Evidence Report Disposition of Comments Report

Research Review Title: Behavioral Programs for Diabetes Mellitus

Draft review available for public comment from February 23, 2015 to March 23, 2015.

Research Review Citation: Pillay J, Chordiya P, Dhakal S, Vandermeer B, Hartling L, Armstrong MJ, Butalia S, Donovan LE, Sigal RJ, Featherstone R, Nuspl M, Dryden DM. Behavioral Programs for Diabetes Mellitus. Evidence Report/Technology Assessment No. 221. (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-2012-00013-I.) AHRQ Publication No. 15-E003-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2015. www.effectivehealthcare.ahrq.gov/reports/final/cfm.

Comments to Research Review

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each research review is posted to the EHC Program Web site in draft form for public comment for a 4-week period. Comments can be submitted via the EHC Program Web site, mail or E-mail. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft comparative effectiveness research review.

Comments on draft reviews and the authors' responses to the comments are posted for public viewing on the EHC Program Web site approximately 3 months after the final research review is published. Comments are not edited for spelling, grammar, or other content errors. Each comment is listed with the name and affiliation of the commentator, if this information is provided. Commentators are not required to provide their names or affiliations in order to submit suggestions or comments.

The tables below include the responses by the authors of the review to each comment that was submitted for this draft review. The responses to comments in this disposition report are those of the authors, who are responsible for its contents, and do not necessarily represent the views of the Agency for Healthcare Research and Quality.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Commentator & Affiliation	Section	Comment	Response
Peer Reviewer 1	General	This report describes the process and findings of an extensive systematic review and meta-analysis of previous primary studies conducted to test behavioral interventions in persons diagnosed with T1DM or T2DM. This review updates and extends previous similar systematic reviews/meta analyses by including a larger number of studies, systematically separating analyses according to type of diabetes, and conducting specific analyses to "tease out" potential factors that influence intervention outcomes, e.g., intervention intensity. This review is extremely important and useful, given the growing recognition of the importance of behavioral interventions in addressing the worldwide global diabetes epidemic. While the findings were somewhat disappointing in terms of overall effectiveness of behavioral interventions, the authors provide important information to guide future research studies in this area.	Thank you for this comment. No changes required.
Peer Reviewer 2	General	The report is clinically meaningful. The comprehensiveness of the report is outstanding and offers much insight into what areas are lacking in terms of high quality evidence around behavioral programs for both type 1 and type 2 diabetes. The authors have considered all relevant aspects particularly in Type 1 as there are distinctions in the approach to self-management for young people/younger adults with type 1 and the evidence for type 1 is weak and would benefit from a synthesis to inform research priorities. The outcomes of interest from the six key questions were also well outlined moving from behavioral outcomes to utilization to harms.	Thank you for this comment. No changes required.
Peer Reviewer 3	General	This report is clinically meaningful and the key questions are clearly stated. The PICOTS are clearly defined.	Thank you for this comment. No changes required.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 4	General	The report reviewed the effectiveness of Behavioral programs on diabetic care. The investigators made excellent efforts to make the report clinically meaningful and defined clinically meaningful differences when possible. The target population defined to be community settings and the key questions are explicitly stated. There is no explanation why harms are not included in key questions for Type 2 diabetes.	Thank you for this comment. We have added a sentence into the section on Rationale for the Review in relation to not examining harms in T2DM which focused on moderating effects. "Because of our focus on moderation of effectiveness for T2DM, we did not examine harms as we did for T1DM." This section appears in both the full report and executive summary (section Rationale for Evidence Review).
TEP 1	General	Overall, the report describes the analyses undertaken and results obtained in a clear manner. The key questions are addressed and the report is clinically meaningful.	Thank you for this comment. No changes required.
TEP 1	General	The analyses included the change (primarily) in A1c and body weight at 12/24 months. There is no mention that most of the programs delivered were less than 6 months in duration and often did not include a maintenance phase of intervention following the intensive intervention phase. Only the total number of contact hours is given in the tables but these contact hours usually are dispersed very differently from beginning of intervention to final followup. Recommend adding this point to the Discussion section of the Excecutive Summary and full report. The impact of these interventions on A1c is usually greater immediately following the intensive intervention phase. We need more research to identify programs and strategies for maintaining change in outcomes during the maintenance (often no contact) phase of the study. The need for this research also could be added to the table regarding Research Gaps. This point needs to be made as it places the findings in context.	Thank you this comment and suggestion. All programs categorized as DSME plus support, and most lifestyle programs, had a maintenance or support phase although often of short total duration. We added discussion on these matters in the Limitations of the Comparative Effectiveness Review sections of the executive summary and full report (Discussion section). We agree that mechanisms (especially ones with low resource requirements) for maintaining outcomes are important but our identification of research needs was specific to our key questions which did not address these interventions.
TEP 2	General	I thought it was generally well done. My big issue is the fact that educational and behavioral programs are considered the same. They absolutely are not. The title of the document is very misleading as the results are all about educational programs and therefore it should be renamed.	Thank you for this comment. We agree that the terminology used can impact the readership and thus use of results. The title was chosen in consultation with the technical expert panel, to examine programs incorporating behavioral approaches to achieve changes in multiple

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		In addition, there is a much broader level of interest in educational programs than behavioral so it would also expand the readership.	behaviors important for managing diabetes. Our operational definition reflects this aim, and several studies were excluded based on failure to explicitly state incorporation of training using behavior change techniques. Changing the title to educational programs would fail to capture this important aspect of the review, as well as its inclusion of programs (classified as lifestyle due to the predominance of programs focusing on diet and exercise) that did not focus on education.
TEP 3	General	Lifestyle programs do not typically target people with diabetes. They target those at risk. Therefore, including lifestyle programs in the same category as DSME and DSMS may bias the results. Specifically, the main outcome of a DSME/DSMS p program is typically change in A1c, while the main outcome in a lifestyle intervention is typically change in weight or change in fasting glucose levels.	Thank you for this comment. As stated in the report, we used the term lifestyle programs to reflect the large number of studies that were not considered DSME (having a primary focus on diabetes-specific self care) but that still met our operational definition of behavioral programs by focusing on multiple behaviors and targeting patients with diabetes. Many of these studies had HbA1c as one of their primary outcomes and all focused on people with diabetes. While many lifestyle programs target people without diabetes, lifestyle interventions are also part of the management of diabetes.
TEP 3	General	A1c results at 12 months are not typically significant if there isn't ongoing support involved, which is the case in this review. It is critically important to present the results stratified by the categories of interventions (DSME, DSME+ DSME, Lifestyle) to tease out if the programs that offered support were effective at, at least MAINTAINING improvements achieved at 6 months.	We agree that stratifying the programs based on these components was important and our moderator analysis captured the relative effects of these programs. Because of the large variability between programs in number of phases and intensities, we were unable to capture the effects specific to a support phase which would be necessary to address this point accurately.
TEP 4	General	The report includes a lot of very important data but I think there need to be more call out boxes or summaries with implications for practitioners and healthcare systems to make it more clinically meaningful.	Thank you for this comment and suggestion. We expanded the clinical implications of the results. We have also submitted 2 manuscripts related to this report, which provide a more succinct reporting of results and we hope will capture a broad readership.
TEP 5	General	The actual 77 page document is quite thorough and provides both the results and the limitations of those results. In contrast, parts of the Executive Summary and abstract are somewhat overstated and needs to be a	Thank you for this comment and the related ones. We agree that the report provided a better description of some of the limitations and these are also important to reflect in the executive summary.

