Comparative Effectiveness Research Review Disposition of Comments Report

Research Review Title: *Comparative Effectiveness of Therapies for Children with Autism Spectrum Disorders*

Draft review available for public comment from July 09, 2010 to August 09, 2010.

Research Review Citation: Warren Z, Veenstra Vander-Weele J, Stone W, Bruzek JL, Nahmias AS, Foss-Feig JH, Jerome RN, Krishnaswami S, Sathe NA, Glasser AM, Surawicz T, McPheeters ML. Therapies for Children with Autism Spectrum Disorders. Comparative Effectiveness Review No. #. (Prepared by the Vanderbilt Evidence-based Practice Center under Contract No. 290-02- HHSA 290 2007 10065 I). Rockville, MD: Agency for Healthcare Research and Quality. November 2010**.** Available at: 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 comparative effectiveness 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.

Section	Comment	Response
General	This is an extremely detailed report that, to a large extent, summarizes well the state of the science with respect to what is known about the efficacy of therapies for children with autism spectrum disorders. It is a clinically-meaningful report. One weakness, however, is its failure to capture the large amount of important and clinically-relevant empirical evidence that is reflected in single-subject design studies. The authors should be prepared for some people to argue that single subject studies should have been included in this review, because many well-accepted systems for evaluating empirical evidence for efficacy of treatments consider multiple studies using single-subject designs to be a valid way to evaluate the efficacy of a treatment method. The area this affects the most is certain types of behavioral interventions for challenging behavior, specifically, functional behavior analysis and positive behavior support which have a wealth of single-subject data to support their efficacy in reducing severe challenging behaviors. It is recommended that this limitation of the review be explicitly recognized	We understand that single subject design studies are commonly used in behavioral research in children with Autism Spectrum Disorders (ASDs). Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the Agency for Healthcare Research and Quality (AHRQ) Evidence based Practice Centers (EPC) study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness. We have explained our inclusion criteria and rationale further in the report's Methods chapter.

Section	Comment	Response
General	There are a few places where general statements are made which arguably go beyond the existing data (described in more detail later in this review). Furthermore, there are places where there seem to be inconsistencies between the quality of the evidence supporting a particular treatment modality (e.g. social skills training) and the conclusions that are drawn about the state of the science.	We have attempted to revise the text as you suggest in your later, more detailed, comments. We have also reviewed our assessment of the strength of evidence for the interventions included in the report and made changes to the ratings for behavioral and educational interventions.
General	The rules for evaluating evidence don't always appear to have been uniformly applied throughout the paper. Nevertheless, this report represents a tremendous amount of work and an extremely comprehensive (although not exhaustive) review of the existing literature on the efficacy of therapies for children with ASD. This report will be very useful as a benchmark for assessing progress in the field in the future	We reviewed our study quality and strength of evidence assessments to ensure uniform application and made corrections to the ratings for behavioral and educational interventions.
General	Minor comments: Instead of “medical,” perhaps the term “biomedical” should be used to describe pharmacological and other biomedical interventions, since behavioral interventions are considered part of medical intervention according to some classification systems.	We acknowledge that the term ‘biomedical’ is often used to describe certain treatments in the autism community, but it lacks a clear definition that differentiates it from ‘medical.’ To avoid the confusing connotations of the term ‘biomedical,’ which is sometimes employed by providers advocating particular underlying theories of autism etiology and treatment response and other times used synonymously to ‘medical,’ we have left our terminology as "medical" throughout the text. Treatments described in this category are focused on delivering an exogenous substance, typically a medication or supplement, and in our a priori categorization of therapies we determined that such treatments would be categorized as medical. We recognize that there are multiple approaches to categorizing treatments. More information about our categorization approach can be found in the report’s Methods and Discussion sections.

Section	Comment	Response
General	This report does not seem to focus explicitly on comparative effectiveness of different treatments or responses of different subgroups to the same treatment. In fact, this is noted as a topic on which there is a paucity of data. Thus, the title does not quite fit the content of the report	This review is considered a comparative effectiveness review and adheres to the AHRQ's standards for such reviews; however, we were limited in making comparisons among therapeutic approaches by a lack of data available in the published literature. We tried to identify elements such as subgroups of populations which may benefit from interventions but data are not yet available to do so with confidence. The Discussion section of the report contains more information on this issue.
General	This is an ambitious undertaking. The complexity and multi-disciplined nature of the literature will likely produce differences of opinion regarding the quality of the review regardless of its quality. That being said, this reviewer was asked to give an opinion regarding this literature review. The strengths of the review are many (multiple judges, systematic review protocol, the presence of guiding questions, use of advisory panel, the attempt to differentiate good-quality studies from poor quality studies, emphasis on replication of findings, presenting on which quality criteria studies were (not) credited, etc.). However, for the sake of attempting to provide input that may aid the authors in revising the paper or may aid the editor in making a decision regarding publication of the paper, the remainder of the review will focus on the aspects that might be improved.	Thank you for your comments.
General	The review is likely to provide valuable information for consumer, providers, and families. The questions to be addressed in the review are stated concisely. Procedures for conducting the	Thank you for your comments.

Section	Comment	Response
	literature search, including or excluding reports, extracting data, rating the quality of studies, and synthesizing the evidence are generally clear. Methods and findings from individual studies are described accurately. The conclusions are for the most part consistent with the evidence.	
General	However, the categories of interventions (behavioral, educational, etc.) do not always appear logical, and there are some inaccuracies in the description of particular intervention approaches.	It is challenging to categorize interventions for ASD, and no one categorization approach is uniformly accepted. While we understand your concerns, we do not feel that any other approach would be acceptable to all experts in the field, and so we will maintain our current organization, and maintain as much clarity as possible about the choices we made. We have also attempted to make this concern transparent and further delineate that we explicitly examined what, if any, impact alternate organizational approaches would hold for evaluating the strength of existing evidence.
General	Also, an explanation is not given for some of the inclusion/exclusion criteria and for some of the categories for rating quality.	We have added text to the Methods chapter to clarify our inclusion/exclusion criteria and quality rating criteria.
General	Moreover, there are some discontinuities between the criteria for rating individual studies and criteria for rating overall evidence	We have reviewed our approaches to quality assessment and rating the strength of evidence for various interventions to ensure consistent application. We have made corrections to the ratings for behavioral and educational interventions in the Results and Discussion sections of the report. We appreciate the reviewer's attention to detail and knowledge of the literature.
General	In addition, there are some sweeping statements, often containing the word "uniformly," that do not withstand scrutiny.	We have re-examined such statements throughout the review and revised them.

Section	Comment	Response
General	Clarity and Usability: The report is written clearly. It provides a valuable summary of the current state of the science in autism treatment research, and it offers directions for future research, although it does not present recommendations for practice decisions. In its current form, it is usable by doctoral-level professionals but would be inaccessible to most others.	Our partner center, the Eisenberg Center, will be producing materials from this report designed specifically for patients, providers and policy makers.
General	The report is very worthwhile and directly addresses clinically meaningful questions. Little information was available for many of the questions but the lack of information was addressed systematically and fairly.	Thank you.
General	Clarity and Usability: Very well-structured and organized. A masterpiece! A really, really impressive effort. Well-written and thoughtful beyond a usual review. Definitely moves the field up a step in terms of knowledge.	Thank you.
General	Though the content of the report would be clinically important, the discrepancy between the actual findings and how they are summarized in the abstract/executive summary lessen the clinical utility of the report. The key questions are appropriate. However, there is little to no mention of the subquestions anywhere other than the Introduction.	We have revised the abstract and executive summary to reflect more accurately the findings of the report. We have clarified in our introductory text that the sub-questions are not addressed individually, but as components of the key questions. This is because, although we intended to address them separately, upon review of the literature itself, it became clear that these distinctions were not reflected in the available data.
General	As discussed above, the report needs significant editing to increase clarity and usability. The main points are not clearly presented and need significant revision prior to being able to inform policy and/or practice decisions.	We have revised and edited the report throughout to improve clarity.

Section	Comment	Response
General	I found this report to be meaningful, but it generally confirmed my personal knowledge of the ASD scientific literature and my personal experience over 30 years. I did not find any information that was contradictory. I thought the key questions were clear and thorough.	Thank you.
General	The report is formidably long, but well organized and structured. I found it logical and easy to understand. The salient information and conclusions are clearly stated. As I previously stated, this report reinforces my current knowledge and clinical approach; I don't think it will generate much change in my practice.	Thank you.
General	Target population and audience is explicitly defined. Questions are appropriate.	Thank you.

Section	Comment	Response
General	Despite the exclusion of many important articles, the manuscript is well-written and well-organized. A considerable amount of work was put into the manuscript, and it does provide a helpful guide for many common interventions.	We restricted the review to those studies published from 2000 forward in consultation with our Technical Expert Panel and Task Order Officer and based on the following points: -Good coverage of older literature in several published systematic reviews (e.g. National Standards Report,[3] *Clinical Evidence* review[4]) -Diagnostic shifts following the 2000 revision of the DSM-IV, which refined the definition of Pervasive Developmental Disorder Not Otherwise Specified to correct an error that allowed this diagnosis to be ascribed when there was impairment in only one developmental area (i.e., social interaction, communication, or stereotyped behaviors, interests, or activities). The definition was clarified to require fundamental core social impairment in addition to either/both communication impairment or the presence of stereotyped behaviors, interests, or activities. -Changes in available assessment methodologies and the introduction of 'gold-standards' of ASD assessment during this same time period; specifically, the commercial release of the Autism Diagnostic Observation System in 1999 and the revised version of the Autism Diagnostic Interview - Revised in 2003 allowed researchers to introduce metrics of sample comparison relative to core characteristics of ASD itself during this time period. As such, focusing the review on this decade of research allowed us to speak to the inclusion of such measurements and, when included, specific behavioral differences relative to core symptoms that may affect thinking about the key elements of interventions and therapies. These points were noted in Chapter 2 of the review; however, we have expanded on that text to clarify our rationale. In addition, the report does not typically include single subject design studies. We understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the

Section	Comment	Response
		study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.
General	The report is well structured and organized. I think that the conclusions could be helpful in informing policy and practice decisions, especially if it is combined with other more thorough reviews such as the National Standards Project and the NRC report.	We have added references to those other reports to the Introduction to assist readers in locating them.
General	this is a very exhaustive review, and will be very beneficial to researchers, and hopefully to clinicians. I believe that most lay people will only read the executive summary, so that section is very important. see my comments above [in executive summary comments].	We agree that the executive summary is particularly important for a lay audience and appreciate your attention to it. We have extensively revised the Executive Summary to ensure that key findings of the review are represented.

Section	Comment	Response
General	We appreciate the opportunity to comment on the Comparative Effectiveness of Therapies for Children with Autism Spectrum Disorders by the Vanderbilt Evidence-Based Practice Center. We recognize the importance of this topic and the need for solid information on the best treatments for children with autism spectrum disorders. We disagree, however, with the conclusion on page 3 of the Executive Summary, that "Despite methodological challenges, the existing research suggests, with moderate strength of evidence, that intensive UCLA/Lovaas*…interventions (greater than 30 hours per week) may confer greater improvements in cognitive performance, language skills, and adaptive behavior skills than the broadly defined eclectic treatments that are often available to children in the community. At present, these approaches have demonstrated effect in RCTs when implemented by both professionals and parents."	We thank the reviewer for her general comment and will provide specific responses to individual items. Of note, additional papers are included in this category in our review than in the previous reviews on the topic. We reviewed our study quality and strength of evidence assessments to ensure uniform application and made corrections in some instances. This category of interventions is considered to have low strength of evidence (SOE) at present.

Section	Comment	Response
General	We also do not agree with the quality rating of "good" for the Sallows and Graupner (2005) paper, as noted on page 38 and elsewhere in Vanderbilt EPC's report. The Blue Cross Blue Shield Association Technology Evaluation Center (TEC) issued a Special Report in February 2009 on Early Intensive Behavioral Intervention Based on Applied Behavior Analysis among Children with Autism Spectrum Disorders (Vol. 23, No. 9, available at http://www.bcbs.com/blueresources/tec/vols/23/23_09.pdf.) It graded the Sallows and Graupner (2005) trial as "fair" quality, because it did not have equal, reliable, and valid measurement. Varied instruments were used for different children to measure outcomes for the same domain. For example, a different instrument might be used pre- and post-treatment because of increased skills at the end of the study, but there was no discussion about how comparable these instruments were in measuring the same domain or change over time. For nonverbal, cognitive functioning, the Leiter-R was used for 11 children and the Merrill-Palmer for 12. CELF-III was used to gauge language in 11 children while the Reynell was used for 12. There are other examples as well. Despite randomization, the percentage nonverbal at the start of the program was 62% in the clinic-directed intervention group and 20% in the parent-directed intervention group. This lack of balance may be due in part to the small size of the trial (n=23).	Our *a priori* developed quality rating scheme did not assess this factor. Based on our scoring system, this paper is considered good quality. The issue of varying measurement strategies is a complex one. While the use of different assessment methodologies certainly influences the nature and complexity of interpreting results, this is not necessarily an inherent weakness in study design. In fact, in order to fully account for developmental and intervention effects, appropriate study designs may in fact necessitate utilizing varying assessments. For instance, the 'gold-standard' of autism assessment, the Autism Diagnostic Observation Schedule (ADOS), utilizes different modules of administration and scoring algorithms to account for such changes and other interventions will routinely utilize varying standardized cognitive assessment for similar purposes. Further, the offered examples are of intellectual ability instruments with normative samples that allow for separate reliable and valid measurement strategies to be compared, although the instruments are not completely equivalent or interchangeable. In addition, the differences in terms of ability level within study are present despite randomization. As the authors' did appropriately randomize based on their internal methodology and report on this difference within a small sample size, this again was not seen as a study design flaw.

