Summary of extraction process for PCAOB materiality data

We have requested the PCAOB retain the programs to convert the raw data and the dataset used for analysis for six years.

1. Source documents. The source documents for most materiality data are engagement profiles (EPs). The EPs originate with the auditor and the PCAOB inspector reviews the data. The auditor completes an EP and sends the completed EP to the PCAOB before the inspection visit. Materiality data are available for inspected engagements in 2005 through 2015 corresponding roughly to fiscal years 2004-2014. The source documents are mostly Word documents from 2005 inspections years – 2012 inspection years; some are PDF files. Beginning in 2012 many source documents are Excel files. For 2014 inspection years to 2015 inspection years, materiality amounts, materiality bases and materiality percentages are available from an internal PCAOB system, with data entered by the PCAOB inspector into fields in IIS (the internal document management system, where PCAOB inspectors save inspection documents) during an inspection. Both EPs and the IIS system provide materiality data for 2014 and 2015; as explained in Section 4 below, we used the 2014 data-redundancy to create an internal check on our data extraction process.

2. Data conversion to SAS files.

The process to convert data from Word files, PDF files and Excel files varies with the nature of the source document. Preeti Choudhary and a PCAOB research assistant (collectively, "we") did the conversions. The initial focus of the conversions was the tabular presentation of materiality data ("materiality table") included.

Word files. We manually opened each Word file and modified the structure of the materiality table to ensure formatting consistency, for example, merging and unmerging cells or moving text to be in a consistent location in the document.

PDF files. We used PDF Converter Enterprise to convert an EP source document in the form of a PDF document to Word, and reviewed each converted file for consistency and correctness (i.e., proper conversion). If a PDF file was not convertible to Word (for example, scanned documents or older PDF files) we created a Word file with the materiality table and manually entered the data into the appropriate fields.

Conversion to Excel. After materiality tables were formatted consistently, the tables were extracted into Excel using an Excel VBA script by the PCAOB data analyst. This step was not necessary for EPs with materiality tables in Excel.

Import into SAS. All Excel spreadsheets were imported into SAS by the PCAOB data analyst.

3. Data fields of interest for this project. The source documents (the EPs) do not require standardized wording to describe materiality amounts. We found 800 variations in wording for fields that report final materiality, planning materiality, materiality for specific accounts, tolerable error and other fields. The amounts reported in the fields are the dollar values applied for the audit engagement. Preeti reviewed

these 800 variations in wording and manually placed all descriptive wording into four main categories, each of which has several subcategories. This categorization was based on a review of approximately 50% of the Word and PDF documents and discussions with a PCAOB staff member (inspector on rotation in the PCAOB's former Center for Economic Analysis, integrated into the Office of Economic and Risk Analysis 2017). The four materiality fields in **bold** in the fourth (Materiality) category are the focus of analysis for this project. These data are available systematically over our sample period. Many of the unbolded items were disclosed for some years or by some audit teams, that is, not systematically included in EP source documents. The unbolded data items are not part of the materiality analysis in this project and are not part of the dataset retained by the PCAOB for this project.

Posting Thresholds: audit adjustments exceeding the posting threshold are typically disclosed to the audit committee and management

1. balance sheet de minimis
2. balance sheet de minimis-preliminary
3. income statement de minimis
4. income statement de minimis – preliminary
5. de minimis posting threshold
6. de minimis posting threshold- Preliminary

Reclassification: audit adjustments exceeding these values must be corrected

1. balance sheet reclass
2. income statement reclass
3. cash flow reclass
4. comprehensive income reclass
5. reclass

Materiality for a specific account or disclosure: occasionally auditors use (and report in their EPs) specific materiality thresholds for specific accounts, disclosures or locations of the audit

1. specific disclosure
2. local materiality
3. specific statement balance sheet
4. specific statement balance sheet – preliminary
5. specific statement cash flows
6. specific statement income statement
7. specific statement income statement – preliminary
8. local tolerable error (the tolerable error amount at a specific location or subsidiary)

Materiality

1. **final materiality**
2. **planning materiality**
3. planning materiality - preliminary
4. planning materiality – balance sheet

