NWX-OS-OGC-RKVL

Moderator: Amy Margolis
December 2, 2010
8:34 am CT

Coordinator: Welcome and thank you for standing by. At this time, all participants are in a listen-only mode for today's conference call. Today's call is being recorded and if you have any objections, you may disconnect at this time.

I would now like to turn the call over to Evelyn Kappeler. Ma'am, you may begin.

Evelyn Kappeler: Thank you. I'm Evelyn Kappeler, the Acting Director for the Office of Adolescent Health, and I'd like to welcome you to the Webinar to talk more about how the Department of Health and Human Services along with Mathematica Policy Research identifies programs that impact teen pregnancy, sexually transmitted infections, and associated risk behaviors.

Before we get started, I have a few administrative issues I'd like to cover. All participants should be able to hear the audio and view the slides. If you were unable to log in to the Netconference to view the slides, please be assured that these slides will be posted along with a recording of the Webinar on the OAH Web site within the next four to five days.

Due to the number of people on the Webinar today, all participants are in a listen-only mode. We will be taking questions throughout the presentation using the Q&A function on the Netconference. I will show you how to ask questions in a minute. We will not however be able to monitor the raise your hand feature, so if you have a question, please type it in the Q&A box. If we run out of time to answer all the questions during today's call, we will post answers to the questions on the Office of Adolescent Health Web site following the call.

In order to ask questions during today's Webinar, you will need to type your question into the Q&A link on the upper left-hand side of the Netconference screen. You will see the Q&A link circled on this slide to ensure that everyone knows how to ask a question.

Our agenda for today is to discuss the background for the HHS Evidence Review of Teen Pregnancy Prevention programs, review the methods and criteria used for the HHS Evidence Review, review the initial findings from the Evidence Review, and discuss plans for maintaining and updating that review.

First, I'm going to present on the background of the HHS review to identify evidence-based teen pregnancy prevention programs. The purpose of our review of evidence-based programs is to systematically review the evidence on programs aimed at reducing teen pregnancy, sexually transmitted infections, and associated sexual risk behaviors, to identify the program models with the strongest evidence of effectiveness, and to help advance the evidence-based and the science-based.

Given the high rates of risky sexual behavior among adolescents, the U.S. Department of Health and Human Services is motivated to identify programs

that have proven effective at preventing teen pregnancy or other associated risk factors through rigorous evaluation. We know that nearly half of all high school students have had sexual intercourse and that adolescents and young adults account for half of all new STI cases in the U.S. every year.

Furthermore, the teen birthrate increased 5% between 2005 and 2007 and then declined by 2% between 2007 and 2008. In addition, given the limited funding available, there is an increased emphasis on evidence-based policy making and using resources to support programs that have demonstrated evidence of effectiveness.

The first review of the teen pregnancy prevention research evidence was conducted in the fall of 2009 and in the winter of 2010 and covered research conducted or published between 1989 and 2009. This initial review identified 28 program models that met the HHS criteria for evidence of effectiveness. Detailed information about each of these 28 program models is available on the Office of Adolescent Health Web site at www.hhs.gov/ash/oah.

The Web site also includes a database with all of the studies that were reviewed in this first reviewed, but did not make the evidence criteria. The results of this initial review were released in the spring of 2010 in conjunction with the Office of Adolescent Health Teen Pregnancy Prevention Program grant announcements as well as the Administration of Children and Families' state personal responsibility education program grant announcement.

At this point, I'm going to turn the presentation over to Barbara Broman from the Office of the Assistant Secretary for Planning and Evaluation who will talk about the next steps planned for updating the evidence review. Barbara.

Barbara Broman: Thank you, Evelyn. Let me just tell you a little bit about how we're going to be updating the review.

In the fall of 2010, HHS competitively awarded a new four-year contract to Mathematica Policy Research so that we can maintain and update the review on an annual basis. My office, the Office of the Assistant Secretary for Planning and Evaluation, will be managing this new contract in partnership with the Office of Adolescent Health.

We plan to conduct the annual updates to the review and each update will focus on new research that has not been covered in previous reviews of the evidence. And based upon the findings from the review, we will update the program models for inclusion on the HHS list of evidence-based programs.

We are now preparing to begin our second round of review. We will start by identifying new studies to review through a search of the published literature as well as a new call for studies that are being released today. The Mathematica Team will discuss the call for studies in more detail at the end of the presentation.

Once the new studies have been identified, Mathematica and their partner, Child Trends, will review the studies using the criteria established by HHS and Mathematica, and we expect the release of the findings from the review in the spring of 2011.

Over the course of the new project period, we will be disseminating the findings broadly through Web site materials. We also plan to issue research briefs and reports that we believe will inform the efforts to implement and evaluate programs to prevent sexual risk behavior. Throughout the project, we

will be consulting with experts in evaluation design and methodology on the review criteria and procedures we've established.

And finally, as the evidence-based expands, we will consider making revisions to the review criteria. For example, we may require more recent evidence of sustained impacts or require that impacts be sustained for a certain period of time post intervention.