Section	Comment	Response
General	After finding no differences in improvement after treatment between treatment groups, the remainder of the paper is devoted to a comparison of rapid learners from both treatment arms versus moderate learners from both arms. As noted in the Vanderbilt EPC report, the interventions in this study were two versions of an ABA approach, one that was clinic-directed and the other parent-directed. The TEC report, which was reviewed by the external Medical Advisory Panel (see http://www.bcbs.com/blueresources/tec/medical-advisory-panel.html) concluded as follows: "Overall, the quality and consistency of results of this body of evidence are weak. Consequently, no conclusions can be drawn from this literature on how well EIBI [based on ABA] works. Weaknesses in research design and analysis, as well as inconsistent results across studies, undermine confidence in the reported results. It is important to distinguish between certainty about ineffectiveness and uncertainty about effectiveness. Based on the weakness of the available evidence, we are uncertain about the effectiveness of EIBI for ASDs. The cost of continuing the current course of assuming that EIBI works may not be obvious. EIBI is costly financially for society and requires a large time commitment from children, their families, and their teachers or therapists. However, these programs may not appear to pose any harm for the children themselves. Nevertheless, the opportunity costs could be high, indeed, of providing suboptimal care to these children, simply because we as a society do not know what works best. The children may be treated with an intervention that is not as	We appreciate the findings of the rigorously-conducted TEC review and these comments which prompted us to re-examine our placement of the Sallows and Graupner study as addressing key question 1 in our report. We realize that the paper more accurately addresses modifiers of treatment effects and have revised our discussion of the paper as needed as well as revised our SOE ratings for the University of California, Los Angeles (UCLA)/Lovaas approach to "low." We note that additional studies, including those assessing the Early Start Denver Model (ESDM), have been published since the TEC review, and while the strength of evidence for ESDM studies is insufficient at this time, we feel that the addition of these studies lends support to the effects of early intensive intervention approaches. Nonetheless, we stress, and have expanded in the text, that the low strength of evidence is in support of an effect seen in subgroups of children that have yet to be well characterized. We also stress other limits of the available data (i.e., lack of community effectiveness, concerns of moderate response in subgroups). As such, the practicality of implementing these approaches in a population is unknown.

Section	Comment	Response
	effective as the alternatives. And if we accept an intervention because it seems to work, without solid evidence, research on the alternatives or on how it can be improved is likely to be stifled."	
General	children with autism don't show any improvement even they go the special class in public school. para educators hours are always cut due to budget cut, and most kids need one on one.	We hope that our systematic review identifies areas of knowledge and areas for further research to improve the current state of ASD treatment.
General	See comments above [in executive summary comments]. ABA is not an "umbrella term for principles and techniques used in the assessment, treatment, and prevention of challenging behaviors." Genuine, competently delivered early intensive ABA intervention for autism does not consist entirely or mostly of discrete-trial procedures used in highly structured settings, with a little bit of incidental teaching thrown in. Rather, it comprises a wide array of ABA procedures and is delivered in a variety of settings. Nor does early intensive ABA ignore information about typical development, as this report implies. In fact, ABA was originally developed by blending the experimental analysis of behavior and aspects of developmental psychology. Well-trained behavior analysts working in autism have long incorporated developmental "principles" and information from research on human development in their treatment.	The description of Applied Behavior Analysis (ABA) here is used to provide a general description of the numerous approaches that may fall under the term. We also attempt to provide the readers with information about the history of the term ABA, which was coined and utilized decades prior to application to autism intervention. We agree that 'ABA' as practiced in the community is much more eclectic than the brief summary that we are able to provide here. Detailing the variety of approaches used is beyond the scope of this systematic review, which focuses on outcomes of research, rather than the protocols for delivering treatment; however we have attempted to clarify our description of ABA approaches to emphasize that a variety of procedures may be employed.
Executive Summary	estimated average prevalence of one in 110	We have deleted the word "average."
Executive Summary	we drafted the initial key question**s**	We have added an "s" to "question."
Executive Summary	Specify who the reviewers were	The investigators on the team served as reviewers. Peer reviewers are identified in the front matter of the report.

Section	Comment	Response
Executive Summary	Clarify in more detail how quality was assessed for each of the dimensions listed, or let the reader know that this is described in more detail later in the report (page 40).	We have added a sentence referring the reader to the full report for additional detail.
Executive Summary	abstracted data on harms. This seems to come out of nowhere. Clarify what is meant by "harms" and perhaps mention earlier in the review that this was done	The use of the word harms is intended to encompass both adverse events of treatment and broader harms potentially associated with treatment, such as social harms, reduced quality of life, etc. We have added the EPC definition of harms to the report's methods discussion.

Section	Comment	Response
Executive Summary	Not all children receiving early intensive intervention, however, demonstrate rapid gains, with some available data suggesting that many children will continue to display prominent areas of impairment and that subgroups may account for a majority of the changes observed. To date, there have been no published RCTs that have adequately examined the impact of moderator variables based on child characteristics, such as IQ, on response to EIBI. Instead, what has been reported is wide variability in outcome within the treated group. We recently reported at IMFAR (2010) that both low and high IQ children responded significantly to early intervention based on the Early Start Denver Model. The lower IQ children start lower and end up lower, but they nevertheless make significant gains. Thus, I think it would be more accurate to state the following: "Not all children receiving early intensive intervention, however, demonstrate rapid gains, with some available data suggesting that many children will continue to display prominent areas of impairment. There is a paucity of published data on the impact of child characteristics on outcome that have been collected in the context of RCTs."	Thank you for your comment. We agree, and have made this change in the Executive Summary and Discussion sections of the report. We concur that the International Meeting for Autism Research data replicate and add support to the idea that proportions/subgroups of children with autism do not seem to demonstrate robust changes in response to early and intensive intervention (in regards to IQ, Adaptive behavior, and core ASD symptoms). We unequivocally agree that that even small change can be powerful and meaningful; however, it is also important to recognize that a sizable portion of children will not see the same robust change in response to same intervention. Per the reference to the conference presentation, this review does not include conference proceedings that were not published at the time of the review literature search. However, we do appreciate the information.

Section	Comment	Response
Executive Summary	I believe that recent studies suggest that the evidence on the efficacy of social skills training is encouraging. As one example, cited in this review, the RCT of parent-assisted children's friendship training with children with ASD (JADD, 2010, 40: 827-42) showed a positive impact on more than 87% of children. This included children in the older age range of the range examined by this review, which suggests that there is evidence of efficacy for social skills training for older children as well. The friendship training study is described later in the report (page 58) and the positive effects of friendship training post treatment and at follow up are described. I am a bit surprised that the quality of this study was only rated as fair in Table 12, but nevertheless, it is still encouraging. A recent review of the evidence for efficacy of social skills intervention was provided by Reichow and Volkmar (Social skills intervention for individuals with autism: Evaluation for evidence-based practices with a best evidence synthesis framework. JADD, 2010, 40: 149-66), and these authors also concluded that there is moderate empirical evidence for the efficacy of social skills training	We think that there is some confusion on the use of the strength of evidence measures in light of measures of effect and have tried to clarify this. Although individual studies do provide promising, positive results, the strength of that evidence (i.e., our confidence that there exist currently enough good studies so that future research would not alter the estimate of effect) is low because there are too few studies that are consistent in both the intervention and the outcome measures. We have tried to clarify the text so as not to conflate these concepts.
Executive Summary	Line 45 – a recent large RCT showed no benefit of fluoxetine for improving repetitive behavior in ASD. Perhaps the sentence should read: "Other medical interventions also been evaluated in single *published* RCTs and show…..	We have made the suggested change.

Section	Comment	Response
Executive Summary	Our published data from our RCT on ESDM does suggest that changes in IQ account for longer term change in IQ, but we have no published data that predict outcomes. So, I am not sure this should be included in this section that is specifically targeted at the question of what predicts outcomes. The question of when the largest gains are made (early versus late) is a different one than what predicts outcome. These comments also pertain to the section on page 140 on Early Results in the Treatment Phase That Predict Outcomes.	We have reframed our discussion to better reflect that such data are reflective of predictions of change in intelligence quotient (IQ) rather than follow-up outcome. We feel that with the paucity of available data in this regard (outcome as defined as such) it is necessary to discuss factors that are available despite limits, in part to provide the state of what is known, but also to clearly point out where such data are lacking.
Executive Summary	Regarding RCTs for treatments for children below 2, it should be noted that many of the children who participated in the RCT of ESDM were below two years of age (range was 18-30 months)	This study is included in the section on therapies for very young children (Key Question (KQ) 7). We have made that more clear in the report.
Executive Summary	I believe the term “sparse” is a little too harsh. Sparse means meager and I believe we are past that stage (not much, admittedly, but still!). I believe it would be more accurate to say that, “In sum, research to support therapies for children with autism ranges from sparse to more adequate, with a ranging strength of evidence; however, this field is developing.”	We concur and have modified this text.

Section	Comment	Response
Executive Summary	I am concerned that the tone and content of the abstract and executive summary may lead a reader to believe that there is stronger evidence in support of early and intensive behavioral interventions than what the Evidence Report would imply. A majority of clinicians and families will only read the Abstract and Executive Summary, so it is very important to make these reflective of the actual findings in the Evidence Report. There seems to be a significant disconnect between these documents and the Evidence Report. For example, the Abstract Results state "There is evidence to support early and intensive behavioral interventions, including the University of California, Los Angeles (UCLA)/Lovaas and Early Start Denver Model for improving cognitive performance, language skills, and adaptive behavior skills." There are very strongly held misconceptions by family members and even many clinicians about the strength of the evidence for ABA, as well as the degree of expected benefits. As discussed in the review, the only good or fair studies found effects far less substantial than what Lovaas originally described and had very significant limitations. Broad, unqualified statements such as those in the abstract only serve to propagate misconceptions. The Evidence Report does a nice job discussing the findings and limitations of other types of therapies. However, the language in the abstract makes it seem like there is evidence to support the ABA methods and only "some suggestion" that anything else may be useful.	We have taken care to note limits of available evidence regarding these interventions within the executive summary. We note also that UCLA/Lovaas-based approaches alone have low SOE while the SOE for ESDM studies is insufficient at this time. We note also that at this time, researchers have been unable to clearly identify which children are likely to benefit from these interventions, which likely limits their broad utility. It is also important to note, and we have expanded our text on this, that strength of evidence is not the same as degree of effect. Strength of evidence is a measure of our confidence in the results currently observed. It is not an endorsement of the "amount" of change likely to be observed.

Section	Comment	Response
Executive Summary	The summarizations in the Executive Summary are quite confusing. For example, in the section on behavioral interventions, it says that there is moderate strength of evidence that intensive UCLA/Lovaas and Early Start Denver Model interventions confer greater improvements in <a number of outcomes> as compared to eclectic treatments. I take it the authors are using data from the 7 prospective studies or non-randomized trials to make that statement, not data from RCTs. However, the next sentence reads "At present, these approaches have demonstrated effect in RCTs when implemented by both professionals and parents." The Evidence Report only includes 2 RCTs on UCLA/Lovaas treatment, both of which compared professional-delivered ABA-based treatment to parent-delivered ABA-based treatment. As the authors explain in the Evidence Report, one of the RCTs had much more tempered results than the original Lovaas research and the other RCT found no difference between the intervention (expert-delivered ABA) and control arms (parent-delivered ABA). Thus, it seems misleading to say that we have evidence of effect from RCTs. It is true that both the professional-delivered and parent-delivered arms of the second RCT showed improvement, but since there was no difference between the two arms of this study, it would seem that the evidence for improvement can at best be seen as coming from a non-randomized, pre-post comparison or case series, not an RCT.	The reviewer correctly points out that the Sallows study compares professional implementation of an approach with parent implementation of that approach, and therefore this study has been moved to the moderator section. We have thus revised our SOE statement for the effects of UCLA/Lovaas-based approaches to low. We also note that strength of evidence is not a measure of effectiveness, and that we note that there are greater effects seen with ABA approaches compared to eclectic ones, but that these effects are difficult to assess across studies and are observed in a subset of children.

Section	Comment	Response
Executive Summary	Like with the Abstract, the Executive Summary seems to dismiss many of the studies for any behavioral intervention other than intensive EIBI/ABA therapy. Though those studies have many limitations, it is unclear why the very limited body of literature for ABA is summarized as "moderate strength of evidence", while positive findings from any other studies are barely described. I am not suggesting that the evidence is strong for other therapies – there just seems to be a discrepancy between how the evidence is discussed for ABA-based vs. non-ABA-based studies.	We explicitly note the available evidence for 1) parent training for bolstering social communication skills and managing challenging behaviors 2) social skills interventions 3) play and interaction-based intervention and 4) cognitive behavioral therapy (CBT). As we note in the evidence report, the sheer number of studies conducted on early intensive behavioral and developmental approaches means that a majority of the evidence regards such studies. Attention and availability should not be confounded with an implicit endorsement of any sort. We have, however, re-examined our application of SOE criteria to ensure consistent application and revised our rating for UCLA/Lovaas studies as noted above. The SOE for the majority of behavioral interventions included in the report is insufficient at this time due to factors including variability in the interventions assessed and the limited number of studies.
Executive Summary	Overall, there is not enough mention of gaps and limitations of the literature in the Executive Summary. There is also not enough mention of the findings as they relate to the sub-questions listed under each key question (see comments under Results for more details.)	We attempt to acknowledge powerful limitations succinctly and have revised text in the Executive Summary to clarify that significant limitations exist. We have also added text to clarify that we have addressed sub-questions within discussion of the appropriate overarching key question
Executive Summary	There also is not enough information in the Executive Summary about KQ7 – please provide some information about what makes the four studies mentioned promising or limited.	We have revised our text discussing these studies to provide more detail.
Executive Summary	Comprehensive and well done. Opens the door for future comparative, well documented studies. Note: Strong family reports of improved status with Hyperbaric oxygen. Needs study without the conflict of interest component.	Thank you.. We agree that a number of potential interventions need further study, and that potential conflicts may exist in current research. This concept is addressed in the future research section of the report.

Section	Comment	Response
Executive Summary	What are some of the established/successful ways of including parents and siblings in therapeutic efforts?(I have been there during my 41 years of active pediatric practice)	This is an important question. The report does detail some studies that addressed parent delivery of therapy; although there is not enough data to truly draw specific conclusions.
Executive Summary	Because the report is so long, many people will only read the ES. Why does the EX only address results for ASD children under age 2?	The executive summary actually focuses largely on children older than age 2. We have revised the text in the Executive Summary and elsewhere in the report to be sure that the age range included is clear
Executive Summary	Be very careful in the summary when discussing "potential" or "preliminary" evidence that an intervention works, based on non-randomized trials. very high level of placebo response in this population. always address potential harms with potential benefits when talking about unproven interventions.	We agree that conclusions should be based upon the best evidence and have revised text throughout to be especially clear about study design and potential for bias. This is also captured in our quality assessment, which is reported for each included study. Nonetheless, we do believe, based on rigorous assessment, that there are a number of studies that do show preliminary evidence for effect, but that warrant further study with strong study designs.
Executive Summary	General comment for review - recommend having specificity around the indication under review (autistic disorder as opposed to ASD).	We understand the argument that the review should focus solely on 'autistic disorder,' but the data do not offer a clear distinction between sub-groups within the autism spectrum disorders, as reflected in the decision of the Diagnostic and Statistical Manual of Mental Diseases-Fifth edition (DSM-V) subcommittee to re-construct this category as 'autism spectrum disorders' without subdivisions for 'autistic disorder,' 'Asperger's,' or 'Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS)'.
Executive Summary	Page iii Statement "Risperidone and aripiprazole demonstrate improvement in challenging behavior that includes emotional distress, aggression, hyperactivity and self-injury, but have high incidence of harms." Comment – recommend that this statement be consistent with the FDA-approved indications risperidone and aripiprazole have for irritability associated with autistic disorder (tantrums, quickly changing moods, self-injurious behavior, and aggression). Recommend this for similar statements in document.	We considered the use of the term 'irritability' quite vague and confusing in comparison to the actual items measured in the Aberrant Behavior Checklist Irritability Subscale, which we have detailed in the report. The concept of 'Irritability' is quite broad in comparison to the specific items on this subscale. Our task is not to review the Food and Drug Administration (FDA) indication but to instead evaluate the evidence for change in specific domains in response to treatments.

Section	Comment	Response
Executive Summary	Page 4 Statement “Both medications also cause significant side effects, however, including marked weight gain, sedation, and risk of extrapyramidal symptoms.” Comment - It is perhaps reasonable to also note hyperprolactinemia in the case of risperidone	We do describe hyperprolactinemia as a risk elsewhere in the report, but word limits on the executive summary prohibit mention of all side effects.
Executive Summary	ABA therapy is very effective & helpful to autistic children.	Thank you for your comment.
Executive Summary	Most insurance company don't cover treatment for autism.	Thank you for your comment.
Executive Summary	Thanks for an excellent report, well-crafted, good questions, thoughtfully addressed. The evidence reviews and analyses seem well done and provide good support for the conclusions in each section. The report does helpful work by noting the variations in the population of children with ASD and how poorly characterized those variations are - and that they have generally not been considered in studies of treatment efficacy. My main concern is with the very summary statements both in the executive summary and in the structured abstract. Importantly, the review does identify some behavioral interventions as effective (recognizing that much more evidence could address for whom and at what intensity, among other questions). Rather than making the conclusion in the abstract begin with the sparseness of data (sadly the opening line from almost all evidence reports - and hardly a new finding), suggest leading with the finding of effectiveness of certain interventions - ie, the second sentence, which more reflects what may be new in this report. (Isn't the evidence sparse for most mental health conditions - e.g, ADHD, depression, other than some data on meds [and of course even there for depression in adolescents the data are pretty sparse.])	Thank you for your comments. We agree and have revised the summary statements to note evidence supporting the effectiveness of some treatments.

Section	Comment	Response
Executive Summary	Similarly, the conclusion of the executive summary leads with a weak and non-novel finding - and does not state that the review finds good evidence for some behavioral/educational interventions. Similarly, could say that there is quite limited or poor evidence of effects of medications in autism.	We have revised these statements.
Executive Summary	ABA Therapy was needed for the kids who have autism, please vote for them	Thank you for your comment.
Executive Summary	The reviewers are to be commended for undertaking this important task. The review and summary of evidence on medical and CAM interventions are generally very thorough and well done, but here and elsewhere the authors provide inaccurate information about applied behavior analysis (ABA) and early intensive ABA intervention for autism. They have also made the common but egregious mistake of excluding evidence from studies using the single-case experimental designs that are the hallmark of behavior analysis. Those types of studies are not uncontrolled "observational studies," "case studies," or "case series." Rather, when done properly, they are highly rigorous controlled experiments that produce rich, direct, objective evidence about the effects of interventions of various types on the behavior of individuals over time. Indeed, they are the *only* types of experiments that can yield that type of evidence.	We agree that single-case experimental design is important, but it did not fit into the scope of this review. We understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.

Section	Comment	Response
Executive Summary	(As an aside, some leaders of the evidence-based practice movement in medicine agree, and recommend one type of single-case experimental design as the strongest design for evaluating many medical treatments). Large-n group design studies in which mathematical abstractions (group mean scores on some limited measures) are compared statistically cannot reveal effects of interventions on individuals over the course of a study. RCTs in particular have rather serious limitations in that regard as well as generality to real-world situations, not to mention their practical and ethical limitations for evaluating any interventions for children with autism, especially comprehensive, intensive, long-term interventions. A number of autism researchers are beginning to recognize those limitations, and to appreciate that there is more than one way to do good science, and more than one type of research design that can produce credible evidence about treatment efficacy and effectiveness. Since autism is defined and diagnosed behaviorally, and affects individuals differently, single-case experimental designs are well-suited for evaluating the effects of all types of interventions for this population. By using an overly narrow definition of science and excluding a very substantial body of scientific evidence of the effects of a wide array of ABA techniques for building skills and reducing problem behaviors in people with autism of all ages, the reviewers have presented a very incomplete and skewed picture of the effectiveness of autism treatments, particularly the ABA approach.	Thank you for your comments. As noted above, our review is focused on studies that met our entry criteria, including a minimum of 10 subjects. We understand that this does not capture a large number of smaller studies, but these studies were not excluded based upon study design but based upon sample size. We appreciate that different study designs are useful across the phases of treatment development and evaluation.

Section	Comment	Response
Introduction	– I would suggest the following re-wording: ASD involves multiple etiologies involving both genetic and environmental risk factors. Among the environmental risk factors that may contribute to ASD risk are advanced parental age, prenatal maternal infection, and prematurity.	We have expanded our text in this section to note that multiple factors may play into the development of ASDs.
Introduction	Suggested re-wording: … continuum of approaches from highly structured approaches based on an ABA technique referred to as Discrete Trial Teaching to natural/developmental approaches that deliver intervention within natural contexts (Floortime, SCERTS model), some of which integrate both developmental and ABA approaches (Early Start Denver Model).	We have revised this statement to note the development of both highly structured approaches based on and natural/developmental approaches that deliver intervention within natural contexts (Floortime, the Social Communication Emotional Regulation Transactional Support model), some of which integrate approaches (ESDM).
Introduction	Suggested re-wording: strictly defined UCLA/Lovaas method, which relies heavily on one-on-one therapy sessions during which a trained therapist uses discrete trial teaching with a child to practice target skills…	We have changed the statement as noted.
Introduction	– I am surprised you don't explicitly mention the casein-gluten free dietary interventions here.	We focused the introductory section of the report on those interventions discussed in later chapters. Studies assessing the gluten free-casein free diet did not meet our inclusion criteria and are not addressed in the report. We have, however, included references to other recent reviews so that readers can learn more about interventions not included here.
Introduction	Although the difficulties in categorizing interventions are appropriately acknowledged, the classification scheme in the report does not appear optimal. For example, it is unclear why there are separate categories for behavioral and educational interventions. As the review indicates, the interventions that are identified as behavioral are routinely implemented in schools, and interventions that are identified as educational are routinely implemented in the home. Moreover, there is almost complete overlap in intervention	As you note, and other reviews have noted, it is challenging to categorize interventions for ASD, and no one categorization approach is uniformly accepted. While we understand your concerns, we do not feel that any other approach would be acceptable to all experts in the field, and so we will maintain our current organization and be as clear as possible about the choices we made.

Section	Comment	Response
	goals, as well as considerable overlap in intervention techniques.	
Introduction	– Another difficulty arises from categories such as "joint attention interventions." Interventions for joint attention range from highly structured formats such as discrete trial training to child-led, play-based approaches.	As noted previously, an inherent challenge of this review was attempting to develop appropriate categories for grouping bodies of literature that may have varied based on permutations of setting, goal, approach, and participant characteristics, amongst other factors. The focus of this report was to be able to provide information that could be potentially utilized in consumer choice. As such, we created categories often incorporating setting as the primary grouping construct. In addition to this grouping, it was necessary to identify other relevant groupings that would drive intervention (i.e., goal or approach). Thus, the category of joint attention interventions does not assume that children receiving UCLA/Lovaas or ESDM intervention might not show improvements in these areas, but this specific category is for interventions focused on those skills as a target outside of additional treatment settings. We describe our approach to categorizing interventions more fully in the Methods chapter of the report.
Introduction	– My suggestion is to have two overarching categories: (1) broad-band or comprehensive interventions defined on the basis of approach, including ABA (EIBI and PRT), TEACCH, and developmental approaches such as the Denver Model and (2) focal interventions defined on the basis of goals (parent training, social skills, interventions for joint attention, computer-based instruction, etc., noting, where applicable, that a variety of different approaches have been attempted).	To a large extent this represents our organizational scheme (i.e., grouping of broad based approaches and focal approaches) however, as stated previously we also felt it important to attempt capture elements of setting where possible and have carved out educational approaches as appropriate. Again, this approach was informed by the Rogers and Vismara review of comprehensive interventions. Including a review of the parent training opportunities (PRT and other approaches) was deemed appropriate based on the fact that multiple targets were often identified. We describe our approach to categorizing interventions more fully in the Methods and Discussion chapters of the report. Of note, we assessed whether a different organizational scheme would have changed our assessment of the literature and determined that it would not have done so.

Section	Comment	Response
Introduction	– Although the definitions of most interventions are accurate, the explanation of ABA is not quite right. The report states, “ABA is an umbrella term used for principles and techniques for principles and techniques used in the assessment, treatment and prevention of challenging behaviors” (p. 21). In theory and practice, however, most ABA interventions do not focus on challenging behavior. ABA is usually described more generally as an approach to addressing socially important problems.	We agree that this wording could be improved. Our intent with the following sentence: "The goal of ABA is to teach new skills, promote generalization of these skills, and reduce challenging behaviors with systematic reinforcement." was to hit on this same point, but in order to clarify we have revised the text as follows: "ABA is an umbrella term used for principles and techniques used in the assessment, treatment and prevention of challenging behaviors as well as the promotion of new desired behaviors." Further, as the core symptoms of ASDs are often constructed as negative symptoms or absences of certain behaviors, it is important to recognize that this absence itself can be construed as a ‘challenging behavior.’

Section	Comment	Response
Introduction	The introduction was well written and accurately reflected priorities in the field. The cut-offs for sample sizes were not specifically justified; this might be helpful, given that other reviews have used different cut-offs.	We have added rationale about our decisions regarding sample size to the Methods section of the review and readily acknowledge that 1) There is no hard and fast standard for size selection 2) Other reports have included different cutoffs. The minimum number (N) of ten participants was selected in consultation with our content experts and Technical Expert Panel as a minimum threshold for comparing interventions. We felt that given the greater risk associated with the use of medical interventions, it was appropriate to require a greater sample size to accrue adequate data of safety and tolerability, in addition to efficacy. We restricted the review to medical studies with at least 30 participants given that most studies of medical interventions for ASD with fewer than 30 subjects report preliminary results that are replaced by later, larger studies. We feel that this restriction did not eliminate specific medical therapies from the review as treatments are typically assessed in larger studies following their preliminary investigation. Moreover these sample size constraints are not uncommon in the systematic review/comparative effectiveness review literature.
Introduction	An example of when the criterion "further evidence is unlikely to change results" would be helpful.	We felt that this criterion was applicable in cases in which more than one “good” study was available and that estimates for the domains comprising strength of evidence scores were in a positive direction (i.e. precise, consistent, etc.). We do discuss evidence in some areas where further evidence is unlikely to change results, most prominently including secretin.

Section	Comment	Response
Introduction	The classification system used to describe and categorize the behavioral interventions appears to be inconsistent and confusing. For example, in the text, the authors talk about a category of "Early intensive behavioral and developmental interventions", while the table refers to this group as "ABA-based approaches." Some of the interventions included in this section are based on ABA, but some have been developed in response to criticisms of ABA and are not generally thought of by clinicians or families as "ABA-based" (especially the therapies grouped under "developmental and relational approaches"). Calling all of these therapies ABA-based, even if in some sense they do remotely relate to ABA, may be confusing to clinicians and families.	We have revised the table categories to be more clear. They now include the following: Approaches aimed at core symptoms and Approaches aimed at commonly associated symptoms. We have revised our grouping of "ABA-based" approaches under the more encompassing heading of "early intensive behavioral and developmental approaches," which we note may incorporate principles of ABA.
Introduction	Also, if I am understanding correctly, some of the therapies listed in Table 1 under "ABA-based approaches" (such as Floortime and RDI), are discussed in the results section under "Play-/Interaction-based". Similarly, it is unclear why there is a separate category called "parent training" in Table 1, when many of the therapies listed under "ABA-based" or "Early intensive behavioral and developmental interventions" also involve parent training.	This table was meant to describe broadly the types of behavioral interventions in the report. We have revised the table to reflect the actual organization/presentation of behavioral approaches in the review.
Introduction	Also, the table calls these interventions "Parent training", but there is no equivalent label in the text.	This table was meant to describe broadly the types of behavioral interventions in the report. We have revised the table to reflect the actual organization/presentation of behavioral approaches in the review.
Introduction	The text has a category for "Play-/Interaction-based interventions", which is not in Table 1, but which is used in the Results.	This table was meant to describe broadly the types of behavioral interventions in the report. We have revised the table to reflect the actual organization/presentation of behavioral approaches in the review.

Section	Comment	Response
Introduction	I am really confused as to where the Play-/Interaction-based therapies fit into the authors' scheme. Many of the interventions listed in Table 1 as ABA-based, Parent Training, or Social Skills also are play or interaction-based. I am not sure what the best labels would be, since I am not totally sure that I understand the authors' reasoning behind their groupings. If the authors do intend to group traditional ABA interventions with developmental/relationship interventions, then maybe a reasonable categorization would be "high intensity, comprehensive behavioral and developmental interventions", "moderate or low intensity, comprehensive behavioral and developmental interventions", "targeted interventions aimed at social skills", and "targeted interventions aimed at associated symptoms". However, that does not seem to match the groupings in the Results Section, where the authors seem to put early intensive developmental interventions that are not ABA-based as play-/interaction-based therapies.	This table was meant to describe broadly the types of behavioral interventions in the report. We have revised the table to reflect the actual organization/presentation of behavioral approaches in the review. As you note, and other reviews have noted, it is challenging to categorize interventions for ASD, and no one categorization approach is uniformly accepted. While we understand your concerns, we feel that our organization of behavioral interventions is justified and have clarified statements regarding our decisions about categorizing treatment approaches.
Introduction	Some of the text describing the interventions is also a bit confusing. For example, on page 10, 4th paragraph, it states that "Interventions on the other end of the continuum include the Early Start Denver Model, which blends ABA principles with developmental and relationship-based approaches." What continuum are they referring to – is it the continuum with behavioral	The use of the word continuum was confusing, and we have removed that word and concept from the text.

Section	Comment	Response
	interventions on one end and relational/developmental interventions on the other? If so, why would the Denver model be at the other end of the continuum? Wouldn't the strict developmental and relationship-based approaches be at the other end of the continuum and the Denver Model in the middle?	
Introduction	Under Social Skills interventions (bottom of page 10), the authors state "Decreased interest in the social environment has been reported since the earliest descriptions of the disorder and is the unique and essential aspect of ASD that distinguishes it from other childhood disorders." I strongly disagree that the unique and essential aspect of ASD is decreased interest in the social environment. Perhaps the authors could refer to decreased social skills or challenges with social interactions, not decreased interest.	We have changed the text to: "difficulties with social engagement have been reported…"
Introduction	Under Educational Interventions, (bottom of page 13), the section starts by saying "Education through schools and other community settings (e.g. centers) is the primary treatment for children with ASD." I see no reason to get into the debate as to whether the primary treatment should be through the school system or the healthcare system. I doubt the authors intended to enter that debate. It may be best to remove that statement and insert a less controversial one.	Currently, most children with ASD do receive care in some capacity through the educational system. We are not implying nor debating whether this is appropriate.
Introduction	The analytic framework model is a little confusing. For example, KQ3, 4, and 5, as written in the text, do not seem to completely match what is in the figure.	We have simplified the analytic framework based on this and similar comments.
Introduction	Good overview of ASD. Strong case made for need for further research and evidence-based interventions. I wouldn't make substantial changes. I identified no major discrepancies.	Thank you.

Section	Comment	Response
Introduction	excellent	Thank you for your comment
Introduction	the older they get more behavior shows up	Thank you for your comment
Introduction	See previous comments [general & exec. summary comments]. This chapter makes it clear that the reviewers have ignored most of the evidence on the efficacy of many focused ABA interventions for addressing not only the core symptoms of autism, but also the problem behaviors that are often exhibited by people with autism, such as self-injury, aggression, stereotypic and repetitive behaviors, elopement, pica, eating difficulties, and sleep difficulties. This is a glaring and serious omission, especially since those problem behaviors directly affect the health and safety of people with autism. The report implies that providing parents with a little training in behavior management techniques is the only approach to treating problem behaviors that is supported by research, and is sufficient for treating them effectively. That position is not defensible empirically when the full range of scientific evidence is considered, and it is dangerous.	We appreciate the comment and again would stress the usefulness, but also the limitations, of single-case experimental designs, which are frequently used to assess the effects of ABA interventions on problem behaviors. We note that we did not specifically exclude single subject designs but required that studies included in the review include at least 10 participants (for non-medical studies) as our team and expert consultants felt that this was the minimum number of participants needed to allow conclusions on the effectiveness of therapies.
Methods	The design logic used to address the research questions was sometimes dubious. [this is the first in a series of 3 comments broken out in the following rows]: 1. The rationale for question KQ2 is "identifying subgroups of children for whom treatments are more or less effective, as well as factors that increase or decrease effectiveness may help understand which treatments are applicable under what circumstances". The design logic used to	We agree that there are significant challenges in interpreting the existing studies that attempt to address key question 2. Specifically, most studies are limited to correlational analyses that predict which subjects improve over the course of the treatment period but do not explicitly test whether that improvement is due to the treatment or may instead reflect other factors. Even beyond the challenge of having too few RCTs in the ASD literature, the existing RCTs are typically under-powered to assess this type of interaction. We worked throughout the section to avoid making statements that

Section	Comment	Response
	infer characteristics of family and child that were associated with treatment response was usually incorrect. The authors used correlates of change or post-treatment outcome for participants in a single group (even treatment group). This is not the same as correlates of change DUE to a treatment, even if the treatment has been shown to be generally effective. Using correlates of change (or post-treatment level) within the treatment group as a way to "identify subgroups of children for whom treatments are more or less effective" is implicitly assuming that all change is due to the treatment. Even when the mean for the treatment group is significantly different from the mean of the control group on an outcome or change score, one cannot assume that ALL change is due to the treatment. Even when group means differ by 3.0 SDs that is only 50% of the variance in the outcome! Essentially, when one looks at pretreatment correlates of gain within a single group, regardless of whether it is a treatment group, one is using only a correlational design. In terms of scientific logic, it is as if the participants did not receive a treatment because the association of characteristic with outcome or characteristic with gain may have nothing to do the treatment.	pre-treatment characteristics moderated outcomes due to the treatment, where the data didn't support such an assertion. Stating that pre-treatment features correlated with treatment outcome is precisely true. We have included additional text to reinforce the concept that correlation does not imply causation.

Section	Comment	Response
Methods	Two other research designs CAN address this question. In a single study, the most appropriate design is a randomized control trial in which the size of between-group differences on the outcome varies as a function of the "characteristic" or "moderator". Statistically, this is a pre-treatment by treatment group interaction predicting outcome. In a meta-analysis, the "moderator" is a correlate of between-group difference effect sizes that are derived from multiple studies. In both designs, the between-group difference is essential to the logic of identifying "identify subgroups of children for whom treatments are more or less effective".	With regard to meta-analysis as a means of answering KQ2, we respectfully disagree in the context of the existing ASD studies, which are too heterogeneous to allow clear identification of moderators across studies. Even in the presence of more homogeneous treatment outcome measures, treatment delivery, and other study characteristics, meta-analysis may be able to suggest treatment modifiers for further study but would be very unlikely to be definitive.
Methods	Note that MOST of the findings the authors use to "answer" KQ2 do NOT fit these two designs. Therefore, many of the statements used to "answer" KQ2 do not hold water. Examples of using this logic are many in this section. For example, p. 98, lines 21 – 24, 48 – 54, p. 99, lines 3 – 6. The same logic error is on page 99, lines 20 – 31. If authors say they are not interpreting all change in the treatment group as due to the treatment, then how can it be justified to say statements such as those on page 99, line 7-8, "Baseline language/communication skills may also correlate with treatment success, with studies generally suggesting a benefit for communication skills…". Additionally, the fact that one of the few studies that used the appropriate design logic (169 – 170, p 99, lines 11 - 14) was included in a paragraph that denigrates this evidence to mere association with growth indicates the lack of this important distinction.	We have substantially revised KQ2 to more clearly highlight the near-total absence of studies that are designed to directly assess treatment modifiers. We have stressed the correlational nature of most data, which does not allow causal inference in this area. We have specifically highlighted the few studies that do allow direct assessment of treatment modifiers to separate them from the overall assessment that little direct data exist to answer KQ2. We do still view it as important to describes some of the correlational data as potentially identifying areas for further study. Our hope in presenting this correlational data is that it will help elucidate and highlight the weakness the reviewer notes in the available data.

Section	Comment	Response
Methods	Finally, the review missed a moderator that was detected using the recommended research design: object exploration or object interest.	We have expanded our discussion of this study under KQ2.
Methods	The design logic used to address the research questions were sometimes dubious. 2. Indicating a significant change in an outcome for one (or more) groups but not in another group doesn't mean significantly greater change in the former group(s) than the latter group. The notion of significance just means the confidence interval (CI) around the mean gain score didn't include zero. But it does not mean that the CI around the mean gain score for one group does not overlap with the CI for the mean gain score another group. The latter logic is the one we use to infer a treatment effect, not the former. If the authors say they know this, then it would more clear to the readers if the authors would write about between-group differences in growth or outcome, not mere significance of change in one group but not in the other, when intending to point out changes due to the treatment. This error is found in many places of the review (e.g., entries in Table 12).	We have tried throughout the report to identify the change on which significance is measured and to emphasize between group differences rather than pre-post differences where possible. We do not intend to imply that within group change is a measure of treatment effect. However, as a systematic review, our role is to report what is available in the literature and we do not typically calculate additional statistics outside of a formal meta-analysis. Therefore, we have presented what data are available, have included statistics where they exist and have tried to be clear when they refer to between group measures and when they do not.

Section	Comment	Response
Methods	The design logic used to address the research questions was sometimes dubious. 3. When reporting group-based treatments (p. 60-educational interventions), there is no attention to whether the appropriate analysis unit was used. That is, number of groups, not number of participants, is the appropriate basis for the degrees of freedom in the significance tests. When participants are used, instead of groups, elevated probability of type I error can occur, some would say is likely.	Our assessment of statistical analyses was intended to differentiate studies that were good vs. fair, and the issue of appropriate analysis unit would not have changed this scoring. This specific issue arises when studies randomize entire groups (e.g., a classroom) or use a group cohort design, rather than randomizing individuals. The evidence from studies using such a design was judged to be insufficient, and adding in this additional significance criteria would not have changed that evaluation.
Methods	Some of the criteria for assessing methodological quality are questionable:[this is the first in a series of 3 comments broken out in the following rows.] 1. Adequate description of concomitant behavioral interventions is almost impossible. Pro forma attempts to do so don't add much. Equating this with treatment fidelity monitoring, which is important and practicable, doesn't seem reasonable.	Full, adequate, and appropriate methodologies for indexing concomitant interventions in this literature are lacking. However, while challenging, there are potent ways in which failure to describe aspects of certain interventions might represent a threat to the internal validity of a study. The psychopharmacological literature is the easiest to point to where failure to account for an additional medication would certainly cloud interpretation of findings. As such, we feel that it is important for researchers to attempt to do so. We agree that it is difficult and that standards do not currently exist to do so uniformly, but we stand by this measure as important in the quality of ASD research. As you indicate, however, further research is needed on how to best achieve these measures in practice.
Methods	Some of the criteria for assessing methodological quality are questionable: 2. Because there were 8 elements under statistical analysis and fewer under other aspects of the quality rating, statistical analysis was probably weighted more strongly than other categories of evaluation. And yet the perceived importance of the sum of the statistical analysis section does not outweigh	The sub-factors under statistical analysis were not given any weight; rather they were guides for the assessment of the one statistical question, "was the statistical analysis appropriate?" We have clarified this in the methods.

Section	Comment	Response
	some other considerations that have more effect on internal validity of the study (e.g., design elements such as randomization and blindness of staff involved in collecting and coding data). One of the problems with the statistical analysis issues is there is apparent redundancy. For example, "were any variables not under study that affected the causal factors handled appropriately" seems to overlap quite a bit with "were potential confounders and effect measure modifiers handled appropriately". By asking the same question twice, it provides twice the weight for that consideration, or so it would seem.	
Methods	Some of the criteria for assessing methodological quality are questionable: 3. In what way is the diagnostic approach a consideration about internal validity. Isn't diagnostic rigor about external validity issues? My understanding of what the authors are attempting to address when rating "quality of evidence" is internal validity (i.e., the degree to which we can confidently infer that change, or group differences, in the outcome can be attributed, at least in part, to the independent variable).	Quality can be defined as “the extent to which all aspects of a study’s design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error.”[1] (In the EPC methods, core elements of quality include the degree to which participant groups are comparable, as well as methods for selecting participants in observational studies.[2]

Section	Comment	Response
Methods	Sometimes, it wasn't clear how final decision for quality of study was made. a. For example, one of the most internally-valid behavioral treatment studies (i.e., Dawson et al 2009) was rated "Fair". The study was "marked down" for not evaluating for differences in attrition, but attrition was very low so it isn't clear this was necessary. It was marked down for "hold steady concomitant intervention, but potentially confounding interventions are unlikely explanations for the results because the treatment was so intense. It was marked down for not testing for at least 3 mo follow-up, but failure to do so doesn't affect its internal validity.	We re-examined the quality scoring for that particular study and agree that the study should not have been marked down for not reporting on concomitant interventions. This change moves the study from fair quality to good quality. We recognize that studies may sometimes have measured elements that are not reported in the published manuscript (such as attrition, in this case), but we are necessarily limited to consider only published information about a study. We think that the threshold of assessing concomitant interventions (which are not required to be held steady) is a reasonable one. There are certain circumstances where assumed control interventions (e.g., see Sallows, 2005) have proven to be delivered at such high levels of intensity that they may not reflect an appropriate control condition. Not assessing 3-month follow-up did not affect the overall quality score for this study. Follow-up measures are important in assessing whether change persists after intensive interventions have been withdrawn; although we recognize that this is only one of many important components of study design and quality assessment.
Methods	Sometimes the application of the quality criteria was incorrect. a. To illustrate, the Kasari et al study was examined closely. This strong study was marked down because "drop out evaluated for differences" was not reported. But it was reported in one of the primary articles that reported its results. Additionally, it was marked down for "held steady concomitant interventions". However, this is one of the few behavioral treatment studies that did so very well. All participants were in the same all-day center-based program.	We re-assessed the quality scoring on this paper and have corrected an erroneous score which changes the study's quality rating from "poor" to "fair."

Section	Comment	Response
Methods	The emphasis on the "natural environment" is misplaced. a. The setting of the measurement does not ensure "representativeness" of the scores in a scientific sense. "Representativeness" in a scientific sense means "stable, in the group design sense of the word, across relevant contexts". Measuring an outcome in the home or school can result in not controlling multiple influential variables, which in turn, can reduce the probability that scores will reflect children's characteristics, instead of contextual variables. The common come-back to this objection is to use parent-reports. These, too, have problems, particularly when the parent is the treatment-implementer or is not blind to the treatment group assignment.	Measuring outcomes in the natural environment speaks to applicability, which is a core element of the EPC methodology intended to help readers understand the environment in which outcomes were measured in order to understand the applicability of the results. The three questions used to collect data on applicability were not used as part of the quality score. We have clarified this in the methods.
Methods	The conceptual model for the review was problematic. 1. The model in Figure 1 (analytic framework for therapies for children with ASD) calls proximal outcomes with those measured in the clinical setting, distal outcomes with functional outcomes measured outside the clinical setting, and even more broadly-defined outcomes with long-term outcomes. The problem with this sort of artificial lumping of disparate concepts is that it leads to incorrect conclusions. For example, all of the particular skills indicated under "targeted outcome in clinical setting" can also be measured outside of the clinical setting	We agree with the observation that wording in analytic frameworks is inherently difficult given the complex concepts that must be conveyed using little verbiage. The conceptual model proved to be a useful starting point and was shaped by our Technical Expert Panel, as well as engagement with the community. When integrating the conceptual model with the literature, however, we were mindful of the complexity of measuring outcomes across time and across settings. We appreciate the thoughtfulness of the reviewer and acknowledge these difficulties. For the purposes of the current review, we will retain the conceptual model, but we fully agree that a denser, more complex conceptual model would be a useful endeavor for the autism research community to tackle.

Section	Comment	Response
Methods	2. Additionally, the importance of specifying the setting of measurement for behavioral treatments is that it conveys something about the degree of generalization outcome measures indicate. The "clinical setting" is the "treatment setting" for the behavioral treatment and these can be in the home or in school, not just the "clinic".	Our use of the term "clinical setting" was intended to refer to any setting in which treatment occurred, including home or school. We have changed it to "treatment setting" per the recommendation.
Methods	One of the "functional outcomes" listed in the figure (i.e., academic engagement) are not functional in and of itself (e.g., engagement is not sufficient for learning).	We agree that engagement is not necessarily sufficient for learning, but would assert that it is necessary for learning. As such, modification of skill necessary for learning might be viewed as functional.
Methods	The other "functional outcomes" are what intervention researchers call "distal" outcomes, they are not the proximal skills the treatment targeted and they can be measured either inside or outside of the treatment setting.	The nature of this report is to attempt to define outcomes for a lay audience. As such, the term functional seemed more appropriate than the term used by intervention researchers
Methods	The "long-term outcomes" have mixed the concept of "long term" with even broader or more distal outcomes such as "social integration".	We would argue that the concepts do overlap to a large extent in terms of thoughts regarding treatment and intervention planning for children in this age range in terms of common goals.
Methods	The overall result is that it isn't clear that the concepts of "generalization" or "transfer" or "maintenance" have been tested separately from the concepts of "distalness" or "breadth" of outcome. The confounding of these important concepts leads the authors to claim we know less than we do.	Again this framework is to be interpreted through the lens of how consumers might approach treatment choices. The framework generally reflects frequent initial targets, broader goals, and ultimate contribution to functional outcomes. This simple schematic is not meant to adequately capture or operationalize the subtleties of intervention research.
Methods	Some quality criteria need clarification.1. If no random assignment, how was "appropriate comparison group" judged?	We added information to the Methods section about our approach for assessing the comparability of comparison groups.

Section	Comment	Response
Methods	Some quality criteria need clarification. 2. Mere reporting of attrition doesn't indicate whether there was differential attrition according to treatment group. How was this information used?	The use of the quality components is described in the Methods of the report. We understand that not all reviewers or researchers will agree with our approach to assessing quality.
Methods	Some quality criteria need clarification. 3. The requirement to have treatments be described in sufficient detail in a journal article so that they are replicable is entirely unrealistic for complex behavioral treatments. Later in the report, this was acknowledged. How was this information used?	While it may be difficult to describe a complex behavioral intervention in a journal article, we did give "credit" for providing a reference to a published manual or methods paper that would provide this description. Adequate description of the intervention is essential to understanding observed outcomes and to synthesizing the literature, even if the intervention is complex.
Methods	Some quality criteria need clarification. 4. How was "validity" of the outcome variable determined? A variable has validation evidence for a particular purpose in a particular population. Many non-norm-referenced tests are likely to lack the type of "validity" evidence that norm-referenced tests will have, and the type of evidence for the latter class may be irrelevant to determining whether the outcome measure is sensitive to change or intervention effects in children with ASD (the type of validation evidence that is relevant). If validation information was only required for certain types of outcomes, is it accurate to say it was a criterion? If any type of validation evidence was considered as good as sensitivity to change or treatment effects validation evidence, then this doesn't seem reasonable. It is questionable that "any type of validity evidence is better than no evidence of validity" when it comes to judging the quality of the treatment study.	The reviewer is correct in noting that assessing the validity of outcomes measures is a complex process. We attempted to identify which outcomes in use had been studied for validity and reliability, and at minimum required that authors provided a reference for the outcome used.

Section	Comment	Response
Methods	Some quality criteria need clarification. 5. What if the treatment fidelity is low? Is monitoring fidelity sufficient when it is low? Doesn't seem like it should be.	We agree that ideally, fidelity should be high. However, for the purposes of measuring quality of the design and conduct of the study, we determined *a priori*, and with the assistance of our technical experts that researchers' attention to measuring fidelity should be considered a measure of higher quality.
Methods	Some quality criteria need clarification. 6. It wasn't clear what "confounders and modifiers captured" or "confounders and modifiers handled" means.	We have added text to the Methods chapter of the review to clarify how we dealt with effect modifiers and confounders.

Section	Comment	Response
Methods	One component that is likely to be controversial is the decision to include uncontrolled case series but exclude single-case experimental studies. Although this decision has little impact on the conclusions of the report, it will probably attract attention because it removed many studies from the analysis. Reasons for the decision are not discussed in Chapter 2.	We have added text to the Methods section to explain this decision. We understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.

Section	Comment	Response
Methods	Later, it is stated that, "while substantial strides have been made in the analysis of single subject designs, measurement of efficacy and effectiveness requires group studies" (p. 136). However, this statement raises at least two questions: What is the point of analyzing studies with single subject designs if they provide no information on efficacy and effectiveness, and why don't they provide such information? Perhaps it would be useful to discuss what can be learned from such studies (e.g., they can serve as demonstration projects yielding initial evidence that an intervention merits further study, and they can test whether an intervention with support from RCT's is helpful for a particular individual). It may also be helpful to be more specific about why they were excluded (e.g., the studies tend to have small samples of convenience, with highly idiosyncratic interventions and measures, making generalizations difficult).	This is an excellent suggestion, and we have added text to this effect. We have noted in the report's Methods section that "Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants, and studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes, nor are they ideal for external validity without multiple replications. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward utility in the treatment setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness."
Methods	The criteria for rating individual studies do not map clearly onto the criteria for evaluating the evidence in aggregate on interventions. Specifically, it is unclear how "risk of bias", "directness" and "precision" are determined. Clear eligibility criteria, random assignment, and thorough diagnostic testing may factor into "risk of bias," but this is not stated. Evaluations of effects outside of the intervention setting and in natural contexts may be evidence of directness, but this is also not stated. Precision is not discussed at all.	We did not clearly describe the SOE methods that are standard in the EPC program[2]; therefore, we have added further text to the Methods section defining each of the components.

Section	Comment	Response
Methods	It is also unclear how evaluations of effects outside of the intervention setting were scored. Ratings from parents, teachers, or clinicians were apparently considered acceptable in some studies (e.g., RUPP et al., 2002) but not others (e.g., Wood et al., 2009).	'Outside of the intervention setting' was used as an index beyond evaluations conducted by an internal rater within study design. Ratings of functioning outside of that setting (parents, etc.) were considered acceptable. You are correct in noting that the RUPP paper should have received a positive score on this question, and we have changed it accordingly. Scores on external validity questions did not factor into the overall quality score.
Methods	In addition, outcome measures in many early intervention studies are standardized tests such as IQ or diagnostic tests such as the ADOS administered by independent evaluators, and it is uncertain whether these are considered to be evaluations outside the intervention setting. Again, these tests appear to have been considered acceptable in some cases (e.g., Eikeseth et al., 2009) but not others (e.g., Drew et al., 2002).	In addition to type nature and types of tests employed in evaluation, it was important to consider who was completing the evaluation in evaluating outcome measurements (e.g., independent/blinded evaluator vs treatment team member). Further, many studies employed numerous and non-overlapping outcome measurements in addition to core assessment measures which were similar. We considered assessments conducted by independent assessors as conducted outside the intervention setting. The Drew study mentioned by the reviewer was appropriately scored as measuring outcomes outside the treatment setting.
Methods	The ABA/EIBI literature as a whole is considered to have provided direct evidence for most outcomes (p. 111 and Table 37), but most individual studies are given poor ratings for their outcome measures (Appendix H).	This is an accurate statement. While many studies within this category received poor ratings for outcome measurement, there were sufficient numbers of studies with improved quality ratings to warrant the classification of direct evidence. The EPC methodology allows for classification of direct evidence if sufficient amounts of adequate quality outcome measurement are present, regardless of additional studies of poor quality in such measurement.
Methods	The report does not define what constitutes a correct randomization procedure, and the ratings in Appendix H appear somewhat arbitrary. For example, why did the raters conclude that Sallows and Graupner (2005 randomized correctly, whereas Aman et al. (2009) did not?	We have added text to the Methods section to clarify our assessment of randomization. We also reviewed the quality scoring for these papers and found an inconsistency in application. We have corrected the randomization score for the Sallows and Graupner study, which does not alter its "good" rating.

Section	Comment	Response
Methods	It is unclear why “it is uniformly impossible” (p. 32) to establish whether there was a primary outcome measure established a priori. Couldn’t one find this information for at least a few studies by inspecting ClinicalTrials.gov or the methodology papers written while many of the studies were ongoing?	We did, in fact, attempt to identify the primary outcome measure from the papers that could be included in the report and were unable to do so with certainly. It is beyond the scope of this review to contact all of the authors, and in order to maintain a systematic approach, we would not contact only select authors. Nonetheless, we have revised this statement to read "extremely difficult" as you correctly note that we could possibly have obtained information beyond the review process to identify the primary outcome. We do think, however, that this is an important weakness in the literature that could reflect publication bias.
Methods	Sometimes the reasons for why different domains were coded nil, + or ++ were not clear. More systematic clarification of these decisions might be helpful in some cases. The consistent coding and implied discussion of "risk of bias" was impressive.	We have added additional detail about our quality assessment methods in Chapter 2 of the report.
Methods	Note that the text in the methods categorizes interventions using different terms than the tables in the Introduction or the Results sections. Otherwise, the methods section seems appropriate.	We had initially included a table indicating our *a priori* categorization of interventions in the review. However, as not all interventions included in the table were discussed in the report, the table caused some confusion, and we have eliminated it in the final version of the report.
Methods	I felt the inclusion/exclusion criteria, search strategies, outcome measures, and statistical measures were logical and pretty standard for reviews of this type. I question why non-English articles were exclude if they otherwise met criteria. I identified no major discrepancies.	The use of non-English articles would have required the use of translation services for which we did not have resources. We are not aware of studies published in other languages that would have changed the conclusions of the report.

