Transcript: Best Practices for Access Controls and Disclosure Avoidance Techniques

Slide 1:

Baron Rodriguez: Good afternoon everybody. Welcome to the U.S. Department of Education and PTAC webinar, *Best Practices for Access Controls and Disclosure Avoidance Techniques*. Today's webinar will provide an overview of the guidance documents around data disclosure avoidance, and best practice strategies for protecting personally identifiable information from education records in aggregate records. This webinar will provide suggestions on how to ensure the necessary confidentiality requirements are met, including compliance with the Family Educational Rights and Privacy Act, known as FERPA.

Slide 2:

Baron: With me today is Michael Hawes, Statistical Privacy Advisor at the U.S. Department of Education; I am Baron Rodriguez and I direct the activities at the Privacy Technical Assistance Center.

Slide 3:

Baron: This webinar will be approximately one hour, and we want to remind folks to please enter any questions that you have throughout the presentation. We will be tracking those and try to answer as many as we can at the end of the webinar.

We will be recording this webinar. It will be available on the PTAC website in approximately two weeks. And just a reminder that this webinar is only intended to provide a preview of recently released documents, and will not be a comprehensive read new review of these documents that we will discuss today, so please go to our website at http://ptac.ed.gov and read them in their entirety.

Slide 4:

Baron: So why is ED providing this guidance? Remember the privacy of individual student records is protected under FERPA. To avoid unauthorized disclosure of personally identifiable information from education records, or PII, students' data must be adequately protected at all times. For example, when schools, districts, or states publish reports on student achievement, or share students' data with external researchers, these organizations should apply disclosure avoidance strategies to prevent unauthorized release of information about individual students.

Slide 5:

Baron: What are the guidance documents will be talking about today? As you can see on your screen, we will be talking about a case study we created on minimizing access to personally identifiable information. We will specifically go over best practices for access controls and disclosure avoidance techniques, which we will be talking about during today's webinar.

We also will spend some time giving definitions to the basic terms. As we've traveled around the country, and met with you from various states, it became very clear that a lot of times when we talked about disclosure avoidance, suppression techniques, and other things like that, there wasn't common terminology between state agencies, postsecondary institutions. or districts, so we thought it might be good that we all get on the same page around the definitions. And we also have a frequently asked question around disclosure avoidance.

Slide 6:

Baron: So again why is the Department of ED defining what many of you may already know? There are many interpretations of common disclosure avoidance terminology. We want to provide clarification on the distinction between various disclosure avoidance techniques. So with that we're going to go into questions with Michael Hawes.

Slide 7:

Baron: So, Michael, what is the definition of disclosure and disclosure avoidance?

Michael Hawes: Thank you Baron. Disclosure means to provide access to or permit the release, transfer, or other communication of personally identifiable information by any means. Often disclosure can be authorized such as when a parent or eligible student gives their written consent to share their education records with an authorized party, for example a researcher or health provider. But disclosure can also be unauthorized or accidental. Unauthorized disclosure can happen, for example, as a result of a data breach or data loss. Building on this, disclosure avoidance refers to the efforts made to reduce the risks of disclosure, such as applying statistical methods to protect PII in your aggregate data tables. These safeguards are often referred to as disclosure avoidance, and can take many forms, for example suppression, rounding, recoding, etc., and we're going to be talking about some more of those later.

Baron: Thanks Michael. What legal obligation do educational agencies and institutions have to protect PII in aggregate reports?

Michael: Well, under FERPA, educational agencies and institutions that are reporting or releasing data derived from education records are responsible for protecting any PII in those reports from disclosure. In addition, the reports reporting requirements of the Elementary and Secondary Education Act, ESEA, also require states to not disaggregate data went publicly reporting achievement results if those disaggregated results would reveal PII about an individual student. ESEA also requires states to select and implement strategies that they will use to protect the privacy of individual students in those reports. I do want to make a note of when we're talking about disclosure avoidance that needs to be applied to these reports, this is only when issuing your public reports, the things that you are publishing on your website, sending out to the public and so on. This does not typically apply to the reports you are sending directly to the U.S. Department of Education or to your internal stakeholders.

Baron: Is public reporting for data for small groups, i.e. small cells, the same thing as disclosure?

Michael: Well that's a tough question. Reporting characteristics, or counts for small groups, such as the exact number of students that are in a small racial group, or the percentage of that group that is proficient on an assessment, may not in and of itself constitute a disclosure. However, the smaller the cell size the greater the likelihood that someone might be able to identify a specific individual within that cell, and thus the greater the risk of disclosure. Many statisticians consider a cell size of three to be the

absolute minimum needed to prevent disclosure. The larger minimums like five or 10 are often used to further mitigate that disclosure risk. Many people ask what is an appropriate n size for protecting these small groups? Well there's really no right answer. It all boils down to how much risk of disclosure you're willing to accept. Which kind of takes us to our next slide.

