The Question-Behavior Effect: Genuine effect or Spurious Phenomenon? A systematic review of randomized controlled trials with meta-analyses

DOI:
10.1037/hea0000104

The Question-Behavior Effect: Genuine effect or Spurious Phenomenon?

A systematic review of randomized controlled trials with meta-analyses

Angela M. Rodrigues[1], Nicola Hobbs[1], David P. French[2], Liz Glidewell[3] & Falko F. Sniehotta[1]

1 Newcastle University; 2 University of Manchester; 3 University of Leeds

Angela M. Rodrigues and Nicola Hobbs are joint first authors.

In press: *Health Psychology*

Acknowledgements:
The authors thank Information Specialist Judy Wright for her expertise developing the search strategy and all authors who provided missing data. Angela Rodrigues is funded by a PhD fellowship from the Portuguese Science and Technology Foundation (FCT). Falko F Sniehotta is funded by Fuse, the Centre for Translational Research in Public Health, a UK Clinical Research Collaboration Public Health Research Centre of Excellence based on funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council and the National Institute for Health.

Abstract

Background: Simply answering questions about a specific behavior may change that behavior. This is known as the mere measurement effect or the question-behavior effect (QBE).

Purpose: To synthesize the evidence for the QBE on health-related behaviors.

Methods: Included studies were randomized controlled trials which tested the effect of questionnaires or interviews about health-related behaviors and/or related cognitions compared with a no measurement control condition or with another form of measurement. Subgroup analyses were conducted to identify potential moderators.

Results: Thirty-eight papers reporting 41 studies were included assessing a range of health behaviors. Meta-analyses showed a small overall QBE effect (SMD= 0.09; 95% CI= 0.04; 0.13; k=33). Studies showed moderate heterogeneity, variable risk of bias and evidence for publication bias. No dose-response relationships were found from studies comparing more with less intensive measurement conditions. There were no significant differences in QBE by behavior, but QBE effects for dental flossing, physical activity and screening attendance were significantly different from zero. Findings were not altered by whether behavior or cognitions were measured; whether or not attitudes were measured; whether studies used questionnaires or interviews; or whether outcomes were taken objectively or by self-report.

Conclusions: There is some evidence for the QBE in relation to health-related behavior. However, risk of bias within studies and evidence of publication bias indicates that the observed small effect size may be an over-estimate, especially given that some studies also included intervention techniques in addition to just providing questionnaires. Pre-registered high quality trials with clear specification of intervention content are needed to confirm if and when measurement leads to behavior change.

Introduction

Measuring health-related behavior and/or related cognitions may change the behavior under investigation. This has been called the mere measurement effect (Morwitz, Johnson, & Schmittlein, 1993; Sherman, 1980) or, more recently, the "question-behavior effect (QBE)" (Ayres et al., 2013; French & Sutton, 2010; Godin, Bélanger-Gravel, Vézina-Im, Amireault, & Bilodeau, 2012). The QBE has been reported for different types of behavior including consumer and voting behavior (Chapman, 2001; Morwitz & Fitzsimons, 2004; Spangenberg, Sprott, Grohmann, & Smith, 2003). More recently, several studies have examined the QBE on health behaviors such as physical activity, blood donation and cervical screening (Godin, Sheeran, Conner, & Germain, 2008; Sandberg & Conner, 2009; Spence, Burgess, Rodgers, & Murray, 2009). However, evidence for the QBE is not consistent across studies. For example, whilst some studies have shown that answering questions about safe sex behaviors affects subsequently measured safe sex behaviors (Knaus, Pinkleton, & Weintraub Austin, 2000), other studies have not found such effects (Kvalem, Sundet, Rivø, Eilertsen, & Bakketeig, 1996).

Investigation of the QBE on health-related behaviors is important for research as well as for evidence-based practice in healthcare (French & Sutton, 2010). The positive implications of the QBE on behavior for healthcare practice is that many forms of measurement, such as self-report questionnaires, are inexpensive and could be distributed widely. If their completion is found to lead to desirable changes in behavior, then distributing questionnaires could potentially be a viable and cost effective public health intervention. The implications for healthcare research are more challenging. In intervention trials, baseline assessment may affect behavior in a similar way as effective interventions affect behavior. For example, baseline questions about alcohol consumption may increase awareness and subsequently reduce instances of binge drinking because participants may realize that their alcohol intake is excessive through their interaction with a questionnaire. Therefore, in trials where an intervention designed to reduce drinking behavior is tested against a control condition, baseline assessment may mask or reduce observed intervention effects

(McCambridge & Kypri, 2011). Moreover, in some trials, individuals allocated to an intervention group could receive different forms of measurement in order to tailor intervention components to participants. In this case, it may be difficult to disentangle measurement and intervention effects.

The QBE can also limit the external validity of a trial. For example, baseline measurement may stimulate a participant to deliberate about behavior increasing their motivation to engage with the intervention. To better understand the potential interaction between baseline measurement and intervention effects, more sophisticated factorial trial designs are useful, such as the Solomon four-group design. In this design participants are allocated to receive baseline measurement or not to receive baseline measurement, and to receive the intervention or not to receive the intervention (McCambridge, Butor-Bhavsar, Witton, & Elbourne, 2011).

The primary aim of this systematic review was to assess the effect of measurement by asking questions about health-related behaviors on subsequent behavior. This was supplemented by subgroup analyses which examined whether there were differences in effects between studies characterized by lower risk of bias and those with higher risk of bias. This review also explored a possible dose-response relationship in the QBE and explored several possible moderators of effects: features of participants (student vs. other samples), interventions (type of measurement: questions about behavior and/or questions about cognitions; format of measurement: questionnaire vs. interview) and outcomes (type of behavior; objective vs. self-reported).

Methods

The protocol for this review was published in advance of the work commencing in the PROSPERO database (record number: CRD42011001467).

Inclusion criteria

Trials randomly allocating participants to measurement or no measurement control conditions or trials where groups were randomly allocated to different forms of measurement (i.e. differences in

length or content of measures) were included in this review. Studies were eligible for inclusion if they reported health-related behavior as outcomes, defined as behavior judged to reduce the risk or severity of diseases or promote health including preparatory behaviors, such as buying condoms or food (Marteau et al., 2010) . Studies that only reported predictors of behavior (e.g., intention or self-efficacy) as outcomes were excluded. The measurement condition could include assessments of cognitions, behavior, or cognitions and behavior by questionnaire (paper and pencil or online) or interview. Studies that used objective forms of measurement as interventions (e.g. pedometers, blood pressure monitors) were not eligible for inclusion. We included studies with any length of follow-up that reported either objectively assessed or self-reported health-related behaviors.

Search Strategy

The following electronic databases were searched from the earliest available date to December 2012: MEDLINE, Cochrane Central Register of Controlled Trials (CENTRAL), EMBASE and PsycINFO. ERIC database was searched until March 2011 (see Appendix 1). An iterative process was used to develop a sensitive and specific search strategy with guidance from an information specialist. The search included studies providing an English language title and abstract. Publications in any language were eligible for inclusion. Reference lists of included studies were reviewed for additional eligible studies and key authors in the research field were invited to provide any additional published literature that fulfilled the inclusion criteria.

Study Selection and Data Extraction

Two reviewers (AR and NH) independently screened all titles and abstracts to identify eligible studies. There was 100% agreement between the reviewers regarding which papers to retrieve for full text examination. Full texts were retrieved for 63 papers and the two reviewers independently assessed each study for eligibility based on the inclusion and exclusion criteria (kappa = 0.73). For five papers, the reviewers could not decide on inclusion and consensus was reached in discussion with a third reviewer (FFS). Data from each study were extracted independently by two reviewers

(AR and NH) into a data extraction form developed for this review. One reviewer (AR) entered data into RevMan Software (version 5.0) (Review Manager, 2011) and another reviewer (NH) independently verified entries. In cases where statistical data were missing, the authors were contacted and asked to make this data available to facilitate calculation of effect sizes.

Assessment of Risk of Bias and Critical Appraisal

Risk of bias was appraised using the Cochrane collaboration tool (Higgins & Green, 2011). For each of eight criteria (adequate sequence generation, allocation concealment, blinding (participants, personnel and assessors), incomplete outcome data addressed, free of selective outcome reporting, free of other bias) studies were categorized as low, unclear or high risk of bias, scoring 0, 1 or 2 respectively. An overall score between 0 and 16 was computed, where higher scores indicate higher risk of bias. For postal/online studies where no information was available about allocation concealment, studies were classified as 'low risk of bias' for those criteria. When information about blinding was not available and studies included an automated or online outcome assessment (including self-report), studies were classified as 'low risk of bias'. Risk of bias was assessed by two reviewers independently (AR and NH) resulting in very good overall agreement of kappa = 0.92 aggregated over all eight criteria.

Analytic strategy

Odds ratios (ORs) or standardized mean differences (SMDs) with 95% confidence intervals (CI) were calculated for all included studies, with the exception of two studies for which data were not available. Results from comparable studies were pooled using a random effects model (inverse-variance approach based on weighted odds ratios and weighted SMDs, calculated by RevMan version 5.0 software (2011). Dichotomous and continuous outcomes were merged using Comprehensive Meta Analysis software version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005) to produce SMD for all included studies. SMDs were used in all reported analyses, which are equivalent to Cohen's *d*. For behavioral outcomes with more than one time point assessed, data

reported at the first follow-up time point was used for meta-analyses. Where studies reported multiple behaviors as outcomes, the data were merged and the pooled effect was used for the main meta-analyses. Effect sizes for all outcomes were calculated. Heterogeneity across studies was assessed using Cochrane's *Q* statistic and I^2 test statistic to quantify the effect of heterogeneity (Higgins & Green, 2011) and I^2confidence intervals as suggested by Higgins and Thompson (2002).

The main comparison performed was measurement vs. no measurement conditions. Subgroup analyses were conducted to examine whether there were differences in effects on the basis of risk of bias. Studies were grouped into 'higher' and 'lower' risk of bias studies using a median cut-off split (Median = 3) of overall risk of bias score. A secondary comparison was conducted to identify a dose-response relationship comparing the most intensive measurement conditions with the least intensive measurement conditions (i.e. frequency/duration of assessment).

Subgroup analyses were also performed for the following pre-specified factors: features of participants (student vs. other samples), interventions (type of measurement: questions about behavior and/or questions about cognitions[1]; format of measurement: questionnaire vs. interview) and outcomes (type of behavior; objective vs. self-reported). The Cochran Q statistic was used to detect sources of heterogeneity in the subgroup analyses, and when a study had more than two conditions and a significant subgroup difference was observed, *Z* tests were used to determine between which groups the difference existed.

Publication bias was examined by plotting the inverse of the standard errors of effect estimates using 'funnel plots' to explore symmetry. These were assessed visually to see if the effect decreased with increasing sample size and there was evidence of considerable asymmetry. Egger's regression test (Higgins & Green, 2011) was used to formally test for the presence of publication bias.

[1] There were insufficient studies to allow meaningful comparisons for more specific comparisons between constructs.

This report follows the PRISMA guidance for reporting systematic reviews (Moher, Liberati, Tetzlaff, & Altman, 2009).

RESULTS

Description of included studies

Thirty-eight papers reporting 41 studies met the inclusion criteria. The paper by Conner, Godin, Norman, and Sheeran (2011) reported two studies and Levav and Fitzsimons' (2006) paper reported three studies. From the 41 studies, 33 were included in the main meta-analysis, five other studies were compared in the most intensive versus least intensive measurement meta-analysis and the remaining three studies were included narratively.

Figure 1 shows the flow diagram of the study inclusion and exclusion, providing reasons for exclusion[2]. The characteristics of included studies are displayed in Table 1.

[Insert Figure 1 and Table 1 here]

Participants

The review represents a total of 71,362 participants (Range: 31 – 7,008). Seventeen of the included studies involved student samples, with 16 studies including university students and one study with high school students. Fifteen studies took place in healthcare settings; three studies recruited in emergency departments, one in a treatment center for alcohol, one in a center for drug abuse, two in hospitals, three in blood donation agencies, and one in a central agency for cervical screening. Seven studies were conducted within community settings. One study included both community and university samples, and one study recruited participants in a health club.

Measurement manipulations

[2] Two of the included studies (Knaus & Austin, 1999; Knaus, et al., 2000) had to be excluded from the meta-analyses as statistical data were missing and could not be obtained after contact with authors.

Of the 41 studies in total, the majority (n=33) utilized questionnaires as the format of measurement, whilst seven used interviews and one used both questionnaires and interviews. In 14 studies, the measurement condition involved questions about the behavior under investigation. In 12 studies, the measurement condition involved questions about cognitions towards the health-related behavior. In the remaining 15 studies, the measurement condition consisted of questions about both behavior *and* related cognitions. For those studies assessing cognitions, ten used constructs abstracted from the Theory of Planned Behavior.

Outcomes: health-related behaviors

Outcomes included alcohol consumption (n=10) , physical activity (n=5), sex-related behaviors (n=5), blood donation (n= 4), cancer screening attendance (n=4), choice of low or high fat snacks (n=2), dental flossing (n=2), attendance for a health assessment (n=2), uptake of a health plan (n=1), health club attendance (n=1), participation in chlamydia screening (n=1), vaccination uptake (n=1), medication adherence (n=1), and hand washing (n=1). One study assessed and reported multiple behaviors as outcomes, including fruit and vegetable consumption, alcohol consumption and physical activity frequency (Kypri & McAnally, 2005). The majority of studies reported self-reported outcomes (n=29) whilst 12 studies reported objectively assessed outcomes. Outcomes were reported both as a dichotomous measures (n=19) and continuous measures (n=22).

Risk of bias

Table 1 shows risk of bias scores for each included study in this review. Overall there was considerable risk of bias. Eighteen studies reported adequate random sequence allocation of participants to conditions. Twenty-one studies were considered to have utilized appropriate procedures for allocation concealment. Thirty studies stated numbers and reasons for participant dropout or used adequate methods to deal with incomplete outcome data. Six studies had considerable risk of attrition bias. Reporting bias was not a risk for 29 studies, but was considered to be a problem for 12 studies. Nineteen studies stated that participants were blinded to their

allocation. Twenty-four studies reported effective blinding procedures for outcome assessors and 21 studies for intervention providers. It was unclear whether 'other' risk of bias was present in four studies due to missing baseline information about groups/participants (n=2) or information about how the outcome measure was computed (n=2). Only one study (Moreira & Foxcroft, 2008) was pre-registered on a public database, a key requirement of the CONSORT guidance (Schulz, Altman, & Moher, 2010).

[Insert Figures 2 and 3 here]

Does answering questions change behavior?

Comparison of studies with measurement v no measurement conditions (k=33)

For n=33 studies comparing measurement and no measurement conditions, there was an overall small but significant QBE (Figure 3: SMD= 0.09; 95% CI= 0.04; 0.13; n= 37452). Statistical heterogeneity was moderate with an I^2 of 44% and a Q of 57.39 (95% CI = 15.7%, 63.1%; df=32, p=0.004).

Two additional studies did not provide sufficient information for meta-analysis. No significant difference was identified between participants randomized to measurement or no measurement conditions in these studies (Knaus & Austin, 1999; Knaus, et al., 2000).

Long term effects (k=4)

In addition to the Moreira et al (2012) study which only assessed relevant outcomes at 12 months and was entered in the main meta-analysis, three further studies reported additional outcomes at 12 months. In line with Moreira et al (2012), Carey et al. (2006), Godin et al. (2010) and Kvalem et al. (1996) did not find QBE at 12 months. Only Godin et al. (2008) found a sustained significant QBE at 12 months (SMD=0.08, 95% CI = 0.02, 0.14; n= 6835).

Publication bias

Egger's regression test shows that there was significant evidence of publication bias (p=0.01; illustrated in Figure 2). Under the assumption of a normal distribution of effect sizes, there was evidence that studies with smaller or no effects were less likely to be published.

Subgroup analysis by risk of bias of trials

There was no evidence that effects were moderated by risk of bias. There was a significant effect in favor of the measurement condition for studies with a lower risk of bias (SMD=0.10, 95% CI=0.06 to 0.14; I^2=39%, 95% CI= 0.0%, 66.2%; k= 16; n= 32908) and a non-significant effect for studies with a higher risk of bias (SMD=0.07, 95% CI=-0.03 to 0.17; I^2=48%, 95% CI= 8.0%, 70.3%; k= 17; n= 4660). Q-test shows that there were no significant differences between subgroups (Q=0.35, p=.55) by risk of bias.

Comparison of most intensive versus least intensive measurement (k=5)

Meta-analysis of five trials comparing conditions with different intensity of measurement did not find a difference between the most intensive measurement conditions (e.g. brief screening plus full assessment; repeated assessments points) and the least intensive measurement conditions on health-related behaviors (SMD= 0.02, 95% CI=-0.28; 0.33; n= 1262). Statistical heterogeneity was high with an I^2 of 84% and a Q of 25.14 (95% CI= 64.1%, 92.9%; df=4, p<0.001).

Possible moderators of the QBE

1. *Type of participants*

Subgroup analysis comparing student and non-student samples showed small significant QBEs in both, student samples (SMD = 0.17, 95% CI = 0.01, 0.32) and non-student samples (SMD = 0.07,

95% CI = 0.04, 0.11). The difference was not significant between subgroups (*Q*=1.38, *p*=.24) (Table 3).

[Insert Table 2 here]

2. *Interventions: content of measurement*

Subgroup analysis showed no significant effect in favor of measurement condition when only behavior was measured (SMD = 0.11, 95% CI = -0.09, 0.30); a small significant effect when only cognitions were measured (SMD = 0.10, 95% CI = 0.05, 0.15); and no significant effect when both behavior and cognitions were measured (SMD = 0.05, 95% CI = -0.04, 0.14) (Table 2). No significant difference between subgroups was identified (*Q*=1.19, *p*=.55).

2.1. *Interventions: measurement of attitudes*

Subgroup analysis showed no differences (*Q*=0.00, *p*=.98) between measurement conditions when attitudes were measured (SMD = 0.09, 95% CI = 0.05, 0.13) and when no attitudes were measured (SMD = 0.09, 95% CI = 0.01, 0.18) with both subgroups showing significant QBEs on health-related outcomes (Table 2).