5. planning materiality – income statement
6. tolerable error (synonym for performance materiality)
7. tolerable error – preliminary
8. tolerable error – specific account
9. **materiality base (the financial statement item used to compute materiality)**
10. **materiality percent (the percent multiplied by the base to compute materiality)**

4. Data validation processes for four materiality fields

4a. *Initial check.* Once a raw SAS file was merged with proper firm identifiers and fiscal year ends, Choudhary validated the data by manually comparing every 30th observation (86 of the approximately 2600 observations) with the materiality information in the source (EP) document. Preeti found one instance where data were not extracted for an unknown reason, one instance where data were extracted correctly, but the source document noted additional zeros to be added and one extraction error. A PCAOB data analyst investigated and made corrections. Preeti found several instances in which the materiality percentage or base was present in the EP source document but located (reported) in a field different from the field originally extracted. The PCAOB data analyst applied an algorithm to search for key words and percentages related to materiality calculations to locate values that were reported in other places in the source document, presumably because auditors completed the EP form non-systematically. The data analyst and Preeti decided to use the first of the key words and percentage if several words were noted in the fields. Preeti reviewed missing fields manually and performed manual checks to identify potential errors from applying the algorithm to fill in missing data. Additional details are provided in sections 4d, Materiality Percentage and 4e, Missing Materiality Base.

4b. *Comparison of IIS data and EP extracted data for 2014.* 2014 data are available from both the EP extraction process and the (internal) IIS system. Preeti compared values from these two sources and computed differences. A review of differences indicated that more than half the differences arise from differences in the number of zeros. After the PCAOB data analyst identified and corrected an error in the SAS informat used to convert text strings to a numeric format, the correlation between data from the internal system (IIS) and data from the extraction process for 2014 is 0.97. Preeti determined that 26 of 223 observations for 2014 differ between the extracted data and the internal system data; 18 differences are due to differences in the number of zeros, 6 differences are apparently due to rounding, 2 differences occur because the PCAOB inspector entered materiality for the engagement team, not the overall audit, but the EP data correctly contain information for the overall audit. Preeti examined values in the tails of the distributions to correct errors related to an incorrect number of zeros. The process suggests that computing estimated materiality using the base and percentage can be used to identify data errors related to number of zeros. Retaining the larger of the two values seems to correct nearly all the disagreements.

4c. *Final materiality values.* For the entire sample, Preeti evaluated the audit engagements with the highest final materiality values (the highest 1%). The highest value (E to 16) was caused by an auditor entering multiple numeric values into the final materiality cell (e.g., planning and tolerable error). The first (table formatting) step of the extraction process should have corrected this problem but did not.

The next 10 companies with the highest final materiality values tied to the engagement profile (EP) data on the source documents, so Preeti did not continue checking high values. Preeti also checked final materiality values below 100,000 as these were likely to have the incorrect number of zeros. She checked 94 observations and identified (and made) the following corrections: 1 observation was a duplicate EP file with a different name which is a data source error (deleted), 9 files had a processing error that led to duplication of data (the same value for two years within a single EP),[1] and 78 had an incorrect number of zeros.

4d. *Materiality percentage.* Preeti compared the materiality percentage field amounts reported for 2014 in the IIS system with the extracted data from the source documents (EPs). The means of the differences in the percentages from the engagement profile (EP) and IIS data source are near zero, such that information appears to be recorded consistently on average. There are 1,399 observations with materiality percentage data from EP, 857 observations with data from IIS and 245 observations with overlapping data sources, of which 33 observations had differences in materiality percentage values. Twenty-four of the differences observations are attributable to differing numbers of zeros and 3 are small differences (i.e., possibly rounding). After correction by taking the higher of the two percentage values when the higher value is not greater than 10 or by manually correcting mis-entered data, the correlation between the two data sources is 0.90. Errors are identified and corrected manually by reading both the notes describing materiality calculations in the original EP source document and the IIS notes completed by the auditor or inspector.

Preeti checked for errors in materiality percentages by evaluating percentages outside the normal range for each materiality base. The normal range is determined by the distribution of percentages reported in the manuscript and is described below. Some apparent errors arise from misplaced decimal points or from an incorrect materiality base, identified using the notes from the EP or IIS about materiality calculations.