		little more transparent regarding the limitations of some of these summary statements. Specific examples of this are provided below.	We have revised the comments in the Executive Summary and full text in relation to the results for ethnicity and the lack of studies reporting in T1DM to enable valid conclusions about the relative effectiveness based on age. For example, in the results section for the ethnicity subgroup analysis in the Executive summary we have added the sentence "These results need to be interpreted with caution because of the apparent worse baseline glycemic control in studies of minority participants; this factor may account for much of the increase in benefit." We also revised the wording in the discussion of the T2DM findings to this effect." Also in the executive summary, we now state "In the comparisons with active controls, the small number of studies in most subgroups provided insufficient SOE for making any conclusions."
Public Comment (Kelly McDermott, Omada Health Inc.)	General	I do not understand why the title is behavioral programs when you state clearly that this is essentially DSME. Because education is a covered benefit by insurers more often than behavioral programs they are a prime audience but they will have no reason to read this paper. The word education must appear in the title. Also medications are recommended from very near the onset of type 2 diabetes according to the latest standards and consensus statements.	Our statement of "essentially DSME" in the conclusions section of the full report was only in relation to T1DM because of the available evidence (i.e. very few DSME plus support or lifestyle programs were included for T1DM). We have revised this to "especially DSME" to be more specific. Many programs for T2DM were not educational. Our purpose was to assess programs offering behavioral approaches striving to achieve benefits for multiple behaviors regardless of their classification. Our definition is quite similar to that of an earlier Technology Assessment commissioned by the Agency of Healthcare Research and Quality on behavioral therapies for diabetes, which was conducted in partnership with Centers for Medicaid and Medicare – see Matcher http://www.ncbi.nlm.nih.gov/pubmed/25855838). These interventions were considered to be those offering behavioral approaches to benefit physical

			outcomes; the only reason to exclude DSME was because these were already reimbursed therefore not appropriate for a technology assessment. We have revised our wording to be clear that T2DM is managed with medications and/or insulin.
Public Comment (Kelly McDermott, Omada Health Inc.)	Abstract	Currently, the limited information in the type 2 diabetes paragraph of the results section of the abstract could be misleading given the lack of reference to comparison groups (i.e. ≤10 hours). For clarification, we suggest comparison groups and/or the entire participant population be made explicit somewhere in the abstract.	Thank you for pointing this out. We have added details about the interventions and comparators into the abstract (i.e., "..evaluating behavioral programs compared with usual care, active controls (e.g., didactic education), or other behavioral programs." The 10 hour factor was a level within the program intensity variable assessed in the analyses, but not a characteristic of an active control.
Peer Reviewer 1	Executive Summary	In the Executive Summary and report sections on Epidemiology and Burden of Disease, the statistics for T1DM and T2DM are combined. It would be more informative to separate the prevalence and cost information by disease, since they are so different.	Thank you for this suggestion. We have revised this section (in Executive Summary and Report) to be more explicit with respect to type of diabetes; we also added some data for T1DM in relation to prevalence and healthcare costs.
Peer Reviewer 1	Executive Summary	The Executive Summary seems quite long and perhaps would be more useful if shortened.	We realize that the executive summary is quite long. We have submitted separate manuscripts for T1DM and T2DM based on this report, and anticipate this will help disseminate the results in shorter formats to reach a larger audience.
Peer Reviewer 4	Executive Summary	The units for some outcomes are not provided in Table A-C, when using MD.	We have added information about the units in these tables.
Peer Reviewer 4	Executive Summary	ES-24, line 47: "but that more, god quality evidence" □ should be good.	Thanks for pointing out this error. We have made the correction.
Peer Reviewer 1	Introduction	The justification for conducting this review is clearly stated and supported.	Thank you for noting this. No change required.
Peer Reviewer 2	Introduction	The introduction is generally well written. I would suggest a little more distinction around the section type 1 and type 2 under the "Diabetes Care and Self-Management"; perhaps sub-heads. I suggest this given the importance of the distinction with reference to self-monitoring and required insulin use. More and more we	Thank you this suggestion. We have removed mention of self-monitoring in the paragraph (3rd) applicable to T2DM, and revised this whole paragraph to focus on the lifestyle factors (not blood glucose control) also required in T1DM and T2DM for reduction of complications.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		are recognizing that blood glucose monitoring is less important for type 2 and may in fact by increasing diabetes related distress/anxiety (authors looked into this in latter stages).	
Peer Reviewer 2	Introduction	Pg 11; lines 5-15. I would suggest there needs to be some reference to medical nutrition therapy (MNT). Not so sure language like "treated" is appropriate but rather than "manged" is m	Thank for this comment. We have added reference to medical nutrition therapy and revised the wording to management for T2DM.
Peer Reviewer 2	Introduction	Pg 11; lines 41-47. Might be worthwhile to provide more scope/detail in terms of achieving behavioral targets (16% meet DSME targets) since the authors have suggested 45% of adults in the US.	The 45% referred to glycemic targets rather than behaviors, but we agree that these figures may appear discrepant. We decided to focus on the statistics for the US (the 16% was international) so this figure has been removed and we have kept those from the CDC specific to particular behaviors/risk factors.
Peer Reviewer 2	Introduction	Would suggest authors makes some reference to Behavioral theory - I recognize this is beyond the scope of the paper but there have been some important work completed in this area (Avery et al., Diabetes Care 2012).	Thanks for this suggestion. We have added reference to behavior change techniques (taxonomy of Michie et al. Psychol Health. 2011 Nov;26(11):1479-98), because these were required for study inclusion, with or without the explicit use of behavioral theory. Under Scope of Review , we added, "A commonality with all programs was that they incorporated one or more behavior change techniques, with or without an explicit use of a theory or model of behavior change."
Peer Reviewer 3	Introduction	The introduction is clear and well written, though a bit longer than necessary. The two analytic framework models are a bit confusing; primarily, it is unclear why both are needed rather than having a single framework that represents aspects of multiple key questions.	We realize the frameworks appear quite similar. However, the focus was different with T1DM examining overall effectiveness and harm including grading of outcomes, and T2DM examining moderators of effectiveness. It was important to reflect this clearly in the frameworks. We removed the "KQ5" placed near the outcomes in the analytical framework, which was an oversight, to avoid implying our focus was on effectiveness.
Peer Reviewer 4	Introduction	The introduction provides a good description of background, pathophysiology, epidemiology and disease management options with	Thanks for this comment. No changes required.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		appropriate amount of details. Rationale for evidence review is clearly stated with different focus for type I and type I diabetes.	
Peer Reviewer 4	Introduction	Page 4, line 10, Level of glycemic control (HbA1c < 7 vs. >= 7 percent) -- baseline values? Target values? (since it is called level of glycemic control).	We have specified this (as baseline control) in the wording of the key questions and methods sections.
Peer Reviewer 4	Introduction	Maybe a sentence to explain why harms are not included as one of the key questions for type 2 diabetes?	We have added a paragraph in the section on rationale for evidence synthesis. Because of our focus on moderation of effectiveness for T2DM, we did not examine harms as we did for T1DM. This review provides information regarding the effectiveness and harms of behavioral programs (T1DM), and what combination of program components and delivery methods are most effective for implementation of these programs in community health settings (T2DM).
TEP 1	Introduction	Page 10, line 17: Suggest changing diabetic patients to patients with diabetes.	Thank you, we have changed this throughout and apologize if this was offensive. To avoid this in future AHRQ reports, we have mentioned this to AHRQ so they can remove this as an acceptable term in their report guidelines.
TEP 1	Introduction	Page 10, line 18: Throughout the report "support" is used to describe "DSME with or without support." Suggest defining "support" early on for greater clarity. There are numerous kinds of support that could be offered (eg, social, material, financial, educational) and the type of support referred to is not clear. This recommendation is included for each section of the report where the types of studies included are defined.	Thank you. We have meant to use the word support only for the phase of a program extending the DSME (hence our frequent use of DSME plus support), to provide clinical, psychosocial and/or behavioral support. We did not include studies that were only evaluating the support phase of a program. We have reviewed the report to revise any wording where support might have implied more than this phase of the program. We have clarified our terminology and definition of support in the background and Scope and Key Questions sections of the introduction, “For the purpose of this review we developed an operational definition of behavioral programs that encompasses DSME (without or with an additional clinical, psychosocial, or behavioral support phase, i.e., “DSME plus support”) as well as other programs incorporating interactive components that target multiple behaviors (e.g., diet and physical activity)