3. *Interventions: format of measurement*

A small significant effect in favor of the measurement condition was identified when using questionnaires (SMD = 0.10, 95% CI = 0.05, 0.15) but not when using interviews (SMD = 0.03, 95% CI = -0.06, 0.12); however, no significant difference between subgroups was identified (*Q*=2.02, *p*=.15) (Table 2). An additional study that tested the effect of using a questionnaire and an interview separately and thus could not be meta-analyzed as it was not comparable to other studies, (Kalichman, Kelly, & Stevenson, 1997) found no difference between these two modes of assessment on sexual behavior (OR = -0.10, 95% CI = -0.79, 0.59).

4. Outcomes: type of health-related behavior

For dental flossing behavior, a significant medium size effect was found in favor of the measurement condition (SMD = 0.50, 95% CI = 0.18, 0.81). Small but significant effects were also found for physical activity (SMD = 0.20, 95% CI = 0.08, 0.32) and screening attendance (SMD = 0.06, 95% CI = 0.003, 0.12). No effects were found for blood donation (SMD = 0.05, 95% CI = -0.00, 0.10), alcohol consumption (SMD = 0.04, 95% CI = -0.08, 0.16), dietary (SMD = 0.08, 95% CI = -0.68, 0.84) or sexual behaviors (SMD = 0.05, 95% CI = -0.20, 0.31). However, no significant differences between subgroups were identified (Table 2) (Q=13.96, *p*= .052);

5. Outcomes: type of measurement

Small significant effects were found for both objective outcome measures (SMD = 0.08, 95% CI = 0.04, 0.13) and self-report measures of behavior (SMD = 0.10, 95% CI = 0.01, 0.19) (Table 2). There were no differences between subgroups (*Q*=0.14, *p*=.71).

DISCUSSION

This is the first systematic review with meta-analysis synthesizing evidence for the effects of completing measures by questionnaire and interview on health-related behaviors. Previous reviews with more optimistic conclusions were not systematic and did not focus on health-related behaviour (Dholakia, 2010; Sprott, Spangenberg, Knuff, & Devezer, 2006). A recent review on the Hawthorne effect (McCambridge, Witton, & Elbourne, 2014) showed that there is some evidence for participants being aware that they are being studied for the behaviors being investigated, but it did not focus on measurement specifically.

We found evidence of a typically small but significant QBE on health-related behaviors with moderate levels of heterogeneity of effects. Studies comparing more with less intensive measurement conditions did not suggest dose-response relationships. Subgroup analyses were conducted to identify potential moderators of effects. There were no significant differences in QBE

by behavior, but QBE effects for dental flossing, physical activity and screening attendance were significantly different from zero. These findings were not altered in studies where students or other samples were studied; cognitions, behavior or both were measured; attitudes were measured or not measured; questionnaires or interviews were used; or outcomes were taken objectively or as self-reports. After the completion of this review, a new trial was published comparing five different measurement conditions (intention only, interrogative intention, intention plus moral norm, intention plus regret and intention plus self-positive image) and one implementation intention intervention with a no intervention control condition (Godin, Germain, Conner, Delage, & Sheeran, 2013b). The comparison between the five measurement conditions and the control condition yielded an aggregated small effect size of 0.16 (95% CI = 0.09, 0.23). This effect is slightly higher that the main effect size found in the present meta-analysis.

Three key findings of this review need to be highlighted, which may suggest some caution regarding the evidence for the QBE. Firstly, methodological quality of the included studies was variable and several studies showed considerable risk of bias, in particular due to selective reporting (outcomes which suggest a significant QBE might be more likely to be reported), lack of blinding of participants (knowledge of allocation may affect question elaboration or desirability bias in self-reported outcomes) and incomplete outcome data not appropriately addressed. Only seven of the 33 studies entered in the main meta-analysis explicitly stated conducting intention-to-treat analysis, thus introducing the risk that loss to follow-up in different trial arms might differ in terms of numbers or participant features. Higher effects were found in studies with a greater risk of bias but this difference was not statistically significant. It cannot be ruled out that the already small effects found in this review are inflated through systematic methodological bias in the included trials.

Secondly, there was evidence of publication bias. Randomly allocating participants to varying forms of measurement is an inexpensive addition to a range of study designs and implemented for a range of reasons. It is possible that studies with random measurement allocation are less likely to be reported in the published literature, if the different measurement conditions do not result in

differences in behavior (Dwan et al., 2008). In this case, the small effects found in this review might be an artifact of publication bias. With the exception of one study (Moreira & Foxcroft, 2008), which was pre-registered and for which a full protocol has been published (and reported subsequently a null finding), none of the trials included in this review were pre-registered. Thus, there are no safeguards to ensure that comparisons, outcomes and analyses were specified a-priori and that the studies achieved the target sample-size that based on an a priori power analysis. Thirdly, intervention procedures are often insufficiently described and therefore it is difficult to conclude that the measurement conditions in this review were not confounded with other procedures potentially affecting outcomes. For example, it is good practice in survey research to send reminders to those who do not respond to an initial questionnaire (McColl, Jacoby, Thomas, Soutter, & Bamford, 2002). In question-behavior effect studies, larger response rates are thought to lead to higher reactivity effects as more participants engage with the questions (Spence, et al., 2009). Three large randomized controlled trials of measurement on blood donation were included in this review (Godin et al., 2010; Godin, et al., 2008; van Dongen, Abraham, Ruiter, & Veldhuizen, 2012). The Van Dongen et al (2012) and Godin et al (2010) trials showed that completing questionnaires did not change blood donations in two Dutch and one Canadian sample, which is in contrast with the Godin et al (2008) trial that showed a significant effect on blood donations. In their 2008 study, Godin and colleagues sent reminders and 'thank you' notes to participants in the measurement condition, resulting in a return rate of 82%. By contrast the Van Dongen and Godin (2010) trials did not send reminders and observed a return rate of 64-65% and 49.5% respectively. It is impossible to conclude if these procedures relate to QBEs due to the poor standard of reporting in some studies, and the field would benefit from full reporting of procedures and response rates in future studies on QBEs. Based on these considerations, the QBE seems to be influenced by an accumulation of sources of bias in trials, failure to published trials with null findings and reporting trial procedures in insufficient detail.

Findings for alcohol consumption differed slightly from those reported in a recent review of measurement reactivity effects in trials of brief alcohol interventions (McCambridge & Kypri, 2011), which found that measurement does affect Alcohol Use Disorders Identification Test (AUDIT: (Bush, Kivlahan, McDonell, Fihn, & Bradley, 1998)) measures but not other measures of consumption. Our review does not find an overall effect of measurement on alcohol consumption. Differences between both reviews are in the aggregation of outcome data between the AUDIT and other measures of consumption and in the exclusion of one trial in this review which did not use a randomized controlled design (Richmond, Heather, Wodak, Kehoe, & Webster, 1995).

Implications for research and practice

The current evidence base is characterized by variable methodological quality and publication bias. Despite 41 randomized trials in this review, it is still not clear whether the QBE exists, largely due to it being unclear how many unpublished negative trials exist. In our view, the best way of making progress in this area, we strongly recommend to journals the principle of publishing QBE trials only if study protocols have been pre-registered. There are at least three good reasons for this (Dickersin & Rennie, 2003). First, estimates of effect sizes based on the results of trials registered in advance are more likely to be free from publication bias. As previously noted, there are good reasons to suspect that publication bias is a particular issue for this area of research, and the results of this review provide empirical support for this suspicion. Second, the non-publication of trials which find null results would be apparent. Third, pre-registration of trials would allow scrutiny of how analyses reported were different from those pre-specified, which would reduce bias introduced through selective reporting of outcomes, and increase transparency when this occurs. It appears that the current practice of publishing studies that have not had protocols pre-registered produces biased estimates of the QBE. There is no reason to think that the continuation of this practice would result in a different outcome: we are likely to end up with a large body of trials with publication bias and selective reporting.

From a theoretical perspective, there is not sufficient evidence to date to allow synthesizing the effects for different theoretical measures and possible mechanisms at this stage. The majority of the studies assessing cognitions used questionnaires abstracted from the Theory of Planned Behavior (Ajzen, 1991). Studies using 'think aloud' technique (Darker & French, 2009; French, Cooke, McLean, Williams, & Sutton, 2007) have shown that using questionnaires based on the 'Theory of Planned Behavior' can result in participants forming beliefs about topics which they have previously devoted little thought. This may thereby increase the salience of beliefs about specific features or aspects of performing that behavior (Morwitz & Fitzsimons, 2004). In a similar way, measurement can also form attitudes towards the behavior itself and/or make specific aspects of performing a behavior more accessible, thereby fostering performance (Morwitz & Fitzsimons, 2004). It is possible that the mere fact of being measured influences the formation of judgments and/or accessibility of these for respondents (Chandon, Morwitz, & Reinartz, 2005). Research comparing QBEs for different theoretical measures and/or different constructs has been published in recent years (Conner, Godin, Norman, & Sheeran, 2011; Godin, et al., 2008) and it is likely that these comparative trials will enhance our understanding of if, how and when measurement changes behavior. The range of cognitive measures investigated to date has predominantly focused around constructs abstracted from the Theory of Planned Behavior as well as on anticipated regret. Other measures such as identity (van Dongen et al., 2012), self-image (Godin, Germain, Conner, Delage, & Sheeran, 2013a) and more emotion-related measures such as worry may deserve additional attention in future research. Effects may also differ due to features of the study population and the period of follow-up (Godin, et al., 2013a).

Current evidence of small effects with moderate heterogeneity suggests that it might be worthwhile to estimate small increases in control conditions when establishing the required sample size for randomized trials. To date there is no compelling evidence for baseline measurement by intervention interaction effects from Solomon trials (cf. McCambridge, et al. (2011)), suggesting

that there might not be a systematic bias in the evidence base about behavior change interventions as a result of baseline measurement in trials.

Implications for practice are more difficult to identify at this stage. The evidence for sending questionnaires to increase behavioral uptake is limited. However, first robust evidence for a QBE has to be accumulated. Second, before the QBE should be used as a behavior change strategy, it has to be shown to not only exist, but also to produce greater changes in behavior than simply sending reminders to perform the behavior.

In summary, this systematic review advances the field by 1) providing a comprehensive synthesis of the evidence; 2) including evidence from various health-related behaviors; 3) providing quantification of effects sizes with moderator analyses; and 4) identifying and critically appraising potential sources of systematic bias. Small QBEs were found with moderate heterogeneity between studies. Future QBE trials should focus on reducing risk of bias and providing detailed description of procedures in each trial arm. Pre-registration of trials is paramount to allow a more precise assessment of measurement reactivity.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179-211. doi: http://dx.doi.org/10.1016/0749-5978(91)90020-T

Ayres, K., Conner, M., Prestwich, A., Hurling, R., Cobain, M., Lawton, R., & O'Connor, D. B. (2013). Exploring the question-behaviour effect: Randomized controlled trial of motivational and question-behaviour interventions. *British Journal of Health Psychology, 18*(1), 31-44. doi: 10.1111/j.2044-8287.2012.02075.x

Bernstein, J., Heeren, T., Edward, E., Dorfman, D., Bliss, C., Winter, M., & Bernstein, E. (2010). A brief motivational interview in a pediatric emergency department, plus 10-day telephone follow-up, increases attempts to quit drinking among youth and young adults who screen positive for problematic drinking. *Academic Emergency Medicine, 17 (8)*, 890-902.

Berry, T. R., & Carson, V. (2010). Ease of imagination, message framing, and physical activity messages. *British Journal of Health Psychology, 15*(1), 197-211.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-analysis (Version 2). Englewood, NJ: Biostat.

Bush, K. R., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., & Bradley, K. A. (1998). The audit alcohol consumption questions (audit-c): An effective brief screening test for problem drinking. *Archives of internal medicine, 158*(16), 1789-1795. doi: 10.1001/archinte.158.16.1789

Carey, K. B., Carey, M. P., Maisto, S. A., & Henson, J. M. (2006). Brief motivational interventions for heavy college drinkers: A randomized controlled trial. *Journal of consulting and clinical psychology, 74*(5), 943-954.

Chandon, P., Morwitz, V. G., & Reinartz, W. J. (2005). Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research. *Journal of Marketing, 69*(2), 1-14. doi: 10.1509/jmkg.69.2.1.60755

Chapman, K. J. (2001). Measuring intent: There's nothing "mere" about mere measurement effects. *Psychology and Marketing, 18*(8), 811-841. doi: 10.1002/mar.1031

Cherpitel, C. J., Korcha, R. A., Moskalewicz, J., Swiatkiewicz, G., Ye, Y., & Bond, J. (2010). Screening, Brief Intervention, and Referral to Treatment (SBIRT): 12-month outcomes of a randomized controlled clinical trial in a Polish emergency department. *Alcoholism: Clinical and Experimental Research, 34*(11), 1922-1928.

Cioffi, D., & Garner, R. (1998). The Effect of Response Options on Decisions and Subsequent Behavior: Sometimes Inaction is Better. *Personality and Social Psychology Bulletin, 24*(5), 463-472. doi: 10.1177/0146167298245002

Clifford, P. R., Maisto, S. A., & Davis, C. M. (2007). Alcohol treatment research assessment exposure subject reactivity effects: Part I. Alcohol use and related consequences. *Journal of Studies on Alcohol and Drugs, 68 (4)*, 519-528.

Conner, M., Godin, G., Norman, P., & Sheeran, P. (2011). Using the question-behavior effect to promote disease prevention behaviors: two randomized controlled trials. *Health psychology,* (3), 300-309. Retrieved from doi:10.1037/a0023036

Daeppen, J.-B., Gaume, J., Bady, P., Yersin, B., Calmes, J.-M., Givel, J.-C., & Gmel, G. (2007). Brief alcohol intervention and alcohol assessment do not influence alcohol use in injured patients treated in the emergency department: a randomized controlled clinical trial. *Addiction, 102*(8), 1224-1233. doi: 10.1111/j.1360-0443.2007.01869.x

Darker, C. D., & French, D. P. (2009). What sense do people make of a theory of planned behaviour questionnaire?: A think-aloud study. *Journal of Health Psychology, 14*(7), 861-871. doi: 10.1177/1359105309340983

Dholakia, U. M. (2010). A critical review of question–behavior effect research. *Review of Marketing Research, 7*, 145-197.

Dickersin, K., & Rennie, D. (2003). REgistering clinical trials. *JAMA, 290*(4), 516-523. doi: 10.1001/jama.290.4.516

Dignan, M., Michielutte, R., Blinson, K., Wells, H. B., Case, L. D., Sharp, P., . . . McQuellon, R. P. (1996). Effectiveness of health education to increase screening for cervical cancer among eastern-band Cherokee Indian women in North Carolina. [Clinical Trial; Randomized Controlled Trial; Research Support, U.S. Gov't, P.H.S.]. *Journal of the National Cancer Institute, 88*(22), 1670-1676.

Dignan, M., Michielutte, R., Wells, H. B., Sharp, P., Blinson, K., Case, L. D., . . . McQuellon, R. P. (1998). Health education to increase screening for cervical cancer among Lumbee Indian Women in North Carolina. *Health Education Research, 13*(4), 545-556. doi: 10.1093/her/13.4.545

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., . . . Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE, 3*(8), e3081. doi: 10.1371/journal.pone.0003081

French, D. P., Cooke, R., McLean, N., Williams, M., & Sutton, S. (2007). What Do People Think about When They Answer Theory of Planned Behaviour Questionnaires?: A `Think Aloud' Study. *Journal of Health Psychology, 12*(4), 672-687. doi: 10.1177/1359105307078174

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology, 15*(3), 453-468. doi: 10.1348/135910710x492341

Godin, G., Bélanger-Gravel, A., Amireault, S., Vohl, M. C., & Pérusse, L. (2011). The effect of mere-measurement of cognitions on physical activity behavior: a randomized controlled trial

among overweight and obese individuals. *The International Journal of Behavioral Nutrition and Physical Activity*, 2. doi: 10.1186/1479-5868-8-2

Godin, G., Bélanger-Gravel, A., Vézina-Im, L.-A., Amireault, S., & Bilodeau, A. (2012). Question–behaviour effect: A randomised controlled trial of asking intention in the interrogative or declarative form. *Psychology & Health, 27*(9), 1086-1099. doi: 10.1080/08870446.2012.671617

Godin, G., Germain, M., Conner, M., Delage, G., & Sheeran, P. (2013a). Promoting the Return of Lapsed Blood Donors: A Seven-Arm Randomized Controlled Trial of the Question–Behavior Effect.

Godin, G., Germain, M., Conner, M., Delage, G., & Sheeran, P. (2013b). Promoting the Return of Lapsed Blood Donors: A Seven-Arm Randomized Controlled Trial of the Question–Behavior Effect. *Health Psychology*, In Press.

Godin, G., Sheeran, P., Conner, M., Delage, G., Germain, M., Belanger-Gravel, A., & Naccache, H. (2010). Which survey questions change behavior? Randomized controlled trial of mere measurement interventions. *Health Psychology, 29 (6)*, 636-644.

Godin, G., Sheeran, P., Conner, M., & Germain, M. (2008). Asking questions changes behavior: Mere measurement effects on frequency of blood donation. *Health Psychology, 27*(2), 179-184.

Higgins, J. P. T., & Green, S. (2011). Handbook for Systematic Reviews of Interventions Version 5.1. 0 [updated March 2011]. *The Cochrane Collaboration*.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. doi: 10.1002/sim.1186

Hobbs, N., Rodrigues, A., Sniehotta, F. F., French, D. P., & Glidewell, L. (2011). Does measurement change health-related behaviour? A systematic review. from PROSPERO http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42011001467

Kalichman, S. C., Kelly, J. A., & Stevenson, L. (1997). Priming effects of HIV risk assessments on related perceptions and behavior: An experimental field study. *AIDS and Behavior, 1*(1), 3-8.

Knaus, C. S., & Austin, E. W. (1999). The AIDS Memorial Quilt as preventative education: a developmental analysis of the Quilt. *AIDS education and prevention : official publication of the International Society for AIDS Education, 11*(6), 525-540.

Knaus, C. S., Pinkleton, B. E., & Weintraub Austin, E. (2000). The Ability of the AIDS Quilt to Motivate Information Seeking, Personal Discussion, and Preventative Behavior as a Health Communication Intervention. *Health Communication, 12*(3), 301-316. doi: 10.1207/s15327027hc1203_05

Krauss, B. J., Goldsamt, L., Bula, E., Godfrey, C., Yee, D. S., & Palij, M. (2000). Pretest assessment as a component of safer sex intervention: a pilot study of brief one-session interventions for women partners of male injection drug users in New York City. [Clinical Trial; Randomized Controlled Trial, Research Support, Non-U.S. Gov't]. *Journal of Urban Health, 77*(3), 383-395.

Kvalem, I. L., Sundet, J. M., Rivø, K. I., Eilertsen, D. E., & Bakketeig, L. S. (1996). The Effect of Sex Education on Adolescents' Use of Condoms: Applying the Solomon Four-Group Design. *Health Education & Behavior, 23*(1), 34-47. doi: 10.1177/109019819602300103

Kypri, K., Langley, J. D., Saunders, J. B., & Cashell-Smith, M. L. (2006). Assessment may conceal therapeutic benefit: Findings from a randomized controlled trial for hazardous drinking. *Addiction, 102 (1)*, 62-70.

Kypri, K., & McAnally, H. M. (2005). Randomized controlled trial of a web-based primary care intervention for multiple health risk behaviors. *Preventive medicine, 41*(3–4), 761-766. doi: http://dx.doi.org/10.1016/j.ypmed.2005.07.010

Levav, J., & Fitzsimons, G. J. (2006). When Questions Change Behavior: The Role of Ease of Representation. *Psychological Science, 17*(3), 207-213. doi: 10.1111/j.1467-9280.2006.01687.x

Marteau, T. M., French, D. P., Griffin, S. J., Prevost, A. T., Sutton, S., Watkinson, C., . . . Hollands, G. J. (2010). Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. *Cochrane Database Syst Rev, 10*.

McCambridge, J., Butor-Bhavsar, K., Witton, J., & Elbourne, D. (2011). Can research assessments themselves cause bias in behaviour change trials? A systematic review of evidence from solomon 4-group studies. *PLoS One, 6*(10), e25223. doi: 10.1371/journal.pone.0025223

McCambridge, J., & Day, M. (2008). Randomized controlled trial of the effects of completing the Alcohol Use Disorders Identification Test questionnaire on self-reported hazardous drinking. *Addiction, 103*(2), 241-248. doi: 10.1111/j.1360-0443.2007.02080.x

McCambridge, J., & Kypri, K. (2011). Can simply answering research questions change behaviour? Systematic review and meta analyses of brief alcohol intervention trials. *PLoS One, 6*(10), e23748. doi: 10.1371/journal.pone.0023748

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of clinical epidemiology, 67*(3), 267-277.

McColl, E., Jacoby, A., Thomas, L., Soutter, J., & Bamford, C. (2002). Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technology Assessment, 5*(31), 256. doi: 10.3310/hta5310

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D., G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ, 339*. doi: 10.1136/bmj.b2535

Moreira, T., & Foxcroft, D. R. (2008). The effectiveness of brief personalized normative feedback in reducing alcohol-related problems amongst University students: Protocol for a randomized controlled trial. *BMC public health, 8*(113).

Moreira, T., Oskrochi, R., & Foxcroft, D. R. (2012). Personalised Normative Feedback for Preventing Alcohol Misuse in University Students: Solomon Three-Group Randomised Controlled Trial. *PLoS ONE, 7*(9).

Morwitz, V. G., & Fitzsimons, G. J. (2004). The Mere-Measurement Effect: Why Does Measuring Intentions Change Actual Behavior? *Journal of Consumer Psychology, 14*(1-2), 64-74.

Morwitz, V. G., Johnson, E., & Schmittlein, D. (1993). Does measuring intent change behavior? *Journal of consumer research*, 46-61.

O' Sullivan, I., Orbell, S., Rakow, T., & Parker, R. (2004). Prospective Research in Health Service Settings: Health Psychology, Science and the 'Hawthorne' Effect. *Journal of Health Psychology, 9*(3), 355-359.

Review Manager. (2011). (RevMan) (Version Version 5.1) [Computer program]. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Richmond, R., Heather, N., Wodak, A., Kehoe, L., & Webster, I. A. N. (1995). Controlled evaluation of a general practice-based brief intervention for excessive drinking. *Addiction, 90*(1), 119-132. doi: 10.1046/j.1360-0443.1995.90111915.x

Rimer, B., Levy, M. H., Keintz, M. K., Fox, L., Engstrom, P. F., & MacElwee, N. (1987). Enhancing cancer pain control regimens through patient education. *Patient Education and Counseling, 10*(3), 267-277. doi: http://dx.doi.org/10.1016/0738-3991(87)90128-5

Sandberg, T., & Conner, M. (2009). A mere measurement effect for anticipated regret: impacts on cervical screening attendance. [Comparative Study]. *British Journal of Social Psychology, 48*(Pt 2), 221-236.

Sandberg, T., & Conner, M. (2011). Using self-generated validity to promote exercise behaviour. *British Journal of Social Psychology, 50*(4), 769-783. doi: 10.1111/j.2044-8309.2010.02004.x

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials. *Annals of internal medicine, 152*(11), 726-732. doi: 10.7326/0003-4819-152-11-201006010-00232

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology, 39*(2), 211.

Spangenberg, E. (1997). Increasing Health Club Attendance Through Self-Prophecy. *Marketing Letters, 8*(1), 23-31. doi: 10.1023/A:1007977025902

Spangenberg, E., Sprott, D. E., Grohmann, B., & Smith, R. J. (2003). Mass-Communicated Prediction Requests: Practical Application and a Cognitive Dissonance Explanation for Self-Prophecy. *Journal of Marketing, 67*(3), 47-62. doi: 10.2307/30040536

Spence, J. C., Burgess, J., Rodgers, W., & Murray, T. (2009). Effect of pretesting on intentions and behaviour: a pedometer and walking intervention. [Randomized Controlled Trial]. *Psychology & health, 24*(7), 777-789.

Sprott, D. E., Smith, R. J., Spangenberg, E. R., & Freson, T. S. (2004). Specificity of Prediction Requests: Evidence for the Differential Effects of Self-Prophecy on Commitment to a Health Assessment. *Journal of Applied Social Psychology, 34*(6), 1176-1190. doi: 10.1111/j.1559-1816.2004.tb02002.x

Sprott, D. E., Spangenberg, E. R., & Fisher, R. (2003). The importance of normative beliefs to the self-prophecy effect. *Journal of Applied Psychology, 88*(3), 423.

Sprott, D. E., Spangenberg, E. R., Knuff, D. C., & Devezer, B. (2006). Self-prediction and patient health: Influencing health-related behaviors through self-prophecy. *Medical science monitor, 12*(5).

Todd, J., & Mullan, B. (2011). Using the theory of planned behaviour and prototype willingness model to target binge drinking in female undergraduate university students. *Addictive behaviors, 36*(10), 980-986. doi: http://dx.doi.org/10.1016/j.addbeh.2011.05.010

van Dongen, A., Abraham, C., Ruiter, R. C., & Veldhuizen, I. T. (2012). Does Questionnaire Distribution Promote Blood Donation? An Investigation of Question–Behavior Effects. *Annals of Behavioral Medicine*, 1-10. doi: 10.1007/s12160-012-9449-3

van Sluijs, E. M. F., van Poppel, M. N. M., Twisk, J. W. R., & van Mechelen, W. (2006). Physical activity measurements affected participants' behavior in a randomized controlled trial. [Multicenter Study; Randomized Controlled Trial]. *Journal of clinical epidemiology, 59*(4), 404-411.

van Valkengoed, I. G. M., Morré, S. A., Meijer, C. J. L. M., van den Brule, A. J. C., & Boeke, A. J. P. (2002). Do questions on sexual behaviour and the method of sample collection affect participation in a screening programme for asymptomatic Chlamydia trachomatis infections in primary care? *International journal of STD & AIDS, 13*(1), 36-38.

Walters, S. T., Vader, A. M., Harris, T. R., & Jouriles, E. N. (2009). Reactivity to alcohol assessment measures: an experimental test. [Randomized Controlled Trial; Research Support, N.I.H., Extramural]. *Addiction, 104*(8), 1305-1310.

Yardley, L., Miller, S., Schlotz, W., & Little, P. (2011). Evaluation of a Web-based intervention to promote hand hygiene: exploratory randomized controlled trial. *Journal of Medical Internet Research, 13*(4), e107.