With that and before we turn it over to Mathematica, we would like to stop here for a few minutes and allow folks to ask any questions of Evelyn and I on this introductory period, so we will pause for a minute and look for questions.

So the first question that we have is, "Why are you updating this review on an annual basis," and that question is very relevant. And we want to just let you know that we think that we want the evidence - I mean basically we want the evidence reviewed to represent the current science by incorporating new research findings from program evaluations on an ongoing basis. And we plan to issue that call for studies and to review the published reviewed literature annually in order to identify that new research. It will really be a tremendous asset for us to be able to do it on an ongoing basis so that we can stay current and continually provide information at that time.

Let me also just introduce before we take any other questions if we have any who will be speaking next from Mathematica Policy Research. It will be Brian Goesling and Chris Trenholm. So in a few seconds, we will turn it over to them to see if there's anything else.

We did - the next question we got is whether we can provide more details on what our evidence criteria was, and I think Mathematica Policy Research will

be going over that in quite a bit of detail and we'd like to not duplicate the Webinar and let them handle that question, so stay tuned.

Another question that came in is are the evidence criteria going to stay the same for the next round of review, and we have not yet started reviewing studies for this next round. And as part of our review process, we will be evaluating the evidence criteria each year and we may make changes over time that will increase the rigor of the evidence standards. So while we haven't made any decisions yet because we haven't started a review, it's quite possible. For example as the evidence base expands, we may require more recent evidence of sustained impacts.

And let's see. WE have another question that has come in. "Will new programs that are deemed evidence based change what OAH is willing to fund under Tier 1 grantees," and I think I'd turn that one over to Evelyn to answer, so here she is.

Evelyn Kappeler: If there are new programs identified in the evidence-based process, those would then be used in any future funding announcements that the Office of Adolescent Health would issue with regard to the replication of evidence-based models and in particular under our Tier 1 programming.

We're still here waiting for other questions as they come in. Okay, we have a question here that is asking, "Are you only reviewing new programs that they've had an RCT evaluation?" And Mathematica Policy Research can answer that during their presentation, but I do want to let you know that we will be looking at high quality QED and RCT evaluations as well.

It appears that a number of these questions are really going to the substance of the presentation, so at this point I'd like to suggest that we move right into the

presentation by Mathematica about the review methods that were used, about this new call, and our future plans for the evidence review. With that, I'd like to turn it over to Brian Goesling and Chris Trenholm from Mathematica Policy Research. After their presentation, we will take questions and I think many of the questions that we have received to date may be answered in this next section. Thank you so much.

Brian Goesling: Thank you, Evelyn. This is Brian Goesling. I'm a Senior Researcher at Mathematica and I'm leading the team that has been conducting the evidence review.

And in this section of the presentation, I'm going to describe the review methods and criteria and then I'll turn it over to my colleague, Chris Trenholm, who will discuss some of the findings from the first review of the evidence. If you have questions as we go along, feel free to keep typing them in to the questions box and we will get to as many as we can at the end of the presentation.

The review process had four main steps. First, we conducted a comprehensive literature search to identify potentially relevant studies for review. Second, we then screened these studies against a set of inclusion criteria that we had determined ahead of time in coordination with HHS. Third, for each study that met the inclusion criteria, the study was then systematically reviewed by a team of two researchers for the quality and execution of its research design. And then finally for studies that passed the quality bar, we examined their impact findings for evidence of program effectiveness. And so in this next section of the presentation, I'm going to talk a little more about the details of each of these steps.

To identify studies for review, we begin by reviewing a reference list of existing research syntheses including a 2008 report by Advocates for Youth Science and Success, two recent reports from the CDC's Guide to Community Preventative Services, a 2008 Heritage Foundation report on abstinence education programs, the Emerging Answers 2007 Report, and two meta-analyses sponsored by the Campbell and Cochrane collaborations.

We also look for relevant studies through an extensive search of the Web sites of research and teen pregnancy prevention organizations such as the National Campaign to Prevent Teen and Unplanned Pregnancy, the Healthy Teen Network, the National Abstinence Clearinghouse, the National Abstinence Education Association, and several other organizations.

To identify any new or unpublished studies, we also issued a public call for studies to a broad group of research organizations, professional associations, and other groups. Studies for this call were submitted by the review team through email and in total received over 100 submissions through this call for studies.

And as a final check, we also conducted a keyword search of electronic databases such as PsycINFO, Medline, and several others.

All of the studies that were identified through this search were then screened against a set of inclusion criteria that we had determined in coordination with HHS, and so to qualify for review, a study must have looked at the impact of a program using quantitative data and statistical analyses. And the interventions we looked at may have focused on a range of different approaches such as encouraging teens to wait to have sex, providing information on contraception, teaching refusal skills, youth development approaches, or discussing the health consequences of sexual activity.

We excluded studies that lacked this type of focus such as research on dropout prevention, job training, or early childhood education programs. We also excluded studies of state or federal policy changes such as policies affecting access to contraception through Medicaid laws.