1. Percentage applied to net income or pretax income less than 3% or more than 12% - 85 corrections made
2. Percentage applied to asset materiality base less than 0.2% or more than 3% - 25 corrections made
3. Percentage applied to equity base less than 0.2% or more than 3% - 2 corrections made
4. Percentage applied to gross profit base less than 1% or more than 2% - 2 corrections made

[1] We checked records for identical values of final materiality, planning materiality and tolerable error across two time periods (the current and the prior year) within the same source document to identify processing errors. It seems unlikely that two audits of the same client would have identical values for these fields. Upon investigation we identified a processing error. In cases with (apparently) missing materiality data, the data analyst found the code read from an existing temporary dataset containing the materiality data. The fix is to delete this temporary dataset at the end of each loop when reading in the file. This fix was implemented and the data were reprocessed. Preeti compared the statistics from before and after the reloading to ensure that data are consistent.

Preeti reviewed 607 cases with missing materiality percentages and sought information in IIS fields and alternative fields, including a review of notes about materiality. In 145 cases she found the percentage in the notes about materiality; in 36 cases she found the materiality base was updated from the word search described in section 4a.

4e. *Missing materiality base*. Preeti compared materiality base information reported for 2014 in the IIS system with the extracted data from the source documents (EPs). 163 observations in the extracted data were missing base values. Preeti replaced 115 missing materiality bases in data extracted from EP source documents with data from the internal system (IIS). Preeti confirmed there is no information about materiality base for the remaining 48 cases with one exception. Preeti replaced cons_quant_mat_base = "No base provided" as a single description for cases with missing materiality base information. Preeti reviewed the 34 cases that stated "other" as the base and was able to fill in bases for 28 cases.

4f. *Planning materiality.* For the 2004-2014 sample, Preeti checked for data errors in planning materiality by scaling planning materiality by final materiality. She evaluated scaled values exceeding 1.5 or less than 0.5, by comparing the extracted data (from EP source documents) with notes on materiality from EP source documents. 51 scaled values fell outside the range specified. Preeti obtained data on final materiality, planning materiality, percent used and base from source documents and identified 38 corrections.

4g. *Prior year final materiality.* For the 2004-2014 sample, Preeti checked for errors in prior year final materiality by scaling prior year materiality by current year materiality and investigating records with scaled values greater than 2 or less than 0.2. She investigated 66 records and corrected 25 records.

5. <u>Summary of data sources and adjustments from processes described in preceding sections</u>

Items 1 through 4 refer to 2014 data and items 5 through 8 refer to all sample years.

1. We replaced missing final materiality (tolerable error) extracted from source EP document with materiality (tolerable error) value from IIS data. (1 (1) change)
2. If IIS final materiality (tolerable error) values exceed EP values, replaced materiality (tolerable error) values with IIS data to correct zeros problem. (13 (18) changes)
3. If IIS materiality percentage is less than 10 but greater than materiality percentage from EP, replaced EP materiality percentage with IIS materiality percentage. (5 changes)
4. If EP materiality percentage exceeds 10 and is not missing and IIS materiality percentage from IIS is not missing and greater than zero, replaced EP materiality percentage with IIS materiality percentage. (4 changes)
5. If EP materiality percentage is missing, replaced materiality percentage with IIS materiality percentage or information noted in the EP (145 cases)
6. For all years in the sample, if cons_quant_mat_base is missing (No base provided, not described in other, see materiality memo/ audit work paper) replaced with IIS data where available (116 replacements of which 1 was unknown and coded as missing)

7. If cons_quant_mat_base is Pre-tax or net income (not specified) and IIS data exist, replace with IIS data. (38 replacements)
8. For all years in the sample, replaced missing values of planning materiality and final materiality with "." if they were zeros. (3 changes)

6. 2015 data

We received access to 2015 inspection information after we had collected data for 2004-2014. We obtained 2015 materiality data from the IIS system; we used EP data as a supplementary information source to review data for missing or incomplete data in the IIS system. Preeti removed observations that do not pertain to inspected audit engagements conducted by global national audit firms (GNFs) for consistency with prior-years' data and removed duplicate observations and observations missing CIK identifiers. The 2015 IