavy and abstract but worthwhile paper. I found it a little lopsided, as the BCW is really the central product but I'd have like to see more details and application. However, that might require a second paper or splitting this one.

We agree with this. We are currently conducting more applied work, but including this would have lengthened an already long paper. We considered omitting the
application in this paper but thought it important to give some data to demonstrate the usability and reliability of the framework.

Minor Essential Revisions
1. Table 3a, 3b and 4 - the notation here is ambiguous. Maybe better to use a tick or even a tick and cross.

   We have changed crosses to ticks

2. Figure 2 - the elements in the outer wheel (Policy categories) seem out of alignment with the intervention functions, e.g., shouldn't regulation be near restriction? Fiscal near incentivisation? Etc

   We have re-drawn Figure 2 to better illustrate that each segment can be aligned with any of the others in the adjoining ring, depending on the COM-B analysis. Thus the different rings can be independently rotated. We think the new diagram conveys this better.

3. The analysis of the obesity and tobacco control frameworks is not presented. That would have been useful to make this less abstract.

   We could add this to the Additional Material. We have not done this as there are already five pieces of material covering 10 pages. However, we are happy to add this if the Editor thinks this would be a good idea.

Referee 3

1. It is not clear out of all the possible criteria these three were selected to both evaluate existing models, and to use for judging the utility of this new model. These seem to 'make sense' but they seem a bit arbitrary…. And frankly sounds more like they were post hoc criticisms of existing models rather than following from any given framework or meta-position. (This would be fine if this was the case, but I just had a hard time understanding were these 3 criteria- and not others came from). Why not include parsimony, ability to explain outcomes that other frameworks are not, usefulness for generating new, context-sensitive interventions, ability to predict effectiveness of intervention, etc.?

   We agree that other criteria are potentially important. There are two sets of criteria: one set that relates to the frameworks themselves and one that relates to their application. Re. the first set, we limited the criteria to those we considered to form a basis for judging adequacy; parsimony, on the other hand, is a desirable feature but it’s not something that you could set a threshold below which a framework shouldn’t fall. Comprehensiveness, coherence and linkage to a model of behaviour are all ones that could form the basis for a judgment of adequacy.

   We have added the following text to make this point: We limited the criteria to those we considered to form a basis for judging adequacy. There are others, e.g. parsimony, that are desirable features but do not lend themselves to thresholds. Other criteria can be used to evaluate its applicability, e.g. reliability, ease of use, ease of communication, ability to explain outcomes, usefulness for generating new interventions and ability to predict effectiveness of interventions

   Re. the second set, we refer to three criteria of application on p.9 and have added to this three from the reviewer’s list. The new text is: “We limited the criteria to those
we considered to form a basis for judging adequacy. There are others, e.g., parsimony, that are desirable features but do not lend themselves to thresholds. Other criteria can be used to evaluate its applicability, e.g., reliability, ease of use, ease of communication, ability to explain outcomes, usefulness for generating new interventions and ability to predict effectiveness of interventions.”

2. Relatedly, the second criteria of ‘coherence’ seems to be ‘in the eye of the beholder’ or raters. It is not clear why interventions cannot be of different levels of specificity? I wonder if others who do not share the authors' perspectives- or level of understanding of the field would make similar ratings or judgments about coherence.

We agree that this is a conceptually difficult issue and ideally would be argued thoroughly. Thus, for example, having categories at very different levels of specificity is problematic because it can lead to a failure to make important distinctions and a focus on trivial distinctions. As a reductio ad absurdum, consider a classification system in which 90% of exemplars fell into one category and 10% fell into 20 other categories by virtue of the first being very broad and the others very specific. This would not be a coherent classification system. Similarly, having categories that are subgroups of very different superordinate categories leads to a confused and incomplete analysis of the domain in question. A wonderful example of this is the ancient Chinese classification of animals which we now include to make the point more vividly (and perhaps humorously!).

The text added is as follows: “A beautiful example of an incoherent classification system is the Ancient Chinese Classification of Animals: “(a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, and (n) those that resemble flies from a distance” (Luis Borges ‘Other Inquisitions: 1937-1952”).

3. The usefulness of the model would be much greater if it could be used to make some a priori predictions, in addition to just classifying interventions. For example, would the model predict that using components from all three ‘levels’ of the ‘wheel would be more effective; or more sustainable than those with fewer?

Agreed. Current work is doing just this.

4. It would seem that some criterion of the level to which interventions are context sensitive; or adapt to context over time, and how these components are integrated, or how evidence-based they in fact are would enhance the usefulness of the system.

We agree. The hope (and our plan) is that future empirical work will test out these, and related ideas.

5. The various categories and levels vary greatly in their breadth- and in their cost- from a public health perspective, this is very important, but it is not clear how or if the new model addresses these issues.

The model as currently developed does not address these issues; we hope future empirical application will do this.
6. The introduction to the article—prior to stating the purpose on page 8 is too long, and seems—without the context of purpose to be a long trashing of alternative and existing approaches—but without a systematic context, or understanding of why or how the authors are being so critical of these other models. This section could be considerably condensed and re-organized so that some of it included in the review of existing frameworks.

We think that the addition (in Additional Materials) of the Table summarising reasons for excluding 16 identified frameworks addresses this problem. We thought it necessary to give a full explanation of the reasons for introducing a new model and why it is a qualitative improvement on preceding version. We note the other two Reviewers did not consider that the introduction to the article was too long.

More Minor issues; Clarifications needed:

1. Agreement or reliability should be assessed by a more rigorous method than percent agreement, such as kappa or other indices that correct for capitalization on chance and base rates.

We appreciate that kappas are often used as a way of assessing inter-rater agreement that is supposed to adjust for chance levels of performance. However, the idea that it does this is questionable. This is quite nicely stated in the following from a website dedicated to analysing inter-rater agreement: ‘Kappa’s calculation uses a term called the proportion of chance (or expected) agreement. This is interpreted as the proportion of times raters would agree by chance alone. However, the term is relevant only under the conditions of statistical independence of raters. Since raters are clearly not independent, the relevance of this term, and its appropriateness as a correction to actual agreement levels, is very questionable. Thus, the common statement that kappa is a “chance-corrected measure of agreement” is misleading. As a test statistic, kappa can verify that agreement exceeds chance levels. But as a measure of the level of agreement, kappa is not “chance-corrected”; indeed, in the absence of some explicit model of rater decision-making, it is by no means clear how chance affects the decisions of actual raters and how one might correct for it.’

Given chance agreement would in any event be extremely low (there being 9 intervention categories and 7 policy categories) we feel that percentage agreement is a readily interpretable figure that avoids the above issue. Of course if the editor insists, we will present the kappas but we would prefer not to.

2. It is unclear the basis on which the authors excluded over 99% of the potentially relevant articles they located.

We have added a Figure of the flow of studies through the review process to the Additional Materials. We have also included in the Additional Materials a Table of the specific reasons that 16 full texts, selected at abstract stage, were excluded.

3. Pg. 9- The reference to U.S. criminal law seems strange— in general the law and behavioral science are two quite different systems, use vastly different criteria, approaches to evidence, etc. I am not sure that appeal to three criteria that appear—on the surface—to be similar—add much to the argument.

We considered it a strength of the model that it had been independently produced within behavioural science and within legal thought. We made this point on p.11
“The commonality of conclusion from these two separate strands of thought lends confidence to this model of behaviour.”

4. If intervention mapping is included, it is unclear why related paradigms or frameworks such as PRECEDE-PROCEED were not also included.

We have added to the Additional materials, a table of the 16 excluded frameworks, and the reasons for exclusion. For PRECEDE-PROCEED, the reason was “Intervention planning model, no intervention techniques are described.” Thus PRECEDE-PROCEED, like some of the other models, is very valuable (as its extensive application has shown) but it is not a taxonomy of interventions.