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			(see Appendix A).")
TEP 1	Introduction	Page 11, line 5: Calorie/caloric/kcal all refer to the energy content of food. Please use energy content instead of caloric content throughout the report. You could state it as energy (eg, calorie) ..... so readers understand what energy refers to. The energy content could be measured as kcal or kjoul.	Thank you for pointing out this mistake which has been corrected.
TEP 1	Introduction	Page 11, line 9: Suggest changing "...good blood glucose control..." to optimal blood glcuose control.	We have corrected this.
TEP 1	Introduction	Page 11, line28: Suggest changing "Because knowledge acquisition alone is not enough...." to knowledge acquisition alone is insufficient....	This has been changed.
TEP 1	Introduction	Page 12, line 26: Suggest changing "demographics" to demographic characteristics.	This has been changed.
TEP 4	Introduction	On Page 5 under "Objectives" in the structured abstract it is stated that the purpose is to review the effectiveness of behavioral programs for type 1 diabetes and identify factors contributing to effectiveness for type 2 diabetes. Isn't it really to look at effectiveness and factors contributing to effectiveness for both types? The different goals for the different types of diabetes is confusing as stated. I see on page 12 that it explains the reason for the different emphasis but I think this explanation needs to be more clear and included in the abstract.	These are our primary objectives for T1DM and T2DM, such that we don't want to suggest that we focus on, and include a detailed description on the effectiveness for T2DM. We have added the term "focusing on" in the abstract but word limitations prevent further elaboration.
TEP 5	Introduction	In ES on page 2, it should read...People with T2DM are often treated progressively though diet AND PHYSICAL ACTIVITY.	Thank you for pointing out this omission. We have made the change.
Peer Reviewer 1	Methods	The authors conducted a thorough literature search of appropriate databases and employed standard, as well as some novel, approaches to the analyses in order to glean the most information possible from this diverse body of research.	Thank you for this comment. No changes required.
Peer Reviewer 1	Methods	The reliability and validity issues were addressed thoroughly with standard procedures.	Thank you for this comment. No changes required.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 1	Methods	The authors provided clear descriptions of justification for their decisions regarding selected procedures, definitions, and target variables.	Thank you for this comment. No changes required.
Peer Reviewer 1	Methods	Heterogeneity analyses were conducted and presented within the context of the results.	Thank you for this comment. No changes required.
Peer Reviewer 1	Methods	The focus of the literature search was on primary studies published since 1993, which was well justified, and thus reflects the most recent evidence. The justification for including studies reported since 1993 is provided in the body of the report, but not in the Executive Summary. This would be an important point to insert in both places.	We have added the rationale for this decision to the executive summary.
Peer Reviewer 1	Methods	The authors addressed potential publication bias through a variety of methods, e.g., contacting experts in the field, searching for abstracts and contacting authors, and conducting funnel plots and Egger's tests.	Thank you for acknowledging this. No changes required.
Peer Reviewer 1	Methods	The process for assessing risk of bias of the data was thoughtful and logical, albeit stringent.	Thank you for this comment. No changes required.
Peer Reviewer 1	Methods	Setting the clinical importance/significance point of an A1C reduction at 0.4%-age points is justified but this reviewer questions whether this is actually a sufficient reduction. The authors clarify that there is little agreement on clinical importance thresholds so it would be important for future research for some consensus to be determined.	Thanks for highlighting this important point. We made this point (i.e., for consensus in the field around minimum clinically important differences) in the Potential Research Needs tables in the Discussion. However, we disagree with the reviewer regarding this/her doubts regarding the importance of a 0.4% reduction in HbA1c. The association between HbA1c and diabetes complications is continuous. Since a 1% difference in HbA1c is associated with a 37% relative difference in incidence of major microvascular complications of diabetes, a 0.4% difference, if sustained, would be associated with a 14.8% reduction in incidence of these long-term complications. To us, this difference is definitely clinically significant, although we agree that a larger impact would be better still.
Peer Reviewer 1	Methods	It is not clear whether the mean differences were weighted by sample size and variance, which is standard procedure. If not, then small and large studies	Studies were weighted by sample size and variance. We have added this detail into the section on Synthesis of T1DM in the executive summary and

		contributed equally to the overall MD. If the MDs were weighted, then a statement indicating the weighting needs to be added to the report.	full report.
Peer Reviewer 1	Methods	Categorizing studies as T1DM vs. T2DM based on whether the majority of the sample (>75%) were of one disease is somewhat questionable. It seems that including 25% of the sample with the other type of diabetes still has the potential to confound the results unless the data are reported separately.	We agree that distinguishing between types of diabetes is important (which is often not incorporated into reviews), and in the large majority of cases this was discernible from the authors' explicit statements for study inclusion or the clinical characteristics of the participants (low proportion on insulin). There were very few studies in the T2DM category that mentioned training in insulin injection/pump use. That said, it is important to create a priori decision rules in systematic reviews and this was used to reflect a balance between study exclusion and confounding by type.
Peer Reviewer 2	Methods	Methods are appropriate. There were difficulties in carrying out the full analysis for some aspects due to low numbers - specifically in type 1 and broader outcomes for type 2. Speaks to the lack of data and the need for comprehensive evaluations including implementation trials (quality improvement) that go far beyond examining clinical effectiveness. And I would argue we do not need anymore RCTs on the the impact of physical activity on glycemic control or other diabetes/cardiometabolic outcomes.	Thanks for providing this opinion. We agree that there were limited data on many outcomes including fitness, body composition, serum lipids and blood pressure (especially at followup longer than end of intervention). We also concur that research on implementation and quality improvement is important, and complementary to RCTs, in terms of helping to understand "real-world" implications and effectiveness.
Peer Reviewer 2	Methods	I have used date of inception in the past for systematic reviews; why 1993 as low book-end for bibliographic search? Why not open it up?	It was important to us to ensure that the context of the comparators in the reviewed studies represent (as possible) current guidelines for care. We provided this rationale in the section on inclusion/exclusion, "This date was chosen because of changes to usual care/medical management (the comparator in most cases in this review) resulting from the findings of landmark trials (like the DCCT) published from this date onwards." We have also included it now into the executive summary. This date was also considered during our consultations with Key Informants, and was thought the earliest date to incorporate the shift towards

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			behavioral approaches to DSME which was the focus of this review.
Peer Reviewer 2	Methods	Pg 17 - Eligibility Criteria - I am always interested to hear how usual care is defined. I have after been asked in the peer-review process to define as there are concerns as to where the line might be drawn between an active control and usual care control. The authors indicate "usual medical management" as to mean usual care - hard to know what this entails - suggest the authors may want to comment on this as it relates to interpreting results.	We agree that this could be clarified further in the report, especially since we take care to carefully distinguish between usual care and active controls in the analysis. We defined usual care as the clinical management received by all study participants, regardless of their study participation and how extensive this may be, such that the results between groups could accurately reflect the effects of (only) the behavioral program. Conversely, active controls received an additional intervention that would theoretically (if beneficial itself) reduce the relative effectiveness of the behavioral program. We wanted to avoid classifying as usual care those arms that actually received an intervention that may confound the effects of the behavioral program. We have made edits to the section on Inclusion/Exclusion to help clarify this.
Peer Reviewer 2	Methods	Pg 19: Table A nicely outlines program components and delivery factors.	Thank you. We are glad you found this table helpful.
Peer Reviewer 3	Methods	Methods for this report are very clear described and well justified. Standard procedures are reported and cited appropriately. Statistical methods are appropriate to the key questions for this report.	Thank you for this comment. No changes required.
Peer Reviewer 4	Methods	Inclusion and exclusion criteria clearly stated and defined. The included types of design included information from initial scan and were justifiable.	Thank you for this comment. No changes required.
Peer Reviewer 4	Methods	ES-23, lines 39-42 “although we analyzed change from baseline scores when able, the differential effect of behavioral programs based on these baseline imbalances (e.g., HbA1c, age)—as suggested by our subgroup analyses—cannot be ruled out.” Please note that using change from baseline score	Thank you for pointing this out; we have deleted this comment. We did not perform sensitivity analysis because of the sole use of RCTs and the small number of studies showing baseline imbalances (as assessed during our risk of bias assessment).