Studies also had to focus on impacts on at least one measure of sexual risk behavior or its health consequences, and this included measures of sexual activity, including initiation of sexual activity, frequency or number of partners, measures of contraception use, sexually transmitted infections, and also measures of pregnancies or births.

And then finally, the study had to focus on U.S. youth ages 18 or younger and have been conducted or published since 1989. We did not review international research and we did not review programs that were intended for young adults or those over age 19.

For each study that met the inclusion criteria, the study was then assessed by teams of two trained reviewers from Mathematica or our partners, Child Trend and Concentric Research & Evaluation. The main purpose of these reviews was to assign each study a rating of high, moderate, or low based on the rigor and execution of the research design, and these ratings were designed to identify studies that provided credible estimates of a program's impact.

At this stage of the review, we weren't looking at the study's findings for evidence or positive/negative effects, rather the goal of these initial ratings was to assess whether the study's findings positive or negative represented credible estimates of the program impacts.

The standards for the study quality ratings were developed by Mathematica and approved by HHS, and in developing these ratings, we drew on the

standards used by several other evidence assessment projects both within and also outside of the field of pregnancy prevention. So this included the Campbell collaboration, the Department of Education's What Works Clearinghouse, the CDC's HIV-AIDS Prevention Research Synthesis, SAMHSA's National Registry of Evidence-Based Programs and Practices, Blueprints for Violence Prevention, and several other groups that are listed on the right-hand side of the current slide.

And like the standards used by these other evidence reviews and groups, the study rating system we developed for this project focused primarily on the issue of internal validity and whether the design and execution of the study provided credible estimates of program impact. So again at this stage of the review, we weren't looking for evidence of positive or negative effects, just whether the quality and execution of the study design gave us confidence that the study provided a rigorous test of the program's effectiveness.

The highest study rating was reserved for studies that randomly assigned subjects to the study's research groups. Studies using random assignment assure that there are only chance differences between the research groups, and so these studies provide the strongest evidence that any differences in outcome measures between the treatment and control groups can be contributed solely to the intervention. To receive the highest rating, studies were also required to have low attrition of sample members.

In random assignment studies, a loss of study participants can bias the impact estimates by creating differences in the characteristics of the treatment and control groups and this bias can arise from two different sources. One is the overall level of sample attrition among sample members. That is the percentage of study participants who are lost among the overall study sample.

Bias can also arise from differential attrition, which is the difference in attrition rates between the treatment and control groups.

And so to assess the level of sample attrition, we followed standards established by the U.S. Department of Education's What Works Clearinghouse. We chose to follow the clearinghouse for these standards because clearinghouse is one of the few groups to recognize the importance of both overall level of sample attrition and differential attrition as potential sources of bias in the impact estimates. And for these clearinghouse standards, there's not one particular cutoff value to define low sample attrition, rather the cutoff value can vary depending on a combination of the level of overall sample attrition and also the differential attrition between the treatment and control groups.

And the way this works is that in studies where there's a larger difference in attrition rates between the treatment and control groups this requires a lower level of overall sample attrition. While if there's a small difference in attrition rates between the groups, then this permits a higher level of overall sample attrition. And if you're interested to see the specific cutoff values that are formed by the combination of these two factors, details are available both on the OAH Web site where we provide more information on the details of this review and also more information on these cutoffs is available on the Web site of the What Works Clearinghouse.

To receive the highest study rating, a study also had to meet four other criteria. First, there could be no reassignment of sample members for the analysis. That is all the participants who are initially assigned to the treatment or control groups had to be analyzed with that same group, and failing to do this can undermine the initial random assignment and so weakens the strength of the initial study design.

Second, we also required that there were no systematic differences in data collection between the two research groups. For example, if program staff had collected data from everyone in the treatment group but an independent group of staff had collected data from the control group, this would make it impossible to separate the effects of the mode of data collection from the effects of the intervention. So it is important that the same mode of data collection was used for both groups.

We also required these studies to have at least two subjects or groups in each research condition. For example, if a study was randomly assigning schools to the treatment and control groups, there had to be more than one school in each group. And this is important because if the study had only one school per group, it would be impossible for us to separate the effects of the interventions from the particular effects of the school that was in the one group. So it was important to have at least two schools for subjects in each of the research groups.

And then finally, these studies also had to control for any particularly significant differences in baseline characteristics between the research groups. In general, random assignment ensures that the groups are similar at baseline on average. However, for any particular random draw, it's possible that chance differences could arise, and so it's good for researchers to control for these differences in their analyses. And so in order to meet the criteria for the highest rating, a study had to meet all of the criteria listed on the last two slides.

For the moderate study rating, we considered two types of studies. First, here we considered (quasi-experimental) studies with well-matched comparison groups. In these types of studies, participants are assigned to the research groups through a process other than random assignment. However, if the

groups are shown to come from similar populations, we can have some confidence that the differences in outcomes between groups reflect the effects of the intervention. And so for that reason, these studies were considered for the moderate rating.

At the same time, because we cannot rule out the possibility that the groups may differ on some unmeasured characteristic, (quasi-experimental) studies were not considered for the highest study rating.

To receive a moderate study rating, a (quasi-experimental) study first had to establish that the research groups that were being compared were similar at baseline that is before the intervention started, on demographic characteristics, in particular age, race, and gender, and also at least one relevant outcome measure. And so this was necessary to establish that the groups being compared represented similar populations.

The studies were also required to control for baseline outcome measures in their analysis to ensure that any marginal differences in these measures at baseline did not bias impact estimates observed at follow up.

To qualify for a moderate study rating, (quasi-experimental) studies also had to meet two of the same requirements that I discussed for the highest study rating. First, there could be no systematic differences in data collection between the research groups, and these studies also had to have at least two subjects or groups in each research condition.

So for example if you are conducting a (quasi-experimental) study of schools, we require that you have more than one school in your intervention group and your comparison group.

Finally, we also considered the moderate study rating for random assignment studies that did not meet all the criteria for the highest study rating. For example because the study had high sample attrition or for example because these studies had reassigned sample members in the analysis.

If you were a random assignment with these types of problems, you did not automatically qualify for the moderate rating; rather the study still had to meet the other criteria relevant for differences in data collection, for having at least two subjects or groups in each research condition, and in some cases for establishing baseline equivalents.

And then finally studies that did not meet the criteria either for the high rating or for the moderate rating by default were assigned to the low study rating, and a rating of low does not necessarily mean that the program is not effective. Rather it means that we didn't have confidence from the supporting impact study that we could accurately determine whether the program was effective or not.

For the studies that met all of the criteria either for the high or moderate study rating, the review team then went through and examined the study's impact findings for evidence of program effectiveness. In particular, we looked at the direction and statistical significance of the impact findings, the types of outcome measures examined, the length of follow up, and also the analysis samples used to generate the program impacts.

And at this point of the review, studies with the low quality rating dropped out of the review again not because the low rating indicated that the programs were in effective, but rather because we did not have confidence that the impact estimates from these findings provided credible estimates of program effects.

And then finally based on the information collected from the high and moderate rated studies, we then identified programs meeting the HHS criteria for evidence of effectiveness, which I will show you on the next slide.

To meet the criteria for evidence of effectiveness, a program was required to have evidence of a positive statistically significant impact on at least one key measure of sexual risk behavior or its health consequences. And this included measures of sexual activity including sexual initiation, frequency of sexual activity or number of sexual partners. We also considered measures of contraceptive use, including both one-time measures of contraceptive use and also measures of consistency of use. Measures of sexually transmitted infections either biologically tested or self reported, and also measure of pregnancy or birth.

In addition, the impacts had to be shown for either the full study sample or for one of two priority subgroups, which were defined for this review as gender or sexual experience measured at baseline. So for example if a program only showed evidence of impact for a different type of subgroup not on this list, for example socioeconomic status or race, this was not sufficient to meet the criteria. The impact had to be either for the full sample or for one of these two priority subgroups and I will explain the rationale for this subgroup criteria on a following slide.

These HHS criteria allowed for a broad range of different types of evidence to qualify a program as evidence based. So for example as part of our review, we divided programs into eight different evidence categories, which are shown on this slide. All of these criteria - all of these different categories meet the HHS criteria for an evidence-based program, but in some ways they reflect different types of supporting research evidence.

So for example to meet the HHS criteria, the supporting impact study could have been based on either a high quality or a moderate quality study. So in this slide, the top panel shows four different evidence categories for studies with the highest rating. The bottom panel shows four different evidence categories for studies with a moderate rating.

So this means that evidence from a high-quality randomized control trial was not necessarily required for making the list. We also considered (quasi-experimental) studies that met the criteria for a moderate study rating.

In the same way, the program did not have to show impacts for the full analytic sample. We also considered programs with evidence of impacts only for subgroups. For example, a program that showed impacts for females but not males was still eligible for meeting the HHS criteria.

There was no requirement about the duration of impacts. They could be long-term impacts on pregnancy or they could be shorter-term impacts on contraceptive use or different measures of sexual activity.

And finally, there were no requirements for impacts to be replicated across multiple studies. A single study that received a high or moderate quality rating was sufficient for establishing evidence of program effectiveness and meeting the HHS criteria.

And as we'll see when we get to discussing the results, this range of evidence allowed for a broad range of programs to make the HHS list of evidence-based programs.

As I explained on a previous slide, the HHS criteria required programs to show impacts either for the full study sample or for one of two priority

subgroups, either gender or baseline sexual experience. And the criteria were limited in this way to help address an issue that's known in the literature as multiple comparisons or multiple hypothesis testing.

And the issue here is that as you increase the number of subgroups examined in a particular study, you also increase the probability of finding a statistically significant impact just by chance. And so in order to limit this possibility, researchers either limited the number of subgroups they examined or in some cases they will apply a formal statistical correction to account for the number of subgroups examined in the study.

For this particular review, we did not have any requirement for statistical adjustments for multiple comparisons or multiple hypothesis testing. So instead, we chose to address this issue by limiting the number of subgroups considered as providing evidence of effectiveness.

In addition to that, to ensure that we could have confidence in these subgroup estimates, we had to limit the subgroups considered to those that were defined by characteristics that could not be affected by the intervention. And so this meant either subgroups defined by demographic characteristics in the case of gender or subgroups defined by characteristics that were measured prior to random assignment or the start of the program as in the case of baseline sexual experience.

Although commonly featured in the literature, we did not consider evidence from subgroups defined by characteristics that were measured after random assignment. For example, measures of sexual activity that was assessed at the time of the follow up survey.

And the reason for this was that if the subgroup is defined by measures taken after random assignment the composition of the groups could be affected in some way by the nature of the intervention and so we'd no longer have confidence that we're comparing similar populations across the treatment and control groups. And so any differences in outcomes we observed might reflect differences in the compositions of the groups rather than an effect of the intervention.

And so to give you an example of this, on the following slide this shows just a hypothetical example comparing subgroups defined by sexual experience at baseline on the left hand side of the figure versus subgroups defined by sexual activity at follow-up on the right hand side of the figure.

So on the left hand side of the figure, when we're defining the subgroups at baseline we can have more confidence that the composition of the groups is similar regardless of treatment status and therefore we can also have confidence that any differences in outcome measures can be attributed to the effect of the intervention, not the differences between the groups.

By contrast when you turn to the right hand side of the figure, when subgroups are defined at follow-up we no longer have the same type of confidence the composition of the groups is similar regardless of treatment status.

And this is especially true if the subgroups are defined by sexual experience since many if not all of the programs we're evaluating in this review potentially could have impacts on measures of sexual activity. And so in this case we don't know whether any differences in the outcome measures might reflect differences in the composition of the subgroups rather than ineffective at the intervention.

And so to put it another way, one of the main advantages of conducting a random assignment study or a quasi-experimental study with a well matched comparison group is that you're creating treatment and control groups with similar types of students.

But if you then divide these groups up based on characteristics measured after random assignment and characteristics - especially characteristics that could be affected by the intervention, then you no longer have the same assurance of equivalent groups until you undercut one of the main strengths of the initial design.

And so one solution to this problem is to only form subgroups by characteristics that cannot be affected by the intervention such as demographic characteristics or characteristics measured at baseline.

So that provides an overview of the criteria that were used for the first review of the evidence and now I'm going to turn it over to my colleague Chris Trenholm who is going to discuss some of the findings from the first review.

Chris Trenholm: Hi everyone. As Brian mentioned I'm Chris Trenholm and I'm a colleague of Brian's on the review. I'll just walk you briefly through the summary of what we - of the results.

As part of step 1 about 1000 potentially relevant studies were identified through the search that Brian described. From those 1000, 199 studies met the screening criteria and were subject to review, proceeded to the third step. Of those 199 that proceeded to that third step we identified 93 that met the criteria for either a high or moderate study rating and based on that the study proceeded to the fourth stage of the review.

In that fourth stage of the review, those 93 studies that we reviewed, we identified 28 program models that met the HHS criteria for evidence of effectiveness that Brian laid down earlier.

Shown on this slide is just a list of those 28 programs. Overall the 28 programs reflect a range of different program models. Those include classroom based curricula, community based programs, and clinic based services for example. This list is available on the OAH website along with more detailed information on each program and the evidence behind it so for more information I encourage you to go to the OAH website.

Now just a little bit of context for the strength of the supporting evidence associated with these 28 programs, among the 28 programs on the list, 19 are supported by a random assignment study that received the highest quality rating; 9 are supported by random assignment or quasi-experimental studies that received the moderate study rating; 21 of the 28 programs have evidence of impacts that the full analysis sample in the supporting study; 7 programs made the list through evidence of subgroup impacts.

About half of the programs on the list are supported by evidence of shorter term impacts of less than 12 months. The other half are supported by evidence of longer term impacts of at least 12 months. Both of those are right at 12 months.

Just continuing with some sense of the supporting evidence, only 1 of the 28 programs that is supported by evidence is supported by evidence from more than one high or moderate quality study. The other 27 are supported by evidence from a single study.

In looking at the different outcome measures that were considered or that were identified for the basis of being among these 28, 5 programs showed impacts on initiation of sexual activity meaning initiation - meaning impacts on abstinence. Seventeen programs showed impacts on other key measures of sexual activity, examples being frequency of sexual activity, numbers of sexual partners, or recent engagement in sexual activity.

Nine programs showed impacts on contraceptive use; four programs showed impacts on STIs; and five programs showed impacts on pregnancies or birth. Those numbers add up to more than the 28 programs because in some cases the programs showed impacts on more than one outcome.

Now just to continue to provide you a little bit more context for our review in relation particularly to work that's gone before us, we examined the HHS list to see how it compared with other evidence based reviews.

First we compared the list to the most recent What Works list published by the National Campaign to Prevent Teen and Unplanned Pregnancy. If you look at those two, at the two lists you'll find 18 programs that are common on both.