Section	Comment	Response
Methods	Thank you for the opportunity to read "Comparative Effectiveness of Therapies for Children with Autism Spectrum Disorders". This clearly was a considerable effort. The authors have clearly and adequately discussed the articles. I have several concerns regarding the manuscript. First, excluded from the review were medical studies with fewer than 30 participants and behavioral/educational/allied health studies with fewer than 10 participants. There is no logical rationale for using these exclusion criteria. There are many studies that do not fall into this limited category that have made significant improvements in the treatment for children with Autism Spectrum Disorders.	We have tried to stress our rationale more clearly in the Methods section of the review. We felt that given the greater risk associated with the use of medical interventions, it was appropriate to require a greater sample size to accrue adequate data of safety and tolerability, in addition to efficacy. We restricted the review to medical studies with at least 30 participants given that most studies of medical interventions for ASD with fewer than 30 subjects report preliminary results that are replaced by later, larger studies. We feel that this restriction did not eliminate specific medical therapies from the review as treatments are typically assessed in larger studies following their preliminary investigation. The N of ten was selected in consultation with our content experts and Technical Expert Panel as a minimum threshold for comparing interventions. Moreover these sample size constraints are not uncommon in the systematic review/comparative effectiveness review literature.
Methods	Second, studies using single subject designs were not included. This also does not make sense. Single subject designs are well-accepted in the scientific community and should have been included. Single subject designs have the advantage of selecting individuals with specific characteristics that may be responsive to a particular intervention. By excluding studies with this type of design, it would be difficult to adequately discuss effective therapies.	We understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward

Section	Comment	Response
		their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness. We have added text to the Methods chapter of the review to clarify these points.
Methods	Next, and similarly, I am confused as to why the authors only selected studies that were published from January, 2000 to January, 2010. If the purpose was to compare the effectiveness of therapies for children with ASD, this would certainly limit this goal.	We restricted the review to those studies published from 2000 forward in consultation with our Technical Expert panel and Task Order Officer and based on the following points: -Good coverage of older literature in several published systematic reviews (e.g. National Standards Report,[3] *Clinical Evidence* review[4]) -Diagnostic shifts following the 2000 revision of the DSM-IV, which refined the definition of Pervasive Developmental Disorder Not Otherwise Specified to correct an error that allowed this diagnosis to be ascribed when there was impairment in only one developmental area (i.e., social interaction, communication, or stereotyped behaviors, interests, or activities). The definition was clarified to require fundamental core social impairment in addition to either/both communication impairment or the presence of stereotyped behaviors, interests, or activities. -Changes in available assessment methodologies and the introduction of 'gold-standards' of ASD assessment during this same time period; specifically, the commercial release of the Autism Diagnostic Observation System in 1999 and the revised version of the Autism Diagnostic Interview - Revised in 2003 allowed researchers to introduce metrics of sample comparison relative to core characteristics of ASD itself during this time period. As such, focusing the review on this decade of research allowed us to speak to the inclusion of such measurements and, when included, specific behavioral differences relative to core symptoms that may affect thinking about the key elements of interventions and therapies. These points were noted in Chapter 2 of the review; however, we have expanded on that text to clarify our rationale.

Section	Comment	Response
Methods	To illustrate my point, 3,248 articles were excluded from this review in the first literature search and 513 full text articles were later excluded. That is a very large number of excluded. In fact only 173 articles were reviewed, which is only slightly over 4% of the published articles included. It is difficult to believe that this type of review could yield an unbiased comparative effectiveness.	Systematic reviews by nature must cast a broad net to ensure that relevant studies are not overlooked; however, fundamental to the process of a review is the development of criteria to allow reviewers to effectively and systematically winnow down a diffuse body of literature to the highest quality evidence most germane to the questions addressed by a review. Similarly, the scope of systematic reviews must be manageable in order to provide meaningful synthesis of studies included, and this body of included literature is in line with other comparative effectiveness reviews (CERs). We scanned recent CERs published by the AHRQ. The number of included studies ranged from 45 to 216 with an average of 111 included studies. We are confident that the 183 papers (some studies were added in an update conducted while the report was undergoing peer/public review) assessed in the current review effectively capture recent ASD research.
Methods	Next the authors go on to say that half of the 4% of articles selected are “poor”. Most of these articles were probably published in peer-reviewed journals. While I agree with many of the reasons for describing an article as “poor”, perhaps a better use of energy would be to evaluate which journals have the most rigorous criteria in regard to research methodology.	Regarding the idea of judging quality of individual studies based on which journals they are published in, it is well known that even the best journals may publish studies with a range of quality and rigor and to not look at each individually would be counter to accepted systematic review methods.
Methods	Some minor concerns regard non-politically-correct terminology, such as “parent training” rather than “parent education”.	We appreciate your comment; however, “parent training” is commonly used in the ASD literature, and we typically reflect the terminology used by authors in our reports.

Section	Comment	Response
Methods	I am also not sure that this limited review succeeded in "comparing the benefits and harms of treatment interventions" (page 31, line 15).	This review is considered a comparative effectiveness review and adheres to the AHRQ's standards for such reviews; however, we were limited in making comparisons among therapeutic approaches by a lack of data available in the published literature. We tried to identify elements such as subgroups of populations which may benefit from interventions but data are not yet available to do so with confidence.
Methods	I am not sure that there would be a meaningful difference in the quality scoring algorithm (page 44) and that the specific areas and cut off criteria are important to determining the effectiveness of the intervention.	This comment appears to suggest that it is not important to use specific criteria to determine the quality of evidence. We respectfully disagree. In terms of the specifics of our criteria, we tried to incorporate the most crucial aspects of study design and sought to balance the importance of each individual criterion within the overall assessment of quality.
Methods	Under the applicability section (page 44), how was external validity rated if all treatment was implemented in natural settings?	In terms of applicability, studies were not "rated" as such on external validity; rather information on applicability was collected and is reported. These factors are not, however, incorporated in the quality scoring. That said, a study in which all treatment was conducted in natural settings would have been coded affirmatively. We have clarified the use of applicability information in the methods text.
Methods	Exclusion and Inclusion criteria do not make sense and no justification is provided.	We have expanded our text on inclusion and exclusion criteria, which are described in Chapter 2.
Methods	good description, and sound methods.	Thank you for your comment
Methods	It was not clear from the EPC report why 10 was used as a sample size cutoff for studies of behavioral, educational, allied health, or CAM interventions, while 30 was used as the cutoff for medical studies.	We have tried to stress this more clearly in the Methods section of the review. We felt that given the greater risk associated with the use of medical interventions, it was appropriate to require a greater sample size to accrue adequate data of safety and tolerability, in addition to efficacy. We restricted the review to medical studies with at least 30 participants given that most studies of medical interventions for ASD with fewer than 30 subjects report preliminary results that are replaced by later, larger studies. We feel that this restriction did not eliminate specific medical therapies from the review as treatments are typically assessed in larger studies following their preliminary investigation. The N of ten was selected in

Section	Comment	Response
		consultation with our content experts and advisory panels as a minimum threshold for comparing interventions. Moreover these sample size constraints are not uncommon in the systematic review/comparative effectiveness review literature.
Methods	number one they are frustrated, can not even communicate , government does not help them , very hard for family	We appreciate the difficulty that families of children with ASD experience, which is one of our motivations for doing research in this area.
Methods	See previous comments about the overly narrow definition of scientific research methods and resulting exclusion of a large body of important evidence on autism treatments.	This comment seems to be referencing single subject design studies. As noted, we understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.

Section	Comment	Response
Methods	Also, it appears that the reviewers characterized some studies incorrectly as evaluating "ABA" intervention, when the only information the study authors provided about the intervention was anecdotal, and there was no evidence that the intervention described as "ABA" was actually designed and overseen by professionals with legitimate training and credentials in behavior analysis; indeed, it is very likely that no such professionals were involved. Examples are studies by Sheinkopf & SIegel and Magiati et al.	We have changed our description of the general category of early intensive intervention studies based on ABA principles to Early Intensive Behavioral and Developmental Interventions.
Methods	Some studies that did evaluate bona fide ABA intervention were mischaracterized. For instance, a study on which I was an author (Howard et al, 2005) is described incorrectly as providing early intensive ABA intervention in an "academic" clinic or center. In fact that intervention was delivered in homes, preschools, and a variety of other community settings. There was no clinic or center. I am very familiar with most of the other studies of early intensive ABA, and believe several of them were also mischaracterized and misinterpreted -- the research methods used as well as the nature of the interventions evaluated.	The intervention settings noted in the report for this study were correctly listed as home, school, and community; the practice setting, which refers to the first author's affiliation as provided in the paper, for the study was listed as "academic." In this case, the affiliation is listed as California State University, which we would consider an academic setting. These reporting conventions for practice setting and intervention setting are described in the Methods chapter of the report.

Section	Comment	Response
Results	In this section on Early Intensive Behavioral and Developmental Interventions, it is indicated that the studies on the Early Start Denver Model were included. However, I don't see these studies referred to in the text. For example, in the summary of the literature, there are sections on studies of ABA based approaches and parent-training, but the ESDM is not described in either. Also, the ESDM studies are not included in Table 11. ESDM is listed in the list of abbreviations at the bottom of Table 11, but it is not listed in the table itself. Perhaps the reader should be referred to Table 27 which does include the Dawson et al., 2010 study?	The reviewer is correct - this was confusing. We have moved discussion of the ESDM study to KQ7, including the reference in the table. We have deleted the acronym from the table.
Results	Perhaps this would be a good place to note that many of the studies of behavioral methods often used to treat challenging behaviors, such as functional behavior analysis and positive behavior support, utilize single subject designs and thus were not included in this review.	We have added a statement in the Results section that "Many of the studies of behavioral methods often used to treat challenging behaviors, such as functional behavior analysis and positive behavior support, included fewer than 10 participants with ASDs and thus were not included in this review."
Results	Computer-Based Educational Approaches. I recommend including in this section a study conducted by Whalen et al. of the efficacy of TeachTown. See Whalen et al., (2010) Efficacy of TeachTown: Basics computer-assistant intervention for the Intensive Comprehensive Autism Program in Los Angeles Unified School District. *Autism,* May 18 (Epub ahead of print).	Thank you for pointing out this paper; we have added it to the report.

Section	Comment	Response
Results	Dietary and Other Medical Interventions – Studies of the efficacy of the casein-gluten free diet don't appear to be included, even though this is a very popular intervention among parents of children with ASD. For example, Elder et al conducted a small RCT and found no benefits. See Elder, JH et al., (2006) The gluten-free, casein-free diet in autism: Results of a preliminary double blind clinical trial. JADD, 36: 413-20. Perhaps this study was too small to include in the review?	You are correct in that the Elder study, and other studies of the casein-gluten free diet were too small to be included.
Results	The summaries were sometimes incorrect. 1. The report that RPMT demonstrated shorter improvement on word acquisition isn't accurate-it was PECS that showed the short-term effect over RPMT.	We have corrected this text.
Results	Review of systematic reviews (page 133, line 32 – 41): my understanding is that the "effect sizes" from the Reichow and Wolery reviewer were pre-post change standardized mean changes NOT between group differences. Thus treatment effects were not estimated in that review. This is an example of the slippage in logic that is now prevalent in autism treatment literature (change during treatment phase is being equated with change because of a treatment).	We agree that this is a problem with the literature. We have ensured that the review does not claim that treatment effects were measured.
Results	It just isn't true that single subject experiments cannot be used to test efficacy of a treatment, as long as the dependent variable is (a) rapidly malleable or (b) reversible. The issue is that this review is mostly interested in another type of dependent variable: those that are difficult to change and changes that endure long after treatment is withdrawn.	Thank you. We have modified the text in this section.

Section	Comment	Response
Results	It isn't accurate that there are no treatment comparison studies (page 137, line 32 – 33). The Kasari et al study is one such study. Yoder and Stone, 2006 is another.	We have modified the text in the Results and Executive Summary to state "few comparison studies" and have cited the relevant literature.
Results	The concept that "behavioral interventions studies uniformly failed to measure outcomes beyond the intervention period, and therefore, we cannot assume that effects were maintained over time", may not be right. For example, Kasari, et al, joint attention and symbolic play study measured and found treatment vs control group differences on at least outcome (language) a 1 year later, I believe.	We have modified the text in the Results and Executive Summary to "behavioral intervention studies rarely measured outcomes beyond the intervention period" and have added additional text describing these outcomes.
Results	The lack of information regarding the relation of treatment outcomes to long-term functional outcomes (KQ4) is likely due to that information not being available in the articles you accepted but being available in other sources (e.g., Kasari has at least one paper in which she shows that post-treatment initiating joint attention [i.e., declaratives] outcomes of treatment are related to much later language scores).	Thank you for pointing out this oversight; we have added text describing this study in the KQ4 sections of the Executive Summary, Results, and Discussion sections of the report.
Results	However, there are some ratings that appear inconsistent or questionable, as discussed in connection with the Methods. This problem arises more often in connection with behavioral and educational studies than with studies on medical or allied interventions, perhaps because of the greater range in research designs and outcome assessment. In any case, this inconsistency may contribute to difficulties in grading individual studies as good, fair, or poor. The grades do not always make sense. For example, few reviewers would rate the Penerai et al. (2009) study as good while rating the Dawson et al. (2010) study as only fair. Similarly, few reviewers would rate Boyd	We reviewed our quality scoring for these studies and noted changes that should be made. The Dawson paper should have received a "good" quality rating, while the Kasari paper should be "fair." We have corrected these scores in the review.

Section	Comment	Response
	(2001) as fair while rating Kasari et al. (2006) as poor.	
Results	In addition to clarifying the rating criteria, it may also help to use the GRADE approach of initially ranking RCT's higher than observational studies, and then downgrading RCT's for methodological weaknesses and upgrading observational studies for methodological strengths.	Indeed, there are many similarities between the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach and that adopted by the EPC program. The primary difference is that the EPC approach does not include applicability within the assessment of SOE and allows for multiple approaches for combining the components (risk of bias, consistency, directness, and precision) into an SOE assessment. For a full description of the EPC Strength of Evidence Approach please see the EPC methods guidance.[2]
Results	I found Tables 11 and 12 less helpful than the summary tables, in part because it was hard to keep in mind what the G's meant across studies. I wonder if the G's could be bolded in terms of "target" groups or color coded to represent similar concepts -- nto sure what else to do but because what was in the G's varied the tables not to add much to the written text, in contrast to some of the other tables (e.g., adverse effects; Table 37) which were very, very useful.	Space constraints require shortening of group designations; however, we have attempted to ensure that text and table descriptions of groups are clear and that meaningful outcomes are reported to bolster the utility of the tables.

Section	Comment	Response
Results	The authors nicely spell out the key questions and sub-questions in the introduction. However, in the results, they only discuss the key questions, with little to no synthesis around the sub-questions. For example, KQ1 sub-questions KQ1a and KQ1b relate to short-term effects, while KQ1c and d relate to longer term effects. That is a very appropriate distinction. However, the discussion in the Executive Summary, as well as in the main text of the Results section, does not distinguish between short-term and longer-term effects. There is some information on length of follow-up in some of the tables, but it would be best if there were some summary text discussing the length of follow-up and directly answering (or saying that there is insufficient evidence to answer) the sub-questions.	We have added text to Chapter 1 to clarify that we have addressed sub-questions within discussion of the appropriate overarching key question given a lack of literature distinctly addressing specific sub-questions.
Results	Overall, the results are difficult to read, especially in regards to the behavioral interventions, and again, the groupings of trials and methods seems inconsistent. See comments below about the interpretation of the strength of the literature in support of ABA- based interventions as compared to other behavioral interventions.	We have simplified and clarified the presentation of results within our organizational framework in addition to providing clearer explanation of the framework and underlying rationale. Please see additional specific concerns referenced below.
Results	In particular, it is unclear why a good quality RCT that showed no difference between treatment arms is being used to say that there is moderate strength of evidence in support of ABA. The only support for effectiveness in that study comes from baseline-to-follow-up changes in outcomes. This type of evidence also exists for therapies in the social support, play/interaction-based, and allied health categories, but does not seem to be weighted heavily due to a lack of a control group.	As noted above, we have moved the Sallows and Graupner (2005) paper from this section into the modifiers section. This paper showed group effects in both intervention groups (parent-directed and professional-directed), with no clear difference in treatment intensity or outcome between groups. As also noted above, the overall strength of evidence assessment for UCLA/Lovaas is now 'low.' For the purpose of effectiveness of the UCLA/Lovaas intervention, this study does, as noted, represent a prospective case series.

Section	Comment	Response
Results	The Results for medical therapies and allied health interventions are easier to read than for behavioral interventions. I wonder, however, if there may be a way to make the tables more readable. It is hard to see, without a lot of effort, which studies had positive outcomes and which ones didn't.	We have simplified and clarified the presentation of these results.
Results	As a more minor note, many of the studies report results as changes in IQ scores. There should be some mention of the huge methodological challenges in assessing IQ in children with ASD (given communication challenges and unevenness of skills) and questions as to the use of IQ as an outcome measure.	This is an oversight on our part. We have included a statement about the issues in assessing IQ in children with ASD in the Discussion section of the report.
Results	Also, there is adequate discussion of possible harms of medical interventions, but not of behavioral interventions. I can guess that most of the primary studies did not measure harms, but some note should be made as to the theoretical possibility of harm and the lack of data on harms of behavioral interventions.	This is an oversight on our part. We have included a statement about harms in relation to other studies in the Discussion section of the report.
Results	Also on a minor note, one page 112, the authors state "In addition, the current research focuses almost exclusively on high functioning children with autism, which excludes the majority of children within this diagnostic classification." What proportion of children with ASD are "high functioning" is the basis of significant controversy and many experts would disagree with the notion that a majority of children with ASD are low functioning (especially with no clear definition of "high" or "low" functioning") or even that a majority would have been excluded from these particular trials (some of which only needed an verbal IQ score of 60, for example.) I would suggest tempering that statement and also adding a	We have altered the statement to specify children considered high functioning on the basis of higher IQ scores.

Section	Comment	Response
	reference to a current, data driven study, (not one of the hundreds of available texts that quote high rates of intellectual disability in autism but cannot trace that figure to an actual study.)	
Results	The studies were described and arrayed well in tabular format. Conclusions were clearly reflected in text. Again, all results were closely aligned with my personal knowledge of the ASD scientific literature and my personal experience over 30 years. I identified no major discrepancies.	Thank you.
Results	The amount of detail is good, however since they reviewed a limited number of studies this is difficult to answer. Figures are clear and descriptive. Many excellent studies were excluded (see above).	This comment may address studies that did not meet our inclusion criteria or timeframe. Regarding our inclusion criteria related to study size, we understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness. Regarding our timeframe, we restricted the review to those studies published from 2000 forward in consultation with our Technical Expert panel and Task Order Officer and based on the following points: -Good coverage of older literature in several published systematic

Section	Comment	Response
		reviews (e.g. National Standards Report,[3] *Clinical Evidence* review[4]) -Diagnostic shifts following the 2000 revision of the DSM-IV, which refined the definition of Pervasive Developmental Disorder Not Otherwise Specified to correct an error that allowed this diagnosis to be ascribed when there was impairment in only one developmental area (i.e., social interaction, communication, or stereotyped behaviors, interests, or activities). The definition was clarified to require fundamental core social impairment in addition to either/both communication impairment or the presence of stereotyped behaviors, interests, or activities. -Changes in available assessment methodologies and the introduction of 'gold-standards' of ASD assessment during this same time period; specifically, the commercial release of the Autism Diagnostic Observation System in 1999 and the revised version of the Autism Diagnostic Interview-Revised in 2003 allowed researchers to introduce metrics of sample comparison relative to core characteristics of ASD itself during this time period. As such, focusing the review on this decade of research allowed us to speak to the inclusion of such measurements and, when included, specific behavioral differences relative to core symptoms that may affect thinking about the key elements of interventions and therapies. Systematic reviews by nature must cast a broad net to ensure that relevant studies are not overlooked; however, fundamental to the process of a review is the development of criteria to allow reviewers to effectively and systematically winnow down a diffuse body of literature to the highest quality evidence most germane to the questions addressed by a review. Similarly, the scope of systematic reviews must be manageable in order to provide meaningful synthesis of studies included, and this body of included literature is in line with other comparative effectiveness reviews. We scanned recent CERs published by the AHRQ. The number of included studies ranged from 45 to 216 with an average of 111 included studies. We are confident that the 183 papers (some studies were added in

Section	Comment	Response
		an update conducted while the report was undergoing peer/public review) assessed in the current review effectively capture recent ASD research.
Results	Page 69 Statement “Aripiprazole also recently received FDA approval for irritability in children with ASD” Comment – recommend revising statement to be specific to FDA indication “Aripiprazole also recently received FDA approval for irritability in children (6-17 yrs old) with autistic disorder.”	Thank you for the comment. We have made this change.
Results	Page 75 Statement “The manufacturer of aripiprazole sponsored both studies. The primary outcome for these studies was challenging behavior indexed by the Autism Behavior Checklist-Community Version Irritability (ABC-C-I) subscale” Comment – recommend revising “challenging behavior” to “irritability”	Again, 'irritability' refers to a specific subscale of the Aberrant Behavior Checklist. That subscale indexes problem or challenging behavior. We don't intend to detail all of the specifics of each study. Clinicians seeking to implement one of the treatments reviewed here should refer to the original papers.
Results	Pages 75-76 Statement “The other study used a dose titration schedule with weekly progression from 2 mg to 5 mg, 10 mg, and 15 mg per day following clinical judgment.” Comment – For additional clarity around dosing in this study, recommend revising to “Aripiprazole was flexibly dosed between 2-15 mg/day, with a target dose of 5, 10, or 15 mg day, based on clinical judgment. Dose increases occurred in 5 mg/day increments (except for the increase from 2 to 5 mg) at intervals of no less than one week”	This comment is asking for more specificity in the dosing guidelines. While we understand that detailed instructions on dosing aripiprazole may be of use to some clinicians reading this document, the purpose of our review is not to provide exact specifications on the use of a given treatment but instead to provide an assessment of the evidence for a given treatment. The current language provides appropriate detail on the treatment that was evaluated in this study.
Results	Page 76 Statement “The two aripiprazole RCTs also provided data on harms (Table 20). Both studies reported on weight gain,216-217 which was greater in the aripiprazole arms (1.3-2.0 kg) than in the placebo arms (0.3-0.8 kg), with a	We highlight the most important harms here. Metabolic parameters over the short term convey little additional meaningful information in this context.

Section	Comment	Response
	statistically significant difference reported in both of the studies." Comment – recommend including a statement about additional metabolic parameters. As noted in Owen et al (Pediatrics 2009;124;1533-1540), there were no statistically significant differences in the median change from baseline to end point for fasting triglycerides, low-density lipoprotein, total cholesterol, high-density lipoprotein, or serum glucose levels.	
Results	Page 104 Statement "RCT or case series data detailing either treatment response or adverse events extending beyond 6 months have not been published for aripiprazole or cyproheptadine." Comment – Please note that a 52 week open label study of aripiprazpole in IAD has been presented as two posters (one for safety/tolerability and one for efficacy) and both are currently under review at two journals. These posters can be provided upon request.	We appreciate the offer, but this report focuses on published work.

Section	Comment	Response
Results	Page 135 Statement "Importantly, the marked improvements in challenging behaviors seen with risperidone and aripiprazole support the study of other atypical antipsychotic medications that do not cause as much weight gain or liability to metabolic disorders." Comment - It is acknowledged that weight gain is seen with both of these compounds in this population. However, it is important to point out that all of the antipsychotics are associated with significant weight gain in younger populations, and that this weight gain appears to be on a continuum, with aripiprazole potentially not causing as much weight gain as the others in younger populations (based on the short-term RCT data). Certainly more information would be beneficial to the field"	We appreciate the comment and highlight the differences between individual medications elsewhere in the report.
Results	See previous comments about studies that were excluded and others that were likely misinterpreted, rendering the conclusions about ABA intervention (and possibly other interventions) in need of substantial revision.	This comment seems to be referencing single subject design studies. As noted, we understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial

Section	Comment	Response
		evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness. We have also reviewed our discussion of interventions employing elements of ABA to ensure that it accurately reflects approaches used in studies and made revisions as needed.
Discussion	it is stated that "one of the most powerfully replicated findings in the available literature is that substantial and varying proportions of children do not seem to demonstrate robust changes in response to early and intensive behavioral interventions." I disagree with this statement. The finding that has been replicated is that, at outcome, many children continue to have significant impairments whereas others are only mildly impaired, if at all, and that some children make rapid, robust changes whereas others make slower, smaller changes. To my knowledge, there has been only one study (conducted by Smith) that showed that children with IQ below 35, as a group, did not appear to significantly benefit from EIBI. As mentioned above, we recently reported at IMFAR (2010) that *both* low and high IQ children responded significantly to early intervention based on the Early Start Denver Model. The lower IQ children start lower and end up lower, but they nevertheless make significant gains. Thus, I think	This is a very important and challenging point. We believe that the International Meeting for Autism Research (IMFAR) data again replicate and add support to the idea that proportions/subgroups of children with autism do not seem to demonstrate robust changes in response to early and intensive intervention (in regards to IQ, Adaptive behavior, and core ASD symptoms). We unequivocally agree that that even small change can be powerful and meaningful; however, it is also important to recognize that a sizable portion of children will not see the same robust change in response to same intervention. We have modified the text to note "one of the most powerfully replicated findings in the available literature is that not all children receiving early intensive intervention demonstrate robust gains and many children will continue to display prominent areas of impairment."

Section	Comment	Response
	it would be more accurate to state that not all children receiving early intensive intervention demonstrate robust gains and many children will continue to display prominent areas of impairment. Keep in mind that, what appears to be a relatively minor gain (e.g. a child who may only gain a few words and signs, become toilet trained, and have significant reductions in challenging behavior) may alter that child's life trajectory in a profound way – leading to a less restrictive environment, more opportunities for learning, and a higher quality of life. I believe it is premature to state that substantial proportions of children do not demonstrate robust changes in response to EIBI until we have carefully defined what is meant by clinically significant, robust changes and conducted more high quality RCTs that assess moderator effects.	
Discussion	it is stated that ESDM was associated with "improvements in adaptive behavior." Actually, significant improvements were noted for IQ, language, and adaptive behavior in this study	We have added the following text: One was a good quality study that suggested benefit for the use of ESDM in young children, with diagnostic shifts in close to 30 percent of children (but still on the autism spectrum) and improvements in adaptive behavior, language, and cognitive outcomes.
Discussion	It is stated that there are no studies directly comparing effects of different treatment approaches but on page 15 a study comparing ABA and TEACCH is described	We have revised this text to indicate that few comparative studies exist.

Section	Comment	Response
Discussion	Some recommendations are not yet realistic to the state of the field. 1. The call for using the "same" outcomes across studies is misguided and unnecessary. The likely outcome of such a call will be to require global measures of status (e.g., ADOS or Mullen or VABS) because it can be argued that many different interventions distally focus on skills that impact these scales (at least their raw scores). The problem with such a recommendation is that they are not sensitive to change, do not measure the skills various treatments tend to directly address (i.e., proximal measures) and do not measure the most important outcomes for children with ASD (e.g., core symptoms).	We have changed the text to read "a consistent set of outcome measures specific to the intended target of treatment." Our intent was not to recommend that the field be limited to a small set of general measures, nor that outcomes based on faulty research logic or empirical evidence be selected; nonetheless, the current state, in which fewer than 150 studies yielded more than 100 outcome measures is such that it is difficult to summarize effectiveness in order to make treatment decisions. This makes it particularly difficult to do statistical combination of studies because of high heterogeneity, although this approach would be helpful given the small sample size possible in most studies of treatment for ASDs.