Figure 1: Trial selection flow diagram (adapted from PRISMA (Moher, et al., 2009))

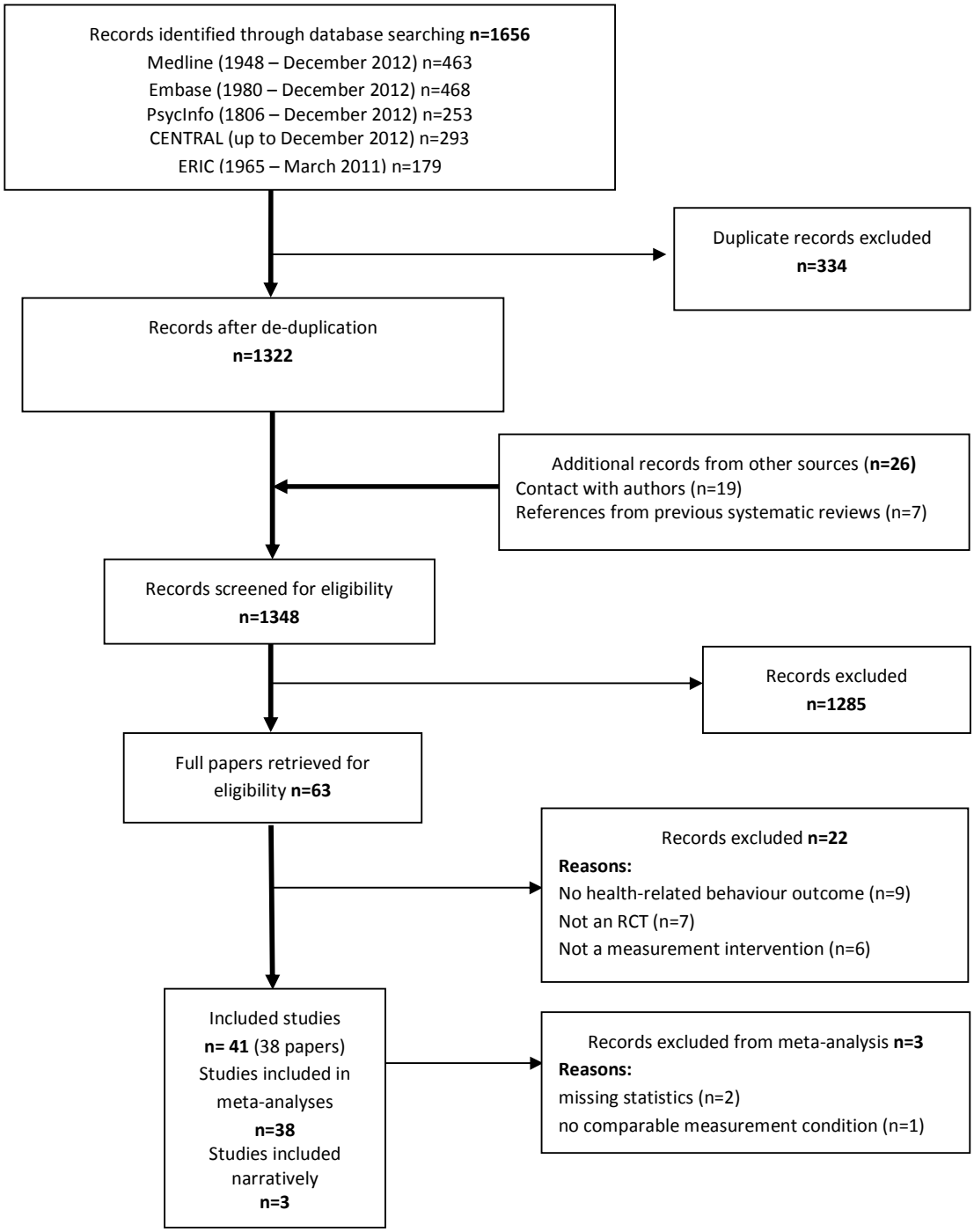


Figure 2: Funnel plot of trials reporting health-related behaviour outcomes

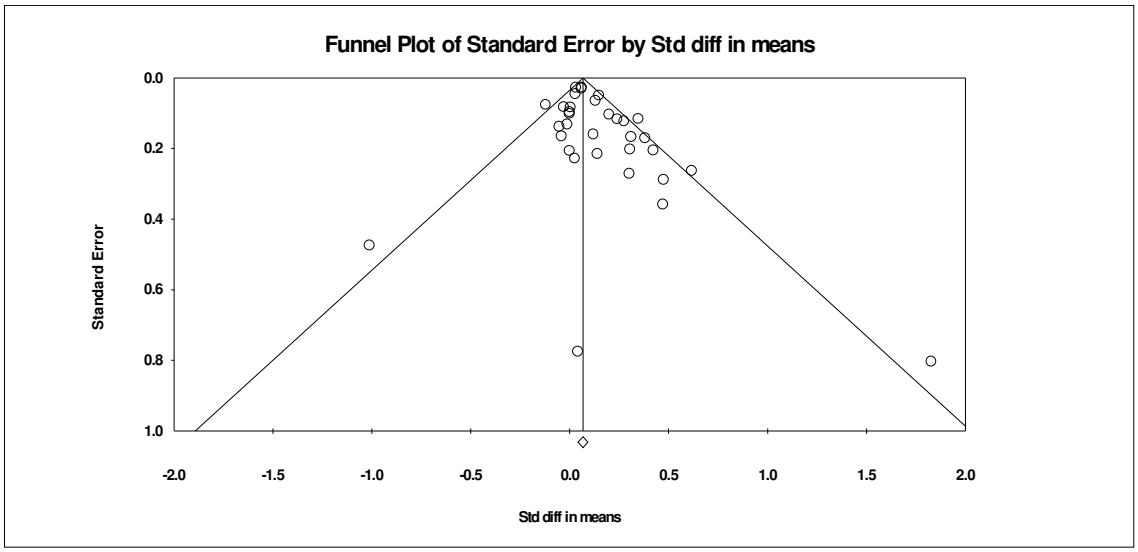
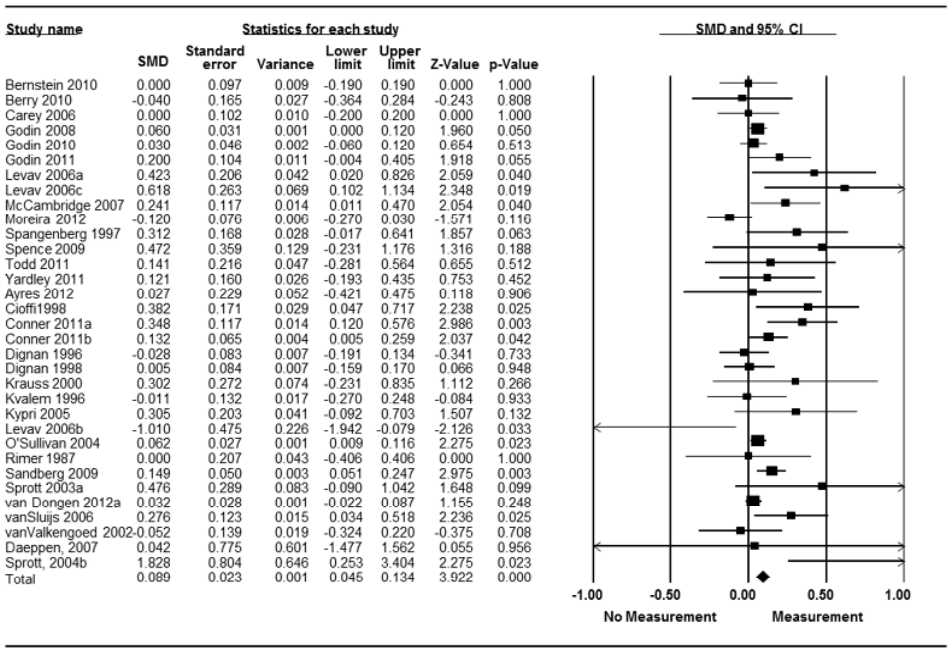


Figure 3: Forest plot of standardized mean differences (SMD) and 95% confidence intervals for health-related behaviors in measurement vs. no measurement conditions



Study name	SMD	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Bernstein 2010	0.000	0.097	0.009	-0.190	0.190	0.000	1.000
Berry 2010	-0.040	0.165	0.027	-0.364	0.284	-0.243	0.808
Carey 2006	0.000	0.102	0.010	-0.200	0.200	0.000	1.000
Godin 2008	0.060	0.031	0.001	0.000	0.120	1.960	0.050
Godin 2010	0.030	0.046	0.002	-0.060	0.120	0.654	0.513
Godin 2011	0.200	0.104	0.011	-0.004	0.405	1.918	0.055
Levav 2006a	0.423	0.206	0.042	0.020	0.826	2.059	0.040
Levav 2006c	0.618	0.263	0.069	0.102	1.134	2.348	0.019
McCambridge 2007	0.241	0.117	0.014	0.011	0.470	2.054	0.040
Moreira 2012	-0.120	0.076	0.006	-0.270	0.030	-1.571	0.116
Spangenberg 1997	0.312	0.168	0.028	-0.017	0.641	1.857	0.063
Spence 2009	0.472	0.359	0.129	-0.231	1.176	1.316	0.188
Todd 2011	0.141	0.216	0.047	-0.281	0.564	0.655	0.512
Yardley 2011	0.121	0.160	0.026	-0.193	0.435	0.753	0.452
Ayres 2012	0.027	0.229	0.052	-0.421	0.475	0.118	0.906
Cioffi1998	0.382	0.171	0.029	0.047	0.717	2.238	0.025
Conner 2011a	0.348	0.117	0.014	0.120	0.576	2.986	0.003
Conner 2011b	0.132	0.065	0.004	0.005	0.259	2.037	0.042
Dignan 1996	-0.028	0.083	0.007	-0.191	0.134	-0.341	0.733
Dignan 1998	0.005	0.084	0.007	-0.159	0.170	0.066	0.948
Krauss 2000	0.302	0.272	0.074	-0.231	0.835	1.112	0.266
Kvalem 1996	-0.011	0.132	0.017	-0.270	0.248	-0.084	0.933
Kypri 2005	0.305	0.203	0.041	-0.092	0.703	1.507	0.132
Levav 2006b	-1.010	0.475	0.226	-1.942	-0.079	-2.126	0.033
O'Sullivan 2004	0.062	0.027	0.001	0.009	0.116	2.275	0.023
Rimer 1987	0.000	0.207	0.043	-0.406	0.406	0.000	1.000
Sandberg 2009	0.149	0.050	0.003	0.051	0.247	2.975	0.003
Sprott 2003a	0.476	0.289	0.083	-0.090	1.042	1.648	0.099
van Dongen 2012a	0.032	0.028	0.001	-0.022	0.087	1.155	0.248
vanSluijs 2006	0.276	0.123	0.015	0.034	0.518	2.236	0.025
vanValkengoed 2002	-0.052	0.139	0.019	-0.324	0.220	-0.375	0.708
Daeppen, 2007	0.042	0.775	0.601	-1.477	1.562	0.055	0.956
Sprott, 2004b	1.828	0.804	0.646	0.253	3.404	2.275	0.023
Total	0.089	0.023	0.001	0.045	0.134	3.922	0.000

TABLE 1: Characteristics of included studies

Study ID	**Format of measurement**	**Type of measure**	**Content of measurement**	**Health-related outcome**	**Follow-up**	**Country**	**Study Setting**	**Population age and gender composition**	**Sample size at follow up**	**Risk of bias score** 0 (low risk) – 16 (high risk)
Ayres, et al. (2013)	Questionnaire	Dichotomous	Intention, attitudes and anticipated regret	Health plan uptake (objective)	Immediately after measurement	UK	Community	Mean age: 53.4 (71.2 % female)	Measurement condition: 67 No measurement condition: 79	0
Bernstein et al. (2010)	Questionnaire	Continuous	Drinking behavior, other health behaviors, patient health questions and PTSD symptoms	Alcohol use (self-report)	12 months	USA	Pediatric emergency department	Age ≤ 17y = 114 ≥18y = 739	Measurement condition: 209 No measurement condition: 198	4
Berry and Carson (2010)	Questionnaire	Continuous	Behavior and attitude	Physical activity (self-report)	7-10 days	Canada	University and community	Students sample: mean age 19.7 (73.7% female) Community sample: mean age 72.0 (75.4% female)	Measurement condition: 117 No measurement condition: 54	7
Carey, Carey, Maisto, and	Interview	Continuous	Behavior	Alcohol use (self-report)	1, 6 and 12 months	USA	University	Mean age: 19.2 (65% female)	Measurement condition: 197	8

Henson (2006)									No measurement condition: 197	
Cherpitel et al. (2010)	Questionnaire	Continuous	Behavior	Alcohol use (self-report)	12 months	Poland	Emergency Department	39% <30 years (16% female)	Screened only: 87 Assessed: 97	4
Cioffi and Garner (1998)	Questionnaire	Dichotomous	Cognitions only	Blood donation behavior (objective)	1-week	USA	University	Not provided	Measurement condition: 277 No measurement condition: 370	3
Clifford, Maisto, and Davis (2007)	Interview	Continuous	Behavior	Alcohol use (self-report)	6 and 12 months	USA	Treatment Centre for alcohol and other drugs abuse	Mean age: 40.01 (37% female)	Intensive assessment: 59 Least intensive assessment: 62	3
Conner, et al. (2011)a	Questionnaire	Dichotomous	Theory Planned Behavior cognitions	Health check attendance (objective)	4 months	England	GP practice	Mean age: 36.4 (52.3% female)	Measurement condition: 199 No measurement condition: 185	0

Conner, et al. (2011)b	Questionnaire	Dichotomous	Theory Planned Behavior cognitions	Vaccination uptake (objective)	2 months	Canada	Public hospital	Mean age: 38.1 (83.4% female)	Measurement condition: 600 No measurement condition: 600	2
Daeppen et al. (2007)[3]	Interview	Dichotomous	Behavior	% of hazardous drinkers (self-report)	12 months	Switzerland	Emergency department	Mean age: 36.7 (21.8% female)	Measurement condition: 277 No measurement condition: 257	3
Dignan et al. (1996)	Interview	Dichotomous	Knowledge, intentions and behavior	Pap smear screening attendance (self-report)	12 months	USA	Tribal community: Cherokee Indian	63.8% <45 years (100% female)	Measurement condition: 448 No measurement condition: 367	7
Dignan et al. (1998)	Interview	Dichotomous	Knowledge, intention and behavior	Pap smear screening attendance (self-report)	12 months	USA	Tribal community: Lumbee Native American	Mean age: 42.4 (100% female)	Measurement condition: 413 No measurement condition:	8

[3] Revman could not compute an effect size for this study as counts and events were equal in both groups. For this reason a value was removed in events for each group.

									426	
Godin, et al. (2008)	Questionnaire	Continuous	Theory Planned Behavior cognitions	Blood donation behavior (objective)	6 and 12 months	Canada	Blood Donors agency	Mean age control: 43.8 (38.7% female) Mean age measurement: 44.7 (38.3% female)	Measurement condition: 2900 No measurement condition: 1772	1
Godin, et al. (2010)[4]	Questionnaire	Continuous	Anticipated regret and intention	Blood donation behavior (objective)	6 and 12 months	Canada	Blood Donors agency	Mean age: 30.4 (53 % female)	Measurement condition: 879 No measurement condition: 888	2
Godin, Bélanger-Gravel, Amireault, Vohl, and Pérusse (2011)	Interview	Continuous	Theory Planned Behavior cognitions, anticipated regret, moral and descriptive norms, self-efficacy, facilitating factors and positive	Physical activity (self-report)	3 months	Canada	Quebec city community	Mean age: 40.2 (47 % female)	Measurement condition: 194 No measurement condition: 180	2

[4] For this study, groups assessing implementation intentions were not included in the analyses.

			feelings							
Krauss et al. (2000)	Questionnaire	Dichotomous	Knowledge, perceived partner risk, behavior	Safe sex Index (self-report)	7 weeks	USA	Community: public spaces	Mean age: 36.7 (100 % female)	Measurement condition: 45 No measurement condition: 28	2
Kvalem, et al. (1996)	Questionnaire	Dichotomous	Behavior	Condom use (self-report)	6 and 12 months	Norway	High school	16-20 years (50 % female)	Measurement condition: 148 No measurement condition: 133	9
Kypri, Langley, Saunders, and Cashell-Smith (2006)	Questionnaire	Continuous	Behavior	Alcohol use (self-report)	6 and 12 months	New Zealand	Primary Health-care clinic	Mean age control: 20.1; Mean age measurement: 20.3 (52.2 % female)	Measurement condition: 126 No measurement condition: 126	0
Kypri and McAnally (2005)[5]	Questionnaire	Dichotomous	Behavior	Fruit and veg consumption, alcohol consumption, and physical activity frequency	6 weeks	New Zealand	University primary Health-care clinic	Mean age: 20.2 (49 % female)	Measurement condition: 64 No measurement condition: 60	2

[5] Outcomes were merged to produce a single health-related outcome.

				(self-report)						
Levav and Fitzsimons (2006)a	Questionnaire	Continuous	Intention to floss	Flossing (self-report)	2-weeks	USA	University	Not provided	Measurement condition: 51 No measurement condition: 46	6
Levav and Fitzsimons (2006)b	Questionnaire	Dichotomous	Behavior	Choice of low or high fat snack (self-report)	Immediately after pre-test	USA	University	Not provided	Measurement condition: 25 No measurement condition: 23	4
Levav and Fitzsimons (2006)c	Questionnaire	Continuous	Intention to floss	Flossing (self-report)	1-week	USA	University	Not provided	Measurement condition: 30 No measurement condition: 30	8
(McCambridge & Day, 2008)	Questionnaire	Continuous	Questionnaire (General Health questionnaire-GHQ, history of trauma scale – HTS, and alcohol use - AUDIT)	Alcohol use – AUDIT (self-report)	2-3 months	England	University	Mean age control: 20.7 (66 % female); Mean age measurement: 20.6 (67 % female)	Measurement condition: 156 No measurement condition: 144	0
Moreira, Oskrochi, and Foxcroft (2012)	Questionnaire	Continuous	Behavior, behavior-related problems,	Alcohol use (self-report)	6 and 12 months	UK	University	58.5% 17-19 years (61 % female)	Measurement condition: 369	4

			perceived norms, positive expectancies)						No measurement condition: 332	
O' Sullivan, Orbell, Rakow, and Parker (2004)	Questionnaire	Dichotomous	Perceptions of susceptibility and severity of colorectal cancer and attitudes and personal beliefs	Colorectal cancer screening uptake (objective)	6-weeks	UK	Community	Age between 50 and 69 years	Measurement condition: 1944 No measurement condition: 10,413	0
Rimer et al. (1987)	Interview	Dichotomous	Behavior and disease-related information, knowledge and concerns about pain regimens, perceived personal control and anxiety	Medication regimens adherence (self-report)	4 weeks	USA	Hospitals	Age: 53.9% more than 60y (44.3 % female)	230 participants	7
Sandberg and Conner (2011)	Questionnaire	Continuous	Theory Planned Behavior cognitions	Physical activity (objective)	2-weeks	UK	University	Mean age: 19.7 (62.0 % female)	TPB only: 192 TPB + regret: 384	2
Sandberg and Conner (2009)	Questionnaire	Dichotomous	Theory Planned Behavior cognitions,	Cervical screening attendance	4 months	England	Central Agency responsible for cervical	Mean age: 39.1 (100 % female)	Measurement condition: 1426	2

			anticipated regret	(objective)			screening		No measurement condition: 1277	
Spangenberg (1997)	Questionnaire	Continuous	Behavior	Health club attendance (objective)	10 days and 6 months attendance	USA	Health club	Not provided	Measurement condition: 73 No measurement condition: 69	4
Spence, et al. (2009)	Questionnaire	Continuous	Behavior, illness perceptions, self-efficacy, intention	Walking behavior (self-report)	1 week	Canada	University	95% <30 years (100 % female)	Measurement condition: 15 No measurement condition: 16	5
Sprott, Smith, Spangenberg, and Freson (2004)b	Questionnaire	Dichotomous	Behavior	Health and fitness assessment attendance (self-report)	Immediately after pre-test	USA	University	Not provided	Measurement condition: 61 No measurement condition: 60	4
Sprott, Spangenberg, and Fisher (2003)a	Questionnaire	Dichotomous	Behavior	Choice of low-fat or higher fat snack (self-report)	Immediately after pre-test	USA	University	Age not provided (100 % female)	Measurement condition: 36 No measurement condition: 44	4
Todd and Mullan (2011)	Questionnaire	Continuous	Behavior, prototypes and Theory Planned	Alcohol use (self-report)	2 weeks	Australia	University	Mean age: 19 (100 % female)	Measurement condition: 44 No measurement	4

			Behavior cognitions,						condition: 42	
van Dongen, et al. (2012)	Questionnaire	Dichotomous	Intention, attitudes (affective and cognitive), subjective, descriptive and moral norms, self-efficacy and role identity	Blood donation behavior (objective)	6 months	The Netherlands	Blood Donors agency: new donors	Mean age: 33.4 (67 % female)	Measurement condition: 3518 No measurement condition: 3490	2
van Sluijs, van Poppel, Twisk, and van Mechelen (2006)	Questionnaire and accelerometers (without display)	Dichotomous	Behavior and barriers to PA, knowledge, health process of change, social support and self-efficacy	Physical activity (self-report)	6 months	The Netherlands	GP practices	Mean age: 55.7 (54% female)	Measurement condition: 155 No measurement condition: 172	3
van Valkengoed, Morré, Meijer, van den Brule, and Boeke (2002)	Questionnaire	Dichotomous	Behavior	Chlamydia screening attendance (objective)	Not provided	Netherlands	Primary care practice	15-40 years (63.2% female)	Measurement condition: 143 No measurement condition: 149	3
Walters, Vader, Harris, and Jouriles (2009)	Questionnaire	Continuous	Behavior, readiness to change, normative beliefs	Peak blood alcohol concentration (self-report)	12 months	USA	University	Mean age: 19.8 (66 % female)	Intensive assessment: 63 Least intensive assessment:	1

									66	
Yardley, Miller, Schlotz, and Little (2011)	Questionnaire	Continuous	Theory Planned Behavior cognitions, perceived risk of infection	Hand washing (self-report)	4 weeks	England	GP practices	Mean age: 49.8 (64 % female)	Measurement condition: 77 No measurement condition: 80	4
Studies excluded from meta-analysis										
Kalichman, et al. (1997)	Interview and questionnaire	Continuous	Behavior	Sexual risk behaviors (self-report)	2 weeks	USA	Community: African American	Mean age: 34.0 (100 % female)	158 participants	10
Knaus and Austin (1999)	Questionnaire	--	Perceptions, self-efficacy, behavior	Sexual risky behavior Index (self-report)	8 weeks	USA	University	Mean age: 19.41 (54 % female)	237 participants	7
Knaus, et al. (2000)	Questionnaire	--	Behavior	Safe sex behaviors Index (self-report)	7-8 weeks	USA	University	Mean age: 19 (53.9 % female)	Measurement condition: 47 No measurement condition: 61	9

TABLE 2: Standardized mean differences (Cohen's *d*) for question-behavior effect by moderator variables[6]

Moderator variable	Measurement group N	No Measurement group N	*k*	I^2(95% CI)	*Q*	SMD	95% CI
Type of participants							
Students	926	1035	14	63% (33.4%, 79.1%)	1.38	0.17	0.01-0.32
Non-students samples	4599	3444	19	20% (0.0%, 54.0%)		0.07	0.04-0.11
Content of measurement							
Behavior only	752	739	9	53% (0.0%, 77.9%)	1.19	0.11	-0.09-0.30
Cognitions only	3860	2736	13	32% (0.0%, 66.6%)		0.05	0.05-0.15
Cognitions plus behavior	923	1004	11	50% (5.9%, 73.7%)		0.05	-0.04-0.14
Measurement of attitudes							
Yes	11193	18392	12	31% (0.0%, 65.5%)	0.00	0.09	0.05-0.13
No	3922	3945	21	50% (17.8%, 70.0%)		0.09	0.01-0.18
Format of measurement							
Questionnaires	4558	3647	27	51% (24.0%, 68.7%)	2.02	0.10	0.05-0.15
Interviews	877	832	6	0% (0.0%, 62.4%)		0.03	-0.06-0.12
Type of health-related					13.96		

[6] For each subgroup analysis, total *k* = 33 with the exception of type of health-related behavior subgroup analysis. In this analysis, k = 29 because while the Kypri 2005 study reports on 3 different behaviors, there are 6 other studies which report on distinct behaviors that could not be combined.

behavior							
Flossing	81	76	2	0% (NA)		0.50	0.18-0.81
Blood donation	7574	6520	4	33% (0.0%, 76.1%)		0.05	0.00-0.10
PA	573	598	6	0% (0.0%, 65.9%)		0.20	0.08-0.32
Screening	4374	12632	5	24% (0.0%, 69.5%)		0.06	0.003-0.12
Drinking	1262	1281	7	35% (0.0%, 72.5%)		0.04	-0.08-0.16
Diet	124	130	3	76% (21.2%, 92.7%)		0.08	-0.68-0.84
Sexual Behavior	193	161	2	7% (NA)		0.05	-0.20-0.31
Type of outcome							
Objective	3852	2729	11	45% (0.0%, 72.6%)	0.39	0.08	0.04-0.13
Self-report	1683	1750	22	46% (11.8%, 67.6%)		0.10	0.01-0.19

Cochrane's Q = heterogeneity for the subgroup differences analysis
Standardised Mean Difference (SMD) = Cohen's d = pooled effect size
**p< .01; NA= Not available

Appendix 1: Database searches
MEDLINE from inception to December 2012

1. randomized controlled trial.pt.
2. controlled clinical trial.pt.
3. randomized.ab.
4. placebo.ab.
5. drug therapy.fs.
6. randomly.ab.
7. trial.ab.
8. groups.ab.
9. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8
10. exp animals/ not humans.sh.
11. 9 not 10
12. interview/
13. Interview, Psychological/
14. questionnaires/
15. health care surveys/
16. exp "Weights and Measures"/
17. (complet* adj3 (measure* or scale* or interview* or survey* or questionnaire* or test*)).tw.
18. "Outcome Assessment (Health Care)"/
19. (panel* adj3 survey*).tw.
20. exp Mass Screening/
21. ("follow up" adj1 (outcome* or measure* or score* or interview* or assessment*)).tw.
22. (behavio?r* adj4 measure*).ti.
23. 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22
24. (behavio?r adj2 measure*).ti.
25. Behavioral Research/
26. Health Behavior/
27. exp patient compliance/
28. exp self-examination/
29. treatment refusal/
30. feeding behavior/
31. fasting/
32. food habits/
33. food preferences/
34. illness behavior/
35. exp reproductive behavior/
36. risk reduction behavior/
37. risk-taking/
38. exp sexual behavior/
39. exp "tobacco use cessation"/
40. motor activity/
41. Alcohol Drinking/
42. ("physical exercise*" or "physical activit*").tw.
43. Alcoholism/
44. (drink* adj1 (alcohol* or pattern* or problem* or addict*)).tw.
45. 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44
46. (panel* adj2 conditioning).tw.
47. (pretest* adj2 (response* or effect* or bias* or reactivity)).tw.
48. (test* adj2 (response* or effect* or bias* or reactivity)).tw.
49. (measurement* adj2 (response* or effect* or bias* or reactivity)).tw.
50. (assessment* adj2 (response* or effect* or bias* or reactivity)).tw.
51. (question* adj2 (response* or effect* or bias* or reactivity)).tw.
52. (interview* adj2 (response* or effect* or bias* or reactivity)).tw.
53. (reactiv* adj2 (response* or effect* or bias* or measure*)).tw.
54. "generated validity".tw.
55. mere measur$.tw.
56. "self prophecy".tw.
57. (solomon adj3 (group$ or design$ or trial$ or study or studies)).tw.
58. (solomon adj2 island$).tw.
59. 57 not 58
60. 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 59
61. 11 and 23 and 45 and 60