There are 12 programs on the What Works list that are not on the HHS list. Summarizing these, 2 of those 12 fell outside the scope for our review; 4 did not meet our criteria for high or moderate quality study rating; and 6 met that criteria for a high or moderate rating but did not show evidence of statistically significant impact on either the full sample or the subgroups defined by gender based on sexual experience.

Finally there are ten programs on our list or the HHS list that are not on the What Works list. We don't know the exact reasons for this but it's quite likely they fell outside of the scope of the What Works review, have not been - they

had not been reviewed yet by the time of that - of the What Works release, or did not meet the criteria used by that assessment.

We also examined how this list compared with CDC's PRS interventions for high risk. Comparing those two lists there are ten programs common to both. In addition there are seven on the CDC list that are not on the HHS list and of these, three fell outside the scope of the HHS review simply because the studies focused on young adults over age 19.

One program didn't meet our criteria for high or moderate quality study rating and three met that criteria again but did not show evidence of significant statistical impact on either that full sample or the two subgroups that were our focus.

There are 18 programs on the HHS list, on the list that we generated that are not on the CDC list. We don't know the exact reasons for each one but the very likely reason for most of them is that the CDC review focused more narrowly on HIV/AIDS prevention so some of the more general pregnancy prevention programs on the HHS list would fall outside the review scope for the work done by CDC.

Now in terms of the common reasons for not making the HHS list just to give you a little bit of sense for that, first of all you can go to the OAH website and get full details on this. The database will indicate whether the study met the review screening criteria, the rating received, and whether the impact findings met the HHS criteria for evidence of effectiveness. So again I encourage you for full details on this to go to the OAH website.

The most common reasons for not making the list were as follows. First, that the study simply didn't meet the screening criteria that Brian described either

because it was for individuals older than age 18 or because the particular program model tested was not covered by the review.

A second reason among those we reviewed is less than half met the criteria for moderate or high quality study rating. Common reasons for not meeting that rating were failure to establish baseline equivalence for the research groups and using a design that featured only one subject or group in each research condition.

So for example a random assignment study that only randomly assigned one school of the treatment and one school of the control would be - would not meet the criteria for moderate or high.

Some studies that received a high or moderate quality rating showed impacts on measures of attitude during tension so there was evidence of impacts on certain outcomes but they did not feature the measures of sexual risk behaviors or consequences that were required to be among the 28.

First as noted in previous slides, in some cases studies received a high or moderate quality rating but did not show impacts for either the full analytic sample or a subgroup defined by gender based on experience.

Finally let me just briefly mention the plans for maintaining and updating the review. So as noted earlier, we're currently looking for submissions to a new call for studies right now. We are distributing this call through an email in the next few days and it will also be available on the OAH website.

This new call for studies will be limited to new studies or manuscripts not previously reviewed. We do not plan to re-review any studies assessed during the first review of the evidence. If you have any questions of whether your

study was previously reviewed you should again look at the searchable database on the OAH website.

For this new call for studies we're using the same inclusion criteria as for the first review of the evidence -- in particular we're looking for quantitative impact studies that focus on measures of sexual risk behavior and their health consequences.

Studies that look only at measures of attitudes and expectations will not be included in the review. We're also focusing the review on studies of U.S. youth ages 19 and younger so we'll not be including international studies or those that focus on young adults over age 19.

Authors may choose to submit new evidence or findings that build on or expand on previously reviewed studies, for example, looking at data from a more long term follow-up survey or using new analytic methods. However, these studies must be written as a new stand-alone paper. We'll not be reviewing simply tables or partial findings that are submitted as add-ons to a previously reviewed study.

The due date for the submissions is January 7, 2011. The email address for submitting studies is listed on the next slide and also will be listed once the call for studies goes out in the next few days.

For more information on the review process again you should visit the OAH website. The website includes more detail on how the review was conducted, a full list of all the studies we reviewed, and more information on the 28 programs that met the HHS criteria for evidence of effectiveness.

In the coming months we'll also be posting a Frequently Asked Questions document that answers many of the common questions we received and might receive today as well.

If your question is not answered by the information posted on the website or is not answered today, please send an email to pprer@mathematica-mpr.com. We will respond to requests as quickly as possible. This is also the email address for the - to use for new submissions to the call for studies. Thank you very much.

Brian Goesling: So now I think we're going to take questions on either the criteria used in the review process or findings from the first review that Chris presented on so we'll wait a few minutes to give people a chance. We'll wait a minute to give people a chance to type in questions and then we'll start answering.

Okay so we have a few questions coming in. One question received, did you review research on clinic based programs? And the answer is yes we did review research on clinic based programs and there were a few different types of programs we covered in the review.

In some cases these were individualized programs where the intervention was integrated directly into existing clinic services and then we also reviewed studies on some small group training or counseling sessions that were offered as separate add-on services but they were offered in a clinic setting.

Another question we have received, who determined the criteria used to define a program as effective? As I think I mentioned during the presentation, the criteria for evidence of effectiveness were defined by HHS but then the Mathematica team of reviewers applied these criteria during the review process.

And we should mention that several agencies within HHS collaborated on the first review of the evidence including OAH and ASPI, also ACF, the CDC, and also staff from OPA.

Another question we received, why does the HHS list of evidence based programs differ from other lists of effective programs I've seen? I think Chris touched on this question briefly during his presentation.

Each evidence review may use a slightly different set of procedures or criteria both to define the scope of the review and then also to assess study quality or evidence of effectiveness.

There is overlap in these criteria so in general you also will find overlap across lists but there are also differences in the criteria used to both screen the studies and assess study quality. And in some cases these differences can lead to differences in a list of programs identified as being effective.

So Chris gave some examples of the comparison of the HHH list - HHS list to the National Campaigns What Works list or to the CDC's TRS list. You could go through and do a similar exercise for any other lists of evidence based programs that may be available.

Here are two related questions. One is how do I determine if my study was reviewed? And the answer to this is that if you go onto the OAH website -- there is a link on the current slide to the OAH website -- there should be a section of the site entitled Database of Studies Reviewed. And this database includes all the studies that we reviewed as a first - for the first review of the evidence.

There is also a search function that you can use that allows you to search by the title of the article or I think by the study author and then once you have found your study the database will tell you whether or not it met the screening criteria, if it met the screening criteria, the rating it received, and if it received a high or moderate rating. It also indicates whether the intervention met the criteria for evidence of effectiveness.

So a related question to that is if you don't see your listed - study listed on the OAH database what should you do or why was the study not reviewed? In general our team reviewed all the studies we uncovered during our 2009 literature search and the call for studies that was conducted in 2009.

So if your call for studies was not uncovered during that search and is not currently listed on the database you should definitely submit it through the upcoming call for studies for the next round of reviews. Okay I think Chris has a couple of questions that he is going to handle.

Chris Trenholm: Yeah there's a question, in studies that reported multiple outcomes for example contraceptive use, sexual activity, pregnancy or births, did you either prioritize the best outcomes and only use that outcome or did you collect all outcomes and meta analyze by these sets of outcomes to determine effectiveness?

The answer to the question is no. We - following the procedure that Brian outlined, we looked only at a set - we looked individually at each one of those outcomes that was - that Brian discussed, each one of those sexual risk behaviors or sexual risk consequences and assessed the statistical significance of each one of those individually.

There is an additional question, which program showed the most impact in the different outcome measures? This is a very good question. Everybody's question is good but it's a difficult one to answer for a couple of reasons.

One is that these measures themselves are different. It's hard to think about equating an outcome like pregnancy, impact on an outcome like pregnancy versus an impact on an outcome like number of sexual partners for example. So it's difficult to assess across the outcome measures.

And even within the measures there is - there are different ways in measuring them and estimating the magnitude of the effect so it's difficult to put them on a common metric. So I would invite anyone who wants to explore that question to actually could potentially look individually at the studies that are supported that supports the 28 programs and could look at them individually and make that assessment.

Brian Goesling: All right I have a couple of more that have come in here. One question is did you look only at U.S. studies? Yes, when we defined the screening criteria in consult - in consultation with HHS there was decision that the review would be limited to U.S. studies only.

Another question, for the evidence criteria you mentioned there was no requirement for duration of time, for example, short term contraceptive use versus long term pregnancy terms. Is this something that is being looked at possibly for an updated review?

As Barbara mentioned during her section of the presentation, yes one of the plans for updating the review is to revisit evidence standards as the field expands and new research emerges and so I think that yes, this was an

example she gave as something, duration of impacts that might be considered or revisited during a future review.

Another question here, are there any plans to evaluate new programs that have never been evaluated versus reviewing existing studies? So far the review has been focused only on evaluating existing evidence so there's a research study and the team evaluates the existing studies.

So if there is a new program that comes out, if there's a study tied to that new program that results in a journal article or a published research report, we would certainly review that. But we're - we have not been reviewing programs in the absence of any supporting research evidence.

Barbara Broman: Brian this is Barbara Broman, let me just jump in here for a minute and just remind folks on the call or let folks know on the call that, you know, there is a federal evaluation requirement and a local evaluation requirement in the OAH grant as well as other federal evaluation activities that we - that are underway or will be underway to evaluate any new programs, you know, as they proceed.

This includes even the new program that's through the Affordable Care Act, the prep program. So just to let folks know that from the federal perspective we have a commitment to undertaking strong federal evaluation. Anyway, that's it for me.

Chris Trenholm: Just a couple more questions that are coming in and one related to what we were just discussing. Do studies have to appear in peer review journals in order to be included in the review? No they don't. The review also considers studies that have not been through a peer review process.

Another question, why wasn't race ethnicity included as one of the priority subgroups? The subgroup criteria was designed with a focus or concern about multiple pairs since they're multiplied plus this testing. The issue or the concern is that you increase - as you increase the number of subgroups examined in a particular study you also increase the probability of finding a statistically significant impact just by chance.