Section	Comment	Response
Discussion	Additionally, it isn't accurate to say that synthesis of effectiveness of interventions is "nearly impossible" when outcomes differ. That is what meta-analysis is for. One can summarize effect sizes for different types of outcomes (e.g., proximal vs distal). Consensus is not necessary to move comparative effectiveness research forward to "provide a sense of expected outcomes of interventions". It can be argued that when "consensus" is based on ill-informed research design logic for intervention studies and psychometric properties of outcomes such consensus can harm the progress of the field. If such consensus decisions are used to make funding decisions, it can be argued that such decisions would be harmful. At least some professionals' experience with such "working groups" is that they result in very vague suggestions based on ill-informed notions that are pushed by a minority of highly influential individuals.	In conducting a meta-analysis it is important that the studies included used the same or very similar outcomes lest the studies be too heterogeneous to be combined. We made an *a priori* decision not to conduct meta analyses given our team's understanding of the heterogeneity of the ASDs literature as a whole.

Section	Comment	Response
Discussion	Some recommendations are not yet realistic to the state of the field. It is premature to ask researchers, as currently understood by most researchers, to "account for concomitant interventions that might confound observed effectiveness". Attempts to do so will likely be pro-forma (e.g., reports of number of hours of non-project treatments attended per month). Instead, there should be a call for investigating how to measure these concomitant interventions well. Only after we understand what "quality treatment" and "active ingredients" and "engagement in" will we be able to measure "child engagement in the active ingredients of high quality treatments" that are concomitant with those being tested in the study. Concomitant treatments are the single most important class of threats to internal validity in behavioral treatment studies because they occur after, and sometimes in response to, randomization. They are not addressed by randomization. This is an extremely important design issue. Calling for more of the types of questionnaires that are currently used is not helpful. This issue isn't even being discussed by the autism treatment leaders or funders of such work!	We agree that it can be difficult to describe concomitant interventions, but we feel that it is important for researchers to attempt to do so. While standards do not currently exist to do so uniformly, we stand by this measure as important in the quality of ASD research, and one that warrants attention. As you indicate, however, further research is needed on how to best achieve these measures in practice.

Section	Comment	Response
Discussion	Potential alienation of a large group of potential readers (i.e., researchers using single subject experiments). 1. By admitting "case study and case-control designs" but excluding all "behavioral studies with sample sizes at or under 10" and "studies with only individual data or graphically presented data", the authors excluded almost all studies using single subject experimental designs (SSED). As SSED can have more internal validity than some of the included designs, one wonders about the wisdom of this choice.	Thank you for your comment. We understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness. We have explicitly acknowledged these issues in the Methods and Results sections of the report.

Section	Comment	Response
Discussion	Not using typical literature review methods. It isn't clear why the study by study reports do not use the typical effect size metrics in displaying effects. For example, on page 53, under "summary of literature," there is discussion of 64.3 vs. 9.1 percent of children losing a diagnosis of anxiety disorder. This type of outcome can be conveyed using relative risk index, which would be more interpretable than merely reporting the two percentages. If the outcome is continuous (e.g., p. 70 when discussing 12.1 to 14.9 point improvements on the ABC-D irritability scale), Cohen's d or Hedges' g would have been more informative than reporting the raw scores. In most other, study by study description, mere significance or lack thereof is used to talk about effects. This is equivalent to merely saying the confidence interval contains or does not contain the mean of the counter-factual group.	The methods used for this report are those of the AHRQ funded Evidence-based Practice Center network and have been developed by the EPCs over more than a decade. It is not our practice to calculate measures that are not presented as such in the included papers. Therefore, while we agree that effect estimates available to us may not always be optimal, those in the report reflect what was calculated and presented in the authors' analyses.
Discussion	The use of the term "prospective cohort study" will not be transparent to many behavioral researchers. This reviewer interprets this design in the context of a treatment study as a "non-randomized pre-post group comparison design." That is the more familiar descriptor to many behavioral treatment researchers. This reviewer is familiar with the "prospective cohort study design" in the context of studying exposure to a disease and considers its use as a descriptor of a treatment study design as less clear than "non-randomized pre-post group comparison design" when studying treatments. It is recommended that other reviewers be consulted on this issue.	Please see the appendix for a description of study designs. We have added here the term you suggested as an alternative, but have left the term "prospective cohort study."

Section	Comment	Response
Discussion	The report states, "No studies analyzed the ability of end of treatment outcomes to predict longer term functional outcomes in children" (p. 130). Such studies are admittedly rare but do exist (e.g., Kasari et al., 2008; RUPP-AN, 2005).	We have revised the Discussion and Executive Summary text to note that few studies predict longer functional outcomes.
Discussion	The report also states, "Behavioral intervention studies uniformly failed to measure outcomes beyond the intervention period" (p. 130). Again, this is not quite true (e.g., Aman et al., 2009; Howlin et al., 2007; Kasari et al., 2008; Yoder & Stone, 2006).	We have changed the statement about measurement of outcomes beyond the intervention period to "few."
Discussion	It is unclear whether the evidence on PECS as a language intervention is considered insufficient. It appears that there is evidence from high-quality studies that children with ASD learn to use PECS with intervention, although generalization and maintenance often fail to occur.	We agree that there is preliminary evidence on the Picture Exchange Communication System (PECS) as a language intervention. The use of the term insufficient is not intended to mean that evidence is low, rather that additional research is needed to draw a clear conclusion. We required at least 2 good studies for a strength of evidence rating of "good" and at least 3 fair quality studies for a rating of "low," in addition to other criteria described in the report's Methods section. At this time, the PECS and Responsive Education and Prelinguistic Milieu Teaching (RPMT) literature lacks studies meeting these specifications.
Discussion	A discussion of how many studies there are and how few even met standards for inclusion seems appropriate.	This information is provided in Figure 1 (disposition of articles).
Discussion	Even more emphasis on the need for a priori targeting of outcomes and then standard outcome measures might be useful (those these are definitely recognized).	We agree with these suggestions and have expanded text discussing outcome measures in the Discussion and Future Research sections.
Discussion	Some discussion of likely roadblocks in doing better research (need for large samples? Costs? training of staff?) might be helpful.	We agree with these suggestions and have expanded text in the discussion and future research sections.

Section	Comment	Response
Discussion	Please see prior comments about the discrepancy in how the authors have summarized data in support of EIBI/ABA interventions vs. other types of behavioral interventions. Otherwise, the Discussion section reads better than the other sections.	We do not intend to dismiss other behavioral studies. We explicitly note the available evidence for 1) parent training for bolstering social communication skills and managing challenging behaviors 2) social skills interventions 3) play and interaction-based intervention and 4) CBT. As we note in the evidence report the sheer number of studies conducted on early intensive developmental and behavioral interventions/ABA intervention means that a majority of the evidence regards these approaches. Attention and availability should not be confounded with an implicit endorsement of any sort.
Discussion	It may be worthwhile to include some of the data in the Discussion section in the Executive Summary.	We have expanded our discussion of issues in ASD research in the executive summary.
Discussion	It would be helpful to have a more detailed section on future research directions.	We have expanded the section on future research.
Discussion	I felt the extent of knowledge and limitation of the review/studies was described accurately. The case for extensive additional research is clearly stated, with an appropriate call for greater subject specificity in both phenotypic and (hopefully) genotypic domains. The current lack of robust and specific therapeutic outcome data is adequately stated.	Thank you.
Discussion	The findings of this limited review are clearly stated. This manuscript didn't seem to focus on future research.	We have expanded the section on future research.
Discussion	See above note [in executive summary comments, established ways of including parents in therapeutic efforts, importance of executive summary]	As noted, some studies in the review address parent and peer training approaches, though there are not enough data to draw specific conclusions. Additionally, we agree that the executive summary is particularly important for a lay audience and appreciate your attention to it. We have expanded our text throughout the Executive Summary to comment on issues in the ASD literature and future research directions.

Section	Comment	Response
Discussion	See previous comments [about the exclusion of studies addressing problem behaviors, interpretation of ABA approaches].	We stress the usefulness, but also the limitations, of single-case experimental designs, which are frequently used to assess the effects of ABA interventions on problem behaviors. As noted, we understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.
Tables	Table 10 – I had a hard time figuring out what the table was illustrating. For example, under RCTs, what does (n = 27) refer to? And what do the numbers in the table refer to? Numbers of studies reviewed? It is not clear.	We have emphasized statements and added footnotes to clarify these tables, which were meant to summarize characteristics of the literature meeting our criteria and addressing KQ1.

Section	Comment	Response
Tables	Table 27 - Dawson et al. 2010 study – It is stated that adaptive behavior improvements were see in both groups. Actually, improvements in adaptive behavior were significantly higher for the ESDM group than the community-based interventions. Also, significant improvements were found for motor skills and daily living skills on the VABS. It should be clarified that diagnostic shift greater toward milder diagnosis (PDD-NOS) for ESDM group.	We have added this additional information.
Tables	Table 37 - There are some inconsistencies between the text and Table 37. For example, the text describes the evidence on the effects of behavioral interventions on IQ as consistent (pp. 109-110 and Table 28), whereas the table describes them as inconsistent.	We have provided correction to table 37 and additional explanatory text to clarify inconsistencies. In part, some confusion is likely based on the use of our operationalized classification of 'consistent' as well as the word consistent within the review. We have made attempts to revise areas causing such confusion both within tables and text.
Tables	Table 37 is confusing. Why are only some intervention categories included in the table under "adaptive behavior"?	In assessing strength of evidence (the focus of this table) EPC authors are instructed to select major intervention/outcome pairs on which to make this assessment.
Tables	Evidence Table - Dawson et al., study - a few corrections are needed (Note that the number of errors found for this study suggests that the other studies listed should be carefully checked as well):• Study was published in 2010 rather than 2009• Intervention setting was the child's home• Assessments: Yearly assessments were conducted by UW examiners blind to group status for both groups (G1 and G2) • Provider: Bachelor's level therapists supervised by Ph.D. level clinician with consultation from clinical	We have corrected these data and verified the data in evidence tables.

Section	Comment	Response
	psychologist, speech pathology, pediatrician, and occupational therapist, as needed. • The following labels/numbers also need to be corrected:Baseline MeasuresCommunication/ Should be labeled: IQ or Early Learning Composite Scorelanguage:MSEL scale score,mean ± SD:Early-learningcomposite:G1: 61.0 ± 9.2G2: 59.4 ± 8.6G1/G2: P = 0.530Receptive language:G1: 21.1 ± 4.7G2: 21.2 ± 3.8G1/G2: P = 0.530 Should be P = 0.920OutcomesCommunication/ Should be labeled: IQ or Early Leaning Composite Scorelanguage:MSEL scalescore, 2 years,mean ± SD:Early-learningcomposite:G1: 24.2 ± 17.6 Should be 78.6 + 24.2G2: 66.3 ± 15.3G1/G2: P = 0.044Also, there are a few instances where there is an end parentheses that looks unnecessary:VABS communicationscore, mean± SD:G1: 68.4 ± 7.6)G2: 69.6 ± 7.3)G1/G2: P = 0.577Overall ratings:ADOS severityscore, 2 years,mean ± SD:G1: 7.0 ± 1.9)G2: 7.3 ± 1.8G1/G2: P = 0.422	
References	See previous comments.	This comment seems to be referencing single subject design studies. As noted, we understand that single subject design studies are commonly used in behavioral research in children with ASD. Because there is no separate comparison group in these studies they would be considered case reports (if only one child included) or case series (multiple children) under the rubric of the EPC study designs. Case reports and case series can have rigorous evaluation of pre- and post- measures, as well as strong characterization of the study participants. Studies using this design that included at least 10 children were included in the review. Studies of this type can be helpful in assessing response to treatment in very short time frames

Section	Comment	Response
		and under very tightly controlled circumstances, but they typically do not provide information on longer term or functional outcomes. They are useful in serving as demonstration projects, yielding initial evidence that an intervention merits further study, and, in the clinical environment, they can be useful in identifying whether a particular approach to treatment is likely to be helpful for a specific child. Our goal was to identify and review the best evidence for assessing the efficacy and effectiveness of therapies for children with ASD, with an eye toward their utility in the clinical setting. With the assistance of our technical experts, we selected a minimum sample size of 10 in order to maximize our ability to describe the state of the current literature, while balancing the need to identify studies that could be used to assess treatment effectiveness.
Appendices	Information about the reviewers and authors would be very instructional.	The report lists individuals participating as peer reviewers and authors; AHRQ reports typically do not include additional author/reviewer information.

ABA=applied behavioral analysis; ADOS=Autism Diagnostic Observation Schedule; AHRQ=Agency for Healthcare Research and Quality; ASD=Autism Spectrum Disorders; CBT=Cognitive Behavioral Therapy; CER=comparative effectiveness review; DSM=Diagnostic and Statistical Manual of Mental Diseases; EPC=Evidence based Practice Centers; ESDM=Early Start Denver Model; FDA=Food and Drug Administration; GRADE= Grading of Recommendations Assessment, Development and Evaluation; IMFAR=International Meeting for Autism Research; IQ=intelligence quotient; KQ=key question; N=number; PDD-NOS=Pervasive Developmental Disorder-Not Otherwise Specified; PECS=Picture Exchange Communication System; RPMT= Responsive Education and Prelinguistic Milieu Teaching; RUPP=Research Units on Pediatric Psychopharmacology; SOE=strength of evidence; TEC=Technology Evaluation Center; UCLA=University of California, Los Angeles;

References

1. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care. 2004 Feb;16(1):9-18.
2. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality [posted November 2008]. Rockville, MD. Available at http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318
3. National Autism Center. National Standards Report. Randolph, MA: National Autism Center; 2009.
4. Parr J. Autism. Clin Evid (Online);2008.