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		does not address the issue of baseline imbalance and the EPC guidance recommends sensitivity analyses using both change score and follow up score to estimate mean difference between intervention groups	
Peer Reviewer 4	Methods	Page 14, lines 56 Report what method is used to impute SD from similar studies.	We have revised this in the main report the section on Data Synthesis, "If computation was not possible they were estimated from upper bound p-values, ranges, inter-quartile ranges, or (as a last resort) by imputation using the largest reported SD from the other studies in the same meta-analysis."
Peer Reviewer 4	Methods	Page 15, line 20-21, Clarify exactly what SD is used to define clinical meaningful difference.	We revised this in the main report the section on Data Synthesis, "..we used a difference of one-half standard deviation (i.e., 0.5 SMD) based on the mean SD from the pooled studies which has been shown to represent a universal, conservative estimate of a meaningful difference"
Peer Reviewer 4	Methods	Page 15, line 33-34, Clarify why 10 minutes is chosen as the duration of calls.	We have added this explanation to the methods in the section on Data Synthesis. When calculating contact hours, we assumed telephone calls (when described in number and serving as more than a reminder/basic followup) would be 10 minutes each if their duration was not reported; this was based on reviewing studies from our preliminary searches that indicated most followup calls were reported as approximately 15 minutes (variable compliance) and that the duration of calls used for providing more substantial content were often not reported.
Peer Reviewer 4	Methods	Choice of Hartung-Knapp-Sidik-Jonkman random effects model a. I understand that there are different opinions about the choice of random effects estimates. While Hartung-Knapp-Sidik-Jonkman estimate is shown to result in more adequate error rates than the DL method in the recent paper by IntHout et al., , its overall performance has been shown not to be better than the DL and profile likelihood methods (Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study.	Thank you for these comments. We recognize that there has been no consensus on one best random effects model and that a different one may be suitable especially for meta-analysis having few studies and/or non-normally distributed effects. It is our preference to keep our model which was chosen a priori, and re-considered and approved during the peer review of our protocol. We do not think our strength of evidence assessments would be affected by changing methods, since most comparison with 1-3 studies (e.g. figures 8-10) were graded as insufficient due to inconsistency (e.g. 1 study as now

		Stat Methods Med Res. 2012 Aug; 21(4):409-26.). The latter paper specifically commented that "It is also of interest to note that … and SJ, despite having been expressly developed to address the limitations of other methods in the estimation of between-study variance, were frequently outperformed by those methods." (The latter paper did extensive simulations of non-normally distributed data, but this is probably true for most data used in MA). b. One major concern about the HKSJ method is that it is too way conservative, in particular, when the number of studies is small (where the method is meant for). The central idea to use a t-distribution with df based on the number of studies makes sense in the context of meta-regression comparing effect measures across study level estimates, but may make less sense when getting a combined estimate, in particular, when randomization has been done on patient level for the included studies. (so we have a lot more df within each study). Such concerns are clearly demonstrated by some of the results shown in this report. For example, Figure 8, adults results, the 95% CI for the combined estimate is much wider than the 95% CI for each included estimate while I2 = 0%. Also, Figure 9, youth results, Figure 10, adult results; Figure 13, Adult results etc. – it does not make sense the combined estimates have much more uncertainty when the included estimates are basically consistent. I would recommend the profile likelihood method, which seems to provide a good balance between power and type I error rates (Kontopantelis, 2012).	reported for adults in fig 8 & 10) or imprecision (from CIs crossing MID thresholds and 0, and/or small sample sizes) together with moderate/high risk of bias. Moreover, we don't feel that it is now appropriate to change methods based on the results (i.e., wide confidence intervals). We have revised our comments about the limitations of this method in the discussion.
Peer Reviewer 4	Methods	For the number of studies for meta-regression, while Cochrane review advised a minimum of 10 studies, the EPC guidance "advise a slightly different rule of thumb than the Cochrane handbook that when the sizes of the included studies are moderate or	Thank you for pointing this out. We acknowledge that it may be important to delineate criteria specific to continuous vs. categorical variables; however, in practice this would not have changed our methods or results. Thus, we have decided to keep our

		large, there should be at least 6 to 10 studies for a continuous study level variable; and for a (categorical) subgroup variable, each subgroup should have a minimum of 4 studies." While all numbers are arbitrary, it makes sense to have some different considerations for continuous vs. categorical study level variables. If a categorical study level variable has 2 levels, it will be different from a study level variable with 4 levels.	reference and methods (defined a priori) as the EPC methods suggest alteration only in the case of moderate to large sample sizes, which was not the case, especially for the only other outcome (generic HRQL) that had between 6-8 studies (8 potentially allowing analysis for a 2-level categorical variable). For T2DM our meta-regression was used for our subgroup analysis in KQ6 which was intended only to be performed for the outcome reported by the most studies. Moreover, the number of studies in these KQ6 subgroups (22 & 31) was still too small for the inclusion of several variables as would benefit from multivariable analysis.
Peer Reviewer 4	Methods	For publication bias, better to call it a small study effect with publication bias being one potential reason.	Thanks for this suggestion. As mentioned in the discussion of the Limitations of the Comparative Review, our assessment of publication bias included several considerations including our comprehensive search for unpublished study results (not yet published potentially due to small effect sizes) and our assessment that many of the studies were small or had unfavorable outcomes. Moreover, we would like to maintain the same terminology as used in the Methods Guide for grading the strength of evidence.
Peer Reviewer 4	Methods	I have a hard time to understand the methods for KQ2 (report text page 17). a. "We searched for subgroup analyses reported by individual trials that focused on whether a particular behavioral program was more or less effective for the outcome with the most data…" -- what does "the most data" mean here? If two outcomes provide subgroup data and one outcome has more than the other, will you only include subgroup data in one of them? Please clarify. Also, are you trying to look at within-study comparison in this search? b. "We also considered the studies themselves as units	Thanks for asking for clarification. We have revised this section to make it explicit that we assessed the same outcome for all subgroup analyses based on our decision rule of focusing on the outcome reported by the most studies which was HbA1c. This is quite common for subgroup analyses in systematic reviews. We added the words within-study and between-study for further clarity.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		for possible subgroup analysis, for example when…" – Are you looking for between-study comparisons here?	
Peer Reviewer 4	Methods	KQ3: univariate meta-regression makes sense. About using data from each study's longest follwoup timepoint -- how much variation of the followup timepoints is there for the included studies? What is the particular reason to include the longest followup? Are the results from longer followup more important than the shorter term effects? Also the impact of attribution will be different for data from different time points and how will it affect the results when the different length of follow ups were combined? No explanation of how to handle studies with more two arms.	Thank you for asking us to clarify these points. We have added some explanation in the section on Limitations of Comparative Effectiveness review about using longest followup, "Our analyses for T2DM should be interpreted based on our approaches to address program durability and the relatively high-level categorization of program components. Our network meta-analyses and subgroup analyses used outcome data at longest postintervention followup, which for the majority of studies was end of intervention (i.e., after all contact between participants and program personnel ceased) or, for fewer, between 1-6 months followup. Only eight trials had followup longer than 6 months. This approach was used to include as many studies as possible (i.e., those that did report data for end of intervention) and also to reflect the durability of the programs in terms of their potential for impacting long-term health." Because few studies in the NMA had followup longer than a few months, we feel that the influence of differential attrition in the studies at longer term followup would have a minimal effect on the results. The differential attrition rates in the included studies were by no means limited to (or higher with) timepoints longer than end of intervention. Our meta-regressions only compared interventions to usual care control arms; if there were more than 2 intervention arms in a study the sample size of the usual care arm was adjusted accordingly.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 4	Methods	It makes sense to conduct a net-work meta-analysis to compare the different the difference interventions/factors given the number of RCTs and scenarios. However, the investigators need to provide more information on model specifications. Currently there is no enough information to evaluate the validity of the specific model. In particular, a. Please provide an appendix to show the model formulation and the WinBUGs codes to fit the model. b. Explain how the model reserves the within-study randomization. The results from studies with more than 2 arms are appropriately included? c. For both KQ5 and KQ6, longest time points were used? See the comment about using data from longest time point above. As shown in the results, effectiveness differed a lot by time points and how do you justify the situations that there might be more differences in earlier time points but not in later time points? The impact of attrition? d. Consistency and heterogeneity: the results indicated that there is very high heterogeneity among included studies and how do you handle this in the network MA? Does this cause inconsistency? There is no mention on results about inconsistency/consistency, and no information on the heterogeneity measure in the results section, either. e. Based on the results, the investigators seem to create a lot of nodes (like dummy variables) to compare the results from specific combinations on program parameters, which is helpful to show which combination works best. However, such an approach does not take absolute difference into account (only relative ranking). KQ5 is asked to evaluate the moderation of effectiveness on program characteristics and given the large number of RCTs, I would recommend fitting a multivariable Bayesian model to specifically evaluate effect modification by program parameters (in that matter, and patient characteristics if possible) while controlling for other	Thank you for your suggestions and questions to help us clarify our rationale and methods. a. We have added a note in the report that the model formulation and WinBUGS codes can be obtained at request of the authors. We have not included this information in previous reports for which we applied network meta-analysis (e.g. Acute Migraine Treatment in Emergency Settings), and do not feel this is of interest to the large majority of readers of this report which is already thought too long. b. We have added to the Methods for KQ5 a statement on how the method preserves the within-study randomization. We also added, “These methods ensure that correlation in multi-armed trials is preserved.” c. We have added a comment in the discussion on why we used longest followup . Only 8 of 112 studies had followup at >6 months and we do not feel that differential attrition in these studies was any worse than in the studies with shorter followup. d. We used the network meta-analysis to try to tease out the reasons for some of the heterogeneity in the pairwise results, which was thought to result largely from variation in the factors we were assessing. This was mentioned in the discussion, “There was substantial statistical heterogeneity in these pairwise meta-analyses, supporting our subsequent analysis for KQs 5 and 6 to determine which program factors, and population characteristics, influence (and optimize) the effects.” We have added the findings for inconsistency for the network meta-analysis to the results section, “A consistency analysis was performed for the HbA1c analysis and it was found that only two quadratic loops (of a total of 43 total quadratic and triangular loops) showed statistically significant inconsistency.” We agree that creating multiple nodes was necessary for our aim to compare the effects of combinations of variables. Our approach focused on

		factors in the model. It will provide estimates (with 95% credible intervals) to characterize the differences by the program characteristics, which could also possibly estimate the mean difference and produce a ranking of the different combination of program characteristics. This approach may also explain some heterogeneity in the data and use all studies for each variable/characteristic. f. Thinning is needed to get relatively independent posterior samples.	looking for synergy between the various program components, rather than trying to control for any factor. KQ6 assessed the effect of patient variables, and including these study-level factors in the main analysis may have further introduced potential bias and heterogeneity in the results. We have added a statement that the results were interpreted in relation to their relative measures and that we did not make any conclusions base on comparisons between single nodes (of which there would be 561 potential comparisons) . We have also added statements in the results for KQ5 on our rationale for only using the variables we did for the groupings of studies, and we don't feel that changing our analysis would increase the number of studies available for each variable, e.g., no other variables were considered reliable or valid for use, and the one which did not use data from all studies (delivery personnel) showed no indication to moderate the effects in the large group of DSME studies within which it was assessed. In summary, we agree that using a multivariable Bayesian model is a potential alternative approach to this question. We feel that our a priori chosen approach using network meta-analysis achieved our objective of identifying the relative effectiveness of different combinations of program components. The utility of this approach and value of the results is reflected in the positive response from the range of peer-reviewers, including comments that the results and presentation were clinically meaningful. f. We did run a sensitivity analysis where we thinned the MCMC to every 10th iteration. The results were virtually identical to our regular run, so independence of iterations did not appear to be an issue. We added a note on this in the methods section.
TEP 1	Methods	Appropriate methodology is used and described.	Thank you for this comment. No changes required.
TEP 4	Methods	No concerns. The methods were very clear.	Thank you for this comment. No changes required.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

TEP 4	Methods	On page 29, line 43--there is a typo. I believe it should say "have a moderate to high.."	Apologies that we cannot locate this typo. A search of the methods section for the words high and moderate did not help.
TEP 5	Methods	Comments were made about "behavioral" programs and then separately for "lifestyle" programs. From the definition provided on page ES-10, it does not appear that "Lifestyle programs" focus on diet and activity alone but may include other components. Because of the success of the behavioral lifestyle intervention from the Diabetes Prevention Program (granted the participants did not have diabetes but were just at high risk) it would be valuable to examine the subgroup of studies focusing on lifestyle behavioral programs alone in the T2DM section.	We classified lifestyle programs as one form of a behavioral program (see Table 3 and the section on data synthesis in the report) and to be included all programs (whether educational or lifestyle) had to include some form of behavioral technique/ approach. We feel our current approach to analysis of KQ5 captures the differences between the various forms of behavioral programs included. We added a sentence into the section on Scope of Review to help clarify: "A commonality with all programs was that they incorporated one or more behavior change techniques, with or without an explicit use of a theory or model of behavior change."
Public Comment (Kelly McDermott, Omada health Inc.)	Methods	The word diabetic should never be used in printed materials. No one who works in diabetes and is credible and no diabetes organizations use that word. People who work in diabetes will immediately dismiss what you have to say as an outsider in this field.	Thank you for pointing out this error and we apologize if it was considered offensive. To avoid this in future AHRQ reports, we have mentioned this to AHRQ so they can remove this as an acceptable term in their report guidelines.
Public Comment (OMada Health Inc.)	Methods	We have two concerns about the *means of communication* variable. First, there is a wide range of technologies and technology uses in behavior change interventions. While we understand that any review must broadly group interventions, we would argue that the categorization of the *means of communication* variable as "in-person vs. some technology" obscures some of the more sophisticated and forward thinking technology uses. As an extension of this, in many cases it may not be appropriate to view technology as an isolated intervention component at all. For example, technology might enable tailoring, or enable high intensity interaction and really should be evaluated in the context of these other variables. Evaluating technology as a stand alone, dichotomous variable threatens to oversimplify its potential impact on behavior change.	We certainly appreciate your concern that our methods did not allow for capturing the potential benefits of technology, which is complex and may offer benefits through various mechanisms. We tried to avoid categorizing program factors which overlapped in meaning, partly by way of our analysis which looked at combinations of factors, but realize this is challenging. This is similar to the reason we didn't include community engagement as a variable in the analysis - because this feature overlaps with delivery personnel using lay providers which are often peers from the community (see explanation added in the section on Detailed Synthesis for Key Question 5). Because we weren't focusing on programs delivered solely by technology, the large majority of the included studies used unsophisticated forms of