To address this issue we working with DHHS, Department of Health and Human Services, chose to limit the number of subgroups considered for providing evidence of effectiveness. This is certainly an interesting subgroup to consider. There are potentially many others but ultimately given this concern HHS chose to focus on gender base and sexual experience as the two priority subgroups.

Here is a question coming in, are there any abstinence programs that are currently on the list? The 28 program models that are currently on the list, they represent a range of different approaches and that includes abstinence. It also includes - there is some comprehensive sex ed programs and programs focused more on HIV, STI prevention and also some more youth development approaches.

And if you want more information on the details of the specific programs or what approach they take, I encourage everyone to go to the OAH website if they haven't already.

On that website there is a section of the site which is entitled Programs for Replication and in that section there is a list of the 28 programs and then you can click on the name of each program. There is a link somewhere there that pulls up a one page description of the program and also a summary of the supporting research evidence.

We received a question on whether quasi-experimental studies were included in the review or if we only limited it to randomized control trials. We didn't have any screening criteria on the type of design the study had to use as long as it was a quantitative program impact study, it was considered for review. But different types of designs though were eligible for different study ratings.

So as I mentioned earlier the highest rating was reserved for random assignment studies but quasi-experimental studies were considered for the moderate study rating and that moderate rating was enough to qualify for the list of evidence based programs.

A couple more questions. Regarding behavioral outcome measure, in the new review or the ongoing review will you include outcomes such as reduction in teen dating, violence, self harm, other risk related behaviors that have a negative impact on adolescent health? This is related to other questions we have received like this which are around the theme of whether we'll be using the same or different criteria the next round of reviews as you walk through those four steps and completing the review.

The answer is that we haven't yet started reviewing studies for the next round. Before starting those reviews we'll recommend a set of criteria to HHS for approval. The same recommendation approval process will happen every time. So the decision on that is - will take place in the future.

I think at this point it's unlikely that the number of behavioral outcome measures may expand into those but ultimately this is the decision for the Department of Health and Human Services.

Does the - another question -- does the age criteria of 19 or younger for program participants refer to the age of the participant at the time of the initial intervention or the maximum age of program participant? In other words, were studies reviewed in which a program participant may have turned 20 during the study period?

Brian you can correct me if I'm wrong but my recollection is that it was 19 or younger at the time that the youth enrolled in the study sample or enrolled in the program. So it's possible that if long term follow-up took place when youth were 21 and there was evidence of effects beyond age 19, that would be - that would merit inclusion in the - that would pass the bar.

Another question, what were the two priority subgroups that you focused on your reviews and your evaluation criteria? As mentioned in the review itself, those two priority subgroups were gender and baseline sexual experience so whether or not youth were sexually active at baseline.

Brian Goesling: So we'll - I think we've gotten through most of the questions we have received so far. We'll wait a couple minutes if anyone has any additional questions.

Evelyn Kappeler: Thank you Brian and thank you Chris. At this point Barbara and I have some questions that have come in that we'll respond to. First question here is who was part of the review panel and how long did it take to come up with the list of effective programs? And just to remind folks, that was in the presentation.

The review was conducted by trained staff from Mathematica Policy Research, Child Trends, and Concentric which screened and reviewed the studies. And all total it took about eight months to do the review and complete

the list and then get it published in conjunction with the funding announcement.

We also had another question that Barbara is going to take with regard to the evidence of the review criteria.

Barbara Broman: We had a question about the evidence criteria and the fact that we had mentioned that there was no requirement for duration of time, short term contraceptive use versus long term pregnancy duration, and this is something that is being looked at being updated in the review.

And at this point that kind of an update is possible. We are going to be convening, reaching out to a broader range of experts who will be - who we will be consulting with on the review and we would like to get their input on the types of things that we would look at in terms of updating the review. Here is Evelyn back.

Evelyn Kappeler: And I would just like to clarify for folks, we have this call for studies that is being issued in the next few days for this year's review of any evidence of new program studies. In the longer term we will be convening an expert panel to help provide us some technical assistance in looking at the review criteria that is used as part of this process.

Which leads me to one of the questions that has been submitted and that is if evidence criteria are changed to become more rigorous, will programs that met the lower standard of evidence be removed from the list?

And I just want to remind folks that part of this process is to keep the science base updated on a regular basis and that as we identify new studies who meet the review criteria they will indeed help inform any future program funding

announcements on teen pregnancy prevention but they would have no impact on the currently funded programs. Those programs, we have a commitment to them and they will continue.

At this point I think we have exhausted all of the questions that have come in. I would like to thank all of you for joining us this afternoon. I would like to remind you that this Webinar is being recorded and that a transcript of the Webinar as well as the slides will be posted on the Office of Adolescent Health website.

We will be developing a set of Frequently Asked Questions including questions that come in after the Webinar and those will also be posted on our website.

I want to thank all of you for participating today and we look forward to hearing from you in response to our call for studies. Thank you so much. Bye-bye.

END