Slide 8:

Michael: It's all about risk. The only completely risk-free method of protecting PII in any public release of data would be to never publish any data at all. Anytime you publish information about individuals in public reports, in aggregate reports, you run the risk that individuals might be re-identifiable in that data. The bigger question isn't "Can I publish it?" but "How much risk am I willing to accept?" Which again brings us to our next slide.

Slide 9:

Baron: So, what standard is used to evaluate disclosure risk, Michael?

Michael: Well, the legal standard for protecting PII from students' educational records is what we call the "reasonable person standard." Namely, can a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, identify a specific individual in the resulting data with reasonable certainty. There is debate as to what constitutes a reasonable person, but generally speaking we mean it to be anyone from the community with a basic knowledge of the enrollment and demographic characteristics of the school.

Slide 10:

Baron: So can a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, identify an individual in the publicly released data with reasonable certainty?

Michael: Right.That is the standard we use, yes.

Slide 11:

Baron: Does the Department of Education require educational agencies and institutions too specific data disclosure avoidance techniques? Say a cell size of six for suppression, for instance?

Michael: The Department does not mandate any particular method, nor does it establish a sufficient threshold for what constitutes sufficient disclosure avoidance. These decisions are typically left up to the individual states and local educational agencies and institutions, and it is up to those institutions to determine what works best with in their specific context. As a general recommendation, in publicly available aggregate reports, whenever possible, data about individual students, for example, proficiency rates presented in cross-tabulated tables, should be combined with data from a sufficient number of other students to disguise the attributes of any specific single individual. When this isn't possible, then data about small numbers of students should probably not be reported and should be suppressed.

Slide 12:

Michael: States have a wide variety of options and techniques at their disposal for protecting PII in there public reporting. But all of these reports can be boiled down to three basic "flavors" of disclosure avoidance: suppression, blurring, and perturbation. On our next slide, we will talk about suppression.

Slide 13:

Michael: Suppression involves removing data to prevent the identification of individuals in small cells, or if those individuals have unique characteristics.

There are several examples of suppression as a technique, the most common are cell suppression or row suppression. It could also involve things when you're dealing with micro data, such as sampling from a broader population. Suppression does have a big impact on the utility of your data. It can result in very little data being available for small populations, in particular, since those are the ones you'd have to be suppressing. And it requires suppression of additional non-sensitive data, also called complementary suppression, in order to further protect those cells that are too small to report.

Is there any residual risk of disclosure when you use suppression methods? Yes, particularly because suppression as a technique can be very difficult to perform correctly, especially if you're dealing with large multi-dimensional tables or hierarchical tables. Applying suppression accurately and effectively through multi-dimensional tables can be very challenging. Also, if any data is available on that population else or on a subset of the population elsewhere, either in related tables or in the community, then the suppressed data might be re-calculatable, and we will show an example of that later.

Slide 14:

Michael: Which brings us to our next method, blurring. Blurring involves reducing the precision of the data that you're reporting or presenting in order to reduce the certainty of identification of a possible individual.

There are various examples of blurring: you can aggregate across groups, you can report data as percentages or percent ranges, you can top and bottom code your data – so for example greater than or equal to 90%, or less than or equal to 10% - or you can round your data to the nearest 10. Blurring also has a sometimes substantial impact on data utility: when blurring is used, the user may not be able to make a very good inferences about small changes in small populations. Also the data utility is also affected in some of these blurring methods, particularly aggregation, it can reduce your ability to perform time series or a cross-sectional analyses across schools, across institutions, or across years.

The residual risk of disclosure with blurring? Generally, it's pretty low but if you're providing row or column totals in your tables, or if they're presented elsewhere, it may be possible to calculate out specific characteristics of specific individuals depending on how the blurring is into implemented. Generally the residual risk of disclosure is fairly low with blaring.

Slide 15:

Michael: The third basic "flavor" of disclosure avoidance is perturbation. Perturbation involves making small changes to the data to prevent identification of individuals, again with those unique or rare characteristics, with any certainty.

Examples of perturbation often include data swapping, where you swap certain values on certain variables across individuals. It can involve introducing noise to your data - a random deviation from a reported value, or it can involve generating synthetic data from your protected information, where you essentially create brand-new data that follows the basic characteristics of the population of the data that you're looking at, but without having any specific records for specific individuals.

What affect does perturbation have on data utility? It can definitely minimize the loss of utility of your data, compared to the other methods, and that's the reason this is a favorite method among statistical agencies, for example. But when you're dealing with program or administrative data, it can be seen as inappropriate at times because it reduces then transparency of the data that is based on, which can have various enforcement and regulatory implications. While it's a favorite message method of statistical agencies you don't often see it in program agencies or administrative agencies because of that transparency effect.

Residual risk of disclosure is typically very low. But, if someone has access to your algorithm or rules, they can possibly use those to essentially reverse engineer the original data. There is some risk but it is also relatively low.

Slide 16:

Baron: Thanks Michael. I'm going to talk a little bit about some of the upcoming guidance. We frequently get the question that's on your screen, specifically will the Department release more in-depth guidance on data disclosure avoidance techniques. The Department of ED and PTAC are in the process of developing some best practices for states to consider when doing public reporting. We're going to talk about some of the draft information that we've come up with so far based on some of the studies we've done.

Slide 17:

Baron: The first thing to remember is that some of the best practices the Department have identified are specific methods on deciding which disclosure avoidance method to apply in which situation. It's important to compare different methods in terms of their effects on the utility of data and the risk of disclosure. The choice of technique will depend on the nature of the data, obviously, and also the level of protection offered by the specific method and the usefulness of the resulting data product, which was essentially what Michael was talking about with the utility of data.

Student-level data: it is important to evaluate whether the proposed release contains any individuals with unique characteristics whose identity can be deduced by the combination of variables in the file. If so, apply disclosure limitation techniques such as data suppression and/or controlled rounding to protect the identity of the students. When you think of subgroups in aggregate public reports or tables, whenever possible combine data about individual students, such as proficiency rates presented in cross-tabulated tables, with data from a sufficient number of other students to disguise the attributes of a single student, such as reporting for results for groups of students combined across grades or years. When this isn't possible, it is a best practice to not published data about small groups of students.

When we say those, remember you still can provide that data back to districts. We're talking about public reports. Be sure to consider what other information about those students may also be available within related tables or in related data releases. The more ways the characteristics of a group of students are sliced and diced, the greater the quantity and precision of information provided, as well,

and the greater the risk of finding some combination of those characteristics that uniquely identifies the specific individuals.

Slide 18

Baron: Some best practices for access controls and disclosure avoidance techniques are to implement in accordance with your state standards. Periodically review your public reports and data tables to ensure that the disclosure avoidance methods you use are implemented correctly, and then consider checking previously released reports and tables to make sure you apply disclosure avoidance techniques effectively.

Slide 19

Baron: Complementary suppression. Most states use are some most states are using some form of suppression. To protect small groups and those with uncommon combinations of characteristics, most states are adopting some form of suppression in their public reports. The challenge with suppression is the suppressed information can often be recalculated by subtracting out the data for the larger reporting groups from the "all student" totals. This vulnerability can be mitigated by suppressing additional non-sensitive data to prevent those values from being calculated. These complementary suppressions serve to protect the re-identification of individuals in the small group. Michael is going to walk you through an example here.

Slide 20

Michael: As we mentioned, suppression of small groups is the most commonly used disclosure avoidance method that we typically see in state reports today, though there is significant variation across states in terms of details and sophistication of how suppression is being applied. As a simple illustration here, let's imagine a school report card that present students' proficiency rates on a particular assessment and is broken down by race and gender. In this example, let's say that the American Indian subgroup has fewer students then the states selected n size. They decided to suppress the data for that subgroup. Often it isn't enough to just protect the PII for that group, and simple math can be used to undo those protections.

Slide 21

Michael: If you add up the numbers of students for the remaining racial categories, you get 85 students. By adding the data reported for gender, on the other hand, we get 86 students. It's clear there is one student in the school who is American Indian.

Slide 22

Michael: If we multiply the reported proficiency rate for each subgroup against the number of students in that group, we can then determine the number of students in each reported group that scored proficient on the assessment.

Slide 23

Michael: Finally, if we add up the numbers proficient for each group, and subtracted from the total proficient reported for the gender categories, we can easily calculate that the one American Indian student in the school was not proficient on the assessment. So how do we protect that information from disclosure?

Slide 24

Michael: By suppressing an additional subgroup, in this case black students, who are the next smallest subgroup in the school, you can prevent re-identification of the protected values for the American Indian subgroup, which was too small to report. This additional suppression is called complementary suppression.

Slide 25:

Baron: Isn't it sufficient to just strip the student ID, name, or Social Security number and get the date out that way? Isn't that de-identification?

Michael: We get a lot of questions on de-identification, and I have to say that one of the driving motivations for producing these guidance documents has been to clarify what de-identification actually means, particularly in the FERPA context.

Many people falsely assume that if you strip off the names, ID numbers, and other direct identifying information from a file that the resulting data has been "de-identified." This is almost always an incorrect assumption and has been the cause of many inadvertent disclosures, both within the education context, but also with healthcare data, commercial data, and so on. FERPA protects the PII from education records, but PII is more than just those identifiers. It also includes any information that alone, or in combination, can be linked to a specific individual. If you've got a series of demographic variables in your file, or course histories, or anything that shows substantial variance across your population, then just removing the identifiers will be typically wholly insufficient to deidentify the data because individuals could still be picked out through one or more combinations of those unusual or unique characteristics. In all of these cases, additional disclosure avoidance, be it suppression, blurring, perturbation, or some combination thereof, will also be needed to effectively deidentify the data.

Slide 26:

Baron: Some of the questions we get frequently at the Department and at PTAC are around what level of access, and who can be provided personally identifiable data, and talking about legitimate need to have access to these data. You can see this question is a question we received at PTAC before.

The best way to illustrate this area is not specifically any FERPA guidance that says you can or can't, but just because something is allowed, or maybe allowed, doesn't mean it's a good idea. I like to give the example of when I was in Oregon, I was taking a look at the folks that had access to student-level data. We actually reduced the number of people who had access significantly, including me, even though I was the director of the data systems there. There wasn't a need for me to have access to that level of data. Just because I could, didn't mean it was a good idea.

The other example I like to give, when you talk about researchers. Many of you get a significant number of requests from PhD students, or outside entities, for data. It's really important to develop a process, either through your data governance program or through some internal policies, on how you're going to give data out for research purposes, and actually doing some sort of evaluation on these research request to determine whether the need for personally identifiable data is truly there. We found through some informal polling with states, that over half of the requests for personally identifiable data by researchers really aren't necessary. They could do the same research with publicly available, deidentified or aggregate information. Michael you have some examples of identifiable data as well, correct?

Michael: Yes, as a former researcher myself, Baron's comment that many of the research request that come in can be taken care of and be sufficiently done with publicly available information that has had disclosure avoidance applied to it.

There are going to be cases in most of your institutions where somebody with a legitimate research interest comes to you, and they might have a legitimate need to have the unsuppressed data for their dissertation or whatever research. If it's a permitted permit a use under one of the standard FERPA exceptions, then you can give them, following the rules, and we have those spelled out in other documents, you can give them access to more identifiable information. Even then, it's really a good practice to not necessarily give them the "keys to the candy shop," so to speak. You can still put some kind of redaction on those files so that you're not giving them all of the most sensitive data that are in those files. Social Security numbers for example: if you have a file in your longitudinal data system that has all of the student demographic information, Social Security numbers, names, and various sensitive information, you probably don't need to be giving Social Security numbers to that researcher. You might consider providing a redacted file, where are you strip out the more sensitive identifiers and replace them with some kind of unique ID. The file is still protected because it hasn't gone through disclosure avoidance, but it's less sensitive, there's less risk of re-identification, there's less risk of misuse from that file. Definitely think about redacting some of the more sensitive identifiers in those cases where you are making access available outside.

Slide 27

Michael: This is a chart from the *Data De-identification: Overview of Basic Terms* document. It highlights some of the distinctions that we've been talking about here. The sensitivity, or risk, associated with any given data set can really be thought of as a spectrum. The dividing line, that horizontal dividing line on the chart, represents that "reasonable person" standard for re-identification that we spoke about before. Everything above that line is protected by FERPA. Everything below the line has gone through disclosure avoidance, in one way or another, and is deemed safe to release to the public. Within those categories, though, above the line and below the line, there is still variation in the risk associated with the file, the risk of misuse, the risk of re-identification, particularly when you're looking at that protected category, as Baron discussed it's really a best practice to consider limiting access to the most sensitive data to as few individuals within your institution as actually need to use it. Then provide, as I was mentioning, either redacted or aggregated information as needed to the rest of your organization and your data users.

Slide 28

Baron: If you have questions that you don't necessarily want to have on the webinar, please do email the PTAC helpdesk. We take those to our FERPA working group and determine one whether it's technical assistance that's needed, or if there is an official answer needed and will work with you to truly

understand the issue, and try to get back the best answer we can. We also on we also offer site visits, and we are going to be releasing several new documents in the area of data disclosure and privacy, as well as some new, updated guidance as it relates to other areas that we've received inquiries from the group.