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		Second, technology based interventions may appeal to a different underlying population. One could argue that committing to regular in-person meetings implies a level of self-discipline and scheduling acumen associated with behavior change success. In other words, technology-delivered behavior change interventions may appeal to a less motivated population. We suggest that authors include more in-depth reflection on these issues around the use of technology in the discussion section.	technology such as the telephone or video conference; with our exclusion of programs having a disease management focus (e.g., monitoring of disease status), technological advances in this respect were not reviewed. We have also added a point to the discussion to this effect. "As stated in the Results chapter, we did not include program tailoring and degree of community engagement in the analysis for KQ5; these factors were considered to overlap in meaning to some extent with delivery method (e.g., use of technology enhancing tailoring) and delivery personnel (e.g., use of nonhealth care providers providing community engagement), and the ones we used were thought to better represent the differences between the programs assessed in this review. With our focus on programs incorporating interaction with program personnel, we cannot comment on the effects of programs delivered entirely by way of technology which may provide sophisticated mechanisms to interact with and motivate participants or closely monitor disease management."
Peer Reviewer 1	Results	The findings are presented from the perspectives of both statistical significance as well as clinical importance. Decisions regarding targeted thresholds for clinical importance of outcomes were determined a priori by expert consensus.	Thank you for this comment. No changes required.
Peer Reviewer 1	Results	In terms of potential harm associated with behavioral interventions, wouldn't hypoglycemia also be considered "harm"? Did any studies report any data on number of hypoglycemic episodes?	We agree that hypoglycemia events (if increased) would normally be considered a harm, although as for several outcomes (e.g. anxiety, quality of life) the direction of effect will largely determine whether they are considered harms or benefits. We classified activity-related injury as a harm that could be directly attributable to the program. Hypoglycemic episodes was included as an outcome for T1DM (see Table 1), and was reported by a few studies as indicated in Table 4. A note on the limited data for this outcome

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			was added to the Key Points for Other Clinical and Behavioral Outcomes, section on Key Question 1. The executive summary also includes a comment to his effect.
Peer Reviewer 1	Results	The total number of citations listed in the flow diagrams (e.g., Figure C and Figure 3) differs from the number presented in the narrative section. This is confusing.	Thank you for pointing this out. Our narrative included the additional studies identified by reviewing reference lists, but we realize that this seemed inconsistent because the additional studies are shown near the end of the PRISMA flow diagram. We have addressed this, also accounting for the numbers from our search update performed after the draft report was submitted.
Peer Reviewer 2	Results	The results are presented comprehensive and in a number of different formats. I did appreciate this approach. The study characteristics we fairly well described - table format does provide this important information.	We are happy you found the formats acceptable. No change required.
Peer Reviewer 3	Results	The results section is long, though the length is due to the detail describing the included studies and the very large number of studies in this report.	We agree that the report is long due to the number of key questions and focus on both T1DM and T2DM.
Peer Reviewer 4	Results	The amount of detail is appropriate and there are good summarizations of important study characteristics.	Thank you for this comment. No changes required.
Peer Reviewer 4	Results	Page 23, lines 20-21, clarify HbA1c > 7 are baseline values?	We have clarified this as baseline data.
Peer Reviewer 4	Results	Good to separate objective vs. subjective measures in ROB assessment.	Thank you, we agree. No changes required.
Peer Reviewer 4	Results	“further, because the 95% CIs included our threshold for clinical importance we cannot rule out benefit for behavioral programs.” – I would only say this for still relatively “precise” estimates. For imprecise estimate, could not rule out either benefit or harm as there is just no adequate information.	Thank you for the comment. These comments were stated when only the clinically important value favoring behavioral programs was included in the 95% CI, hence they were precise enough not to include clinically important effects both for and against behavioral programs (for which we downgraded to insufficient).
Peer Reviewer 4	Results	KQ1 Any insight why comparison to active control has bigger effect size than comparing to usual care?	We had included a potential explanation for this in the discussion (e.g., less bias from lack of blinding leading to co-interventions etc. in these studies).

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 4	Results	KQ1 it seems that the observational studies may have larger effect size? Any insight?	Looking closely at these 3 studies provided little insight as to why 2 of the 3 had clinically important effects. We have added a comment in the Detailed Synthesis for T1DM (HbA1c for usual care comparisons) that the medium risk of bias likely indicted that the results should be interpreted with caution; the only study having low risk of bias found no difference. "These results should be interpreted with caution because of concerns with bias and confounding in observational studies; the only study assessed as having low risk of bias found no difference."
Peer Reviewer 4	Results	Figure 8, Adults, does Weinger 2011 have 74 patients in the behavioral program? The data seem to be double counted here. Given the way the data were analyzed, the two arms of usual care should be combined. This same issue applies to Figure 9 and Figure 10.	Thank you pointing out this mistake – this study had 2 active control arms which are now combined for this analysis.
Peer Reviewer 4	Results	Table 4, units for lipid variables (HDL, LDL, etc) are not provided.	We have provided these units.
Peer Reviewer 4	Results	Figure 14, what is the n for Husted 2014 in the Behavioral program? Missing?	Thank you pointing this out. This has been fixed (n= 37).
Peer Reviewer 4	Results	p44, refer Table 9 in the results of KQ3.	Thank you for pointing this out. We have made the change.
Peer Reviewer 4	Results	P49, line 50-51: (MD 0.90; 95% CI, 0.90 to 0.90)?	This is correct; the studies had the exact same effect size which resulted in the CI calculated using HKSJ.
Peer Reviewer 4	Results	P51, First paragraph, what SD was used to determine clinically important difference?	We have added that the mean SD of the pooled studies for that comparison was used.
Peer Reviewer 4	Results	Results for KQ5 and KQ6: as mentioned in the comments for methods, there was no information on the heterogeneity and consistency of the included data.	We have added this in the Results under Detailed Synthesis for Key Question 5, "A consistency analysis was performed for the HbA1c analysis and it was found that only two quadratic loops (of a total of 43 total quadratic and triangular loops) showed statistically significant

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			inconsistency.”
Peer Reviewer 4	Results	P53, estimates of baseline BMI are obtained by simply averaging over the studies, or from the model? Or??	We added a note that this was the mean baselines from the studies.
Peer Reviewer 4	Results	Table 10, I would suggest to adding the sample size in each node. Some of the nodes only incorporated a very small number of subjects , but have a relatively high ranking with a wider CI (for example, node 14) – would you really trust the ranking for this one?	We have added the sample sizes. We also added a comment on the interpretation of the results for HbA1c (Detailed Synthesis), “When interpreting the results, we relied mostly on the relative ranking of the nodes, and looked for trends in the findings based on program variables that appeared to determine whether the effects would offer clinical benefit. Some nodes had very few studies, small sample sizes, and/or wide credibility intervals, thus we did not make any firm conclusions for a single node (or for differences in 561 potential comparisons) but rather from looking across nodes with similar features.”
Peer Reviewer 4	Results	Also most MDs are not statistically significantly different from 0 (not different from usual care) in tables 10 and 11; for table 11, the top 3 rankings have two MDs not significant and one of them has a wider CrI, which raise concerns about the utility of such ranking.	We were interested in the relative effects based on different combinations of program components; therefore, we relied on more than one node to interpret the findings. Given the results of the pairwise comparisons showing moderate effects at best for these programs, it is not surprising the most MDs were not statistically significant in the network meta-analysis. We believe that the approach we used and the focus on ranking, or relative effectiveness across nodes, helps discern which combinations of program features are likely to yield the most favorable results. Through this approach we were able to interpret that greater program intensity and in-person delivery seem to be moderators of effectiveness.
TEP 1	Results	See comment above about differentiating the intensive phase of interventions from the maintenance phase of interventions. Usually, the impact is greater following the intensive phase and weakens at study end due to limited or lack of continued education and support. It is	We have addressed this with a paragraph in the discussion on Limitations of the Comparative Effectiveness Review. The lack of studies of DSME plus support or focusing on lifestyle (both often including lower intensity phases for support or

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		probably beyond the scope of review to separate the contact hours by intervention phases in the tables. However, a statement regarding the reduced contact hours is needed. Otherwise, it appears that the number of contact hours is equally dispersed throughout the intervention.	maintenance) in T1DM makes this statement most applicable for T2DM.
TEP 1	Results	Page 21, lines 43-49 and Page 32, lines 32-43: When discussing the medium to high risk of bias across studies due to lack of blinding of participants, study personnel, and outcome assessors, need to add a statement that blinding participants and study personnel in behavioral studies is very difficult, if not impossible. Participants are aware of the treatment they are receiving and it is not possible to blind the treatment. This point is addressed in the full report but needs to be added to the Executive Summary.	We have added a sentence to this effect in the executive summary section on limitations of the evidence base, "Blinding of participants and personnel are arguably difficult for trials of behavioral programs especially when the comparator is usual care. According to our decision rules for assessing ROB, a low ROB for participant and personnel blinding was granted if the comparator was an attention or active control and the authors stated some means to blind the study hypothesis from participants, and if there was a structured training and protocol followed for the personnel. Participant blinding in this manner was rarely reported. Similarly, blinding of outcome assessors, highly feasible in any situation, was rarely reported or sufficient."
TEP 1	Results	Page 90, lines 26-40 and throughout the report: Suggest adding the rationale for how the low, medium and high intensity programs are defined. That is, why was < 10 contact hours defined as low intensity? Why 10 hours as the cutpoint? This is an important point as the impact of low intensity programs vs. the moderate/high intensity programs had a differential effect.	This is a good point. We have added a description in the data synthesis section of the report (and a footnote to ES Table A),"The cut-points used for creating the intensity categories were based on practical considerations. The 10-hour "minimal intensity" limit was based on the current number of hours billable for patients eligible for public healthcare administered by the Centers for Medicare and Medicaid Services in the United States; this was described by our TEP as an important practical limitation on implementing programs having higher intensity. The value of 27 hours was based on what would be considered the lower end of highly intense (e.g., at least weekly 1-hour sessions for 6 months)."
TEP 4	Results	Throughout I believe "followup" should be separated into 2 words or a hyphen used.	We have followed the AHRQ style guide.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

TEP 4	Results	Page 32, Table E of research gaps is very important but without the key questions listed it is more difficult to interpret and requires the reader to flip back to the KQs. I recommend you include the questions in the table that identifies the research needed. Same issue on page 113, Table 13.	Thanks for this good suggestion. We have added a short term (e.g., Effectiveness for T1DM, Moderating factors for T1DM) to indicate which question is most relevant to the recommendations in these tables.
TEP 4	Results	Page 33, line 47, the word should be changed to good not god.	Thanks for pointing this out; we have corrected this error.
TEP 4	Results	In several places it is noted that interventions delivered in person seem to achieve more favorable results than those delivered by technology but as far as I can tell this is based on one technology study using Skype. Maybe I missed the section where it detailed the number of studies and types of technology used but if it is just the Skype study that seems like insufficient evidence to base the superiority of in-person treatment.	The study using Skype was specifically noted because it was the only head-to-head (comparative effectiveness) trial for T1DM. The Key Question 5 findings of greater effects for in-person delivery were only relevant for T2DM (i.e. for T1DM the univariate meta-regression analysis in Key Question 3 did not allow for any conclusions in this respect), for which 16 trials used only technology (described in Characteristics of Included Studies) and many other incorporated technology (email, telephones) as a form of delivering the program, as shown in the tables describing the nodes of interventions for the network meta-analyses.
TEP 5	Results	When presenting the results, caution right up front is needed in interpretation of some of the univariate findings. For example, in the statement about programs offered to predominately minority versus predominately non-minority participants, these studies are not stratified by glycemic control. This fact should be mentioned in the results and discussed in the discussion.	We agree that this could be made explicit in the results section and emphasized more in the discussion. We added a sentence in this respect within the Results (Detailed Analysis for KQ6), Discussion (Key Findings and Discussion for KQ6) "All of our results for this KQ relied on between-study rather than within-study comparisons, such that the effect of randomization is removed and the results are considered observational and possibly biased through confounding by other study-level characteristics." We also added a comment in the executive summary. We also added the difference in baseline HbA1c in the ethnicity subgroups as a Key Point for KQ6.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 1	Discussion	Studies were stringently assessed for risk of bias, which is a positive feature of this review. However, it is questionable if total blinding can be accomplished, and should be expected, in behavioral studies. The most that can usually be accomplished is blinding of the assessor(s). So would the authors recommend any changes to the risk assessment, based on their experience with this review?	Thank you for agreeing that we assessed the risk of bias with rigor. While we agree that blinding of participants and providers is difficult, the risk of bias still remains in any study where blinding is not used. We did allow for "equivalents" to blinding as described in the methods section (also see Supplementary materials), and feel that this is one option that accounts for ways some study designs (e.g. active control versus usual care) may have relatively lower risk than others.
Peer Reviewer 1	Discussion	There is very little discussion of measurement issues in the report. It would seem that some of the problematic findings and barriers to conducting some of the intended analyses, plus the high risk of bias in many of the analyses, might have resulted from measurement difficulties. There are few excellent measures to assess behavioral change (e.g., dietary and physical activity changes); and many researchers continue to rely on self-reported measures. These measurement issues are particularly problematic in underserved populations with low literacy rates.	Thank you for your comment. A paragraph in the discussion (Limitations of the Evidence Base) addressed some of these points but we appreciate your point about emphasis. When considering the moderator analyses, we have added , "Considering that behavioral changes are the key mediators to achieving clinical and health outcomes, analysis based on valid outcomes of changes to physical activity or diet would be ideal; greater use of these outcomes, especially via objective means, would be beneficial."
Peer Reviewer 2	Discussion	Yes, the research gaps section and specifically Table E. There is value in synthesizing the evidence to find there is no evidence. All to often we don't embark on these types of evaluations because we have a sense there is no evidence and then we just dont bother. Grouping these 6 key questions helps to build a body of knowledge in the areas. Now, we are also going to require the same approach for gestational diabetes and post gestational diabetes in relation to risk reduction for the development of type 2 diabetes. The authors may want to position themselves to include this in this version or a future version - the data are fairly minimal at this point but perusing clinical trials registries finds many trials underway targeting this important and understudy diabetes population.	We agree that risk reduction in gestational diabetes is also important to evaluate and that there is likely sufficient numbers of published trials starting to emerge to make this an appropriate topic for review.
Peer Reviewer 2	Discussion	Researchers will need to consider the cost of conducting trials to fill the research gaps - longer term follow-up is always a limitation of the behavioral	We agree that this is always an important consideration.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

		research in diabetes.	
Peer Reviewer 2	Discussion	Conclusion is sufficient - spelling mistake "god quality evidence"?	Thanks, we have corrected this.
Peer Reviewer 3	Discussion	Minor edit: ES-24 line 47 refers to "god quality" studies rather than "good quality" studies.	Thanks, we have corrected this.
Peer Reviewer 3	Discussion	The discussion is clearly broken down into sections describing limitations of the review, limitations of the body of evidence, and future research needs. The executive summary provides a succinct overview of these topics, and the main report goes into greater detail and could be pared down in places to more parsimoniously summarize the discussion points.	We realize the report is long although feel this was necessary considering the breadth of the topic and need to follow reporting standards.
Peer Reviewer 4	Discussion	The limitations of review/studies are described and the research gap is clearly listed.	Thank you.
Peer Reviewer 4	Discussion	Report text P68 lines 18-24, starting of third paragraph: Please note that the random effects model used in pair-wise MA is likely more conservative than the Bayesian, and this could cause discrepancy in results between the two approaches.	We agree that the approaches may give discrepant results if we were directly comparing effect sizes and statistical significance levels. We wanted to highlight in this paragraph that several factors may be at play when considering delivery format. We have deleted the sentence about the pairwise results to avoid confusion.
Peer Reviewer 4	Discussion	Report text P74, lines 27-28 “and may have missed some meaning.” -- Unclear text and please clarify.	We agree and we have taken out this sentence since we described the concept better in another.
TEP 1	Discussion	See comment above about adding the need for future research to address strategies for sustaining change in outcomes to the tables regarding research gaps.	This study did not focus on maintenance strategies (i.e. outcomes were not captured specifically for maintenance interventions or time periods), such that we cannot propose a research agenda in terms of this.
TEP 1	Discussion	Page 33, line 47: "god" should be good.	Thank you; this has been corrected.
TEP 4	Discussion	This section was very useful. It might be even more useful if it included a more practical take-away for the practitioner after reviewing the key findings.	Thanks for this comment. AHRQ does not make practice recommendations, so we tried to offer the evidence in a manner that a practitioner and other decisionmakers can review the key findings while also understand some of the limitations of the review and evidence base to make appropriate decisions

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			based on their context.
TEP 5	Discussion	For type 1, we really don't have enough studies to say much of anything. That needs to be clearer in the abstract and ES. For example, we don't seem to have enough evidence to say that, for K2, the effectiveness of behave programs compared with usual care for HbA1c appeared higher for adults than youth. There are too few studies done in adults to say much of anything as identified in the actual 77 page report.	We agree in terms of the comparisons with active controls, and have added a comment in the abstract and executive summary about this. The executive summary now states that "From the subgroup analysis for age in comparison with usual care, adults appear to benefit more at end of intervention than do youth; in comparisons with active controls the SOE was largely insufficient which precluded making any conclusions."
TEP 5	Discussion	Also, the comment about effectiveness of these programs in those with good versus suboptimal control on page ES-19 is a very important one and should be highlighted. One would expect this to be the case but many don't think this through.	We agree this was an important finding. This point is included in the abstract and the conclusions.
TEP 5	Discussion	Also, in ES-17. the statement was made that there was little evidence around the outcome related to changes in physical activity ..etc. Did studies show a significant change in activity so that this comment can be made, or is the issue that activity did not change or was not measured? Very important difference.	Very few studies reported on this outcome. We have revised the wording to this effect.
Public Comment (Kelly McDermott, Omada Health Inc.)	Discussion	Peer support programs are not designed to meet ethnic needs. They are designed to provide ongoing support. The statement that this designed to adapt to a particular ethnicity is not true.	We did not want to give this impression, but rather that those programs offered to minorities were often tailored in ways to make them more effective. On several occasions it was mentioned that the population was thought to prefer group delivery by peers. We have revised this statement (Discussion of Key Findings for T2DM), "Many investigators enrolling a large proportion of ethnic minorities in the trials included in this review adapted programs in ways to make them more culturally and linguistically acceptable—often including peers in the delivery or social support groups—which may have enhanced their effectiveness."

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

Peer Reviewer 1	Clarity & Usability	There are some formatting issues. A number of the figures are impossible to read, e.g., Figures A/B and 1/2 (analytic frameworks). Figures 4 and 15, risk of bias, are difficult to interpret in black and white print and most of the forest plots are too small to read. The forest plot in Figure 17 is much more readable.	We appreciate your point, and have tried to improve the formatting to enhance clarity, without increasing the page numbers of the report substantially. The report will be available electronically and will thus be able to be read in enlarged formats which will be helpful.
Peer Reviewer 2	Clarity & Usability	The report is well structured and clear. The results I believe can drive some research progress; in particular type 1 diabetes. AS the authors point out there is is a lot of room for cost-effectiveness research with the use of HC utilization data and this will have certain influence on policy - authors may want to add more than just one line in the discussion.	We have intentionally not added to the discussion regarding cost-effectiveness as the focus of AHRQ reports is on effectiveness without consideration of cost.
Peer Reviewer 3	Clarity & Usability	Overall this is a clear report using well justified methods. The conclusions appear to be sound and based on the quality assessment, results, and SOE stated in the report.	Thank you. No changes required.
Peer Reviewer 4	Clarity & Usability	Yes very well structured and organized. Main points are clearly presented.	Thank you. No changes required.
TEP 1	Clarity & Usability	The report is well organized and clear overall.	Thank you. No changes required.
TEP 3	Clarity & Usability	The report is well structured and organized. No additional clarifications are needed.	Thank you. No changes required.
TEP 4	Clarity & Usability	There is a lot of information to support the conclusions but a clear and concise executive summary of what was found and the research gaps that remain would make this report a lot more useful to the average reader who wants to get to the bottom-line much more quickly. It is a very dense report and the "key points" are helpful but it is sometimes a challenge to tease out the implications for each KQ. I think the report is very good for an academic but I am not sure it is a user-freindly as it needs to be to use it to policy or practice.	Thanks for sharing this concern. We have also submitted the findings in two manuscripts which will hopefully help in terms of accessibility.
TEP 5	Clarity & Usability	The report is a bit wordy, wish it could be cut down a bit.	Thanks for the comment. We have tried to be concise and have reviewed the report in full to see where this can be enhanced. The focus on both

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015

			T1DM and T2DM using six key questions with multiple outcomes makes this a considerable challenge. There is also some need for repetition of important concepts due to the length.
Public Comment (Kelly McDermott, Omada Health Inc.)	Abbreviations & Acronyns	DSME and DSMS are not the same thing in diabetes and should not be in this document.	Thank you. We realize that DSME and DSMS are not the same thing, although we needed to distinguish between those programs which solely looked at DSME and those that also had an additional DSMS component. Without this separation there could have been substantial bias due to the differences between these programs in content, intensity of contacts etc.
Peer Reviewer 1	Additional	Overall, this report is very well written, well organized, comprehensive, and a valuable addition to diabetes-related clinical guidelines. The limited positive effects of behavioral interventions reported in this review are disappointing and threatens to feed the narrative that behavioral interventions are time and personnel intensive with few clinical advantages. What is needed now is more detailed guidance for future research and more attention on measurement issues, common measures, and consensus on thresholds for clinical importance of targeted outcomes.	We appreciate your comment and time taken to review and provide comments on the report. We hope that this report was not interpreted as a clinical guideline, which it is not. We do hope that the findings and recommendations for future research is of value for researchers and multiple other stakeholders when considering implementation of these programs in their setting and context.
TEP 1	Additional	The statements regarding the impact of interventions for minorities is somewhat over-stated. Since the minority samples usually had poorer glycemic control, it is not clear if the interventions were more effective for minorities or more effective for those in poor glycemic control in general.	We appreciate this comment and have made revisions to the discussion of this outcome to highlight that several factors are likely contributory including poorer glycemic control.
Public Comment (Kelly McDermott, Omada Health Inc.)	Additional	As someone immersed fulltime in the field of diabetes patient education and research this reads like it was written by someone with an academic or peripheral interest in diabetes. It does not reflect the reality of what is happening in this field.	We appreciate your comments. It was our intent by way of using multiple stakeholders throughout the systematic review process to ensure we were capturing information that was considered important for multiple decision makers. Several of the peer reviewers (representing multiple clinical fields) have commented on the relevance and meaning of the report. We have made some revisions to the text, as suggested by the various reviewers, and we hope these clarifications have improved the interpretations

of the methods and findings.

Source: http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=2124
Published Online: September 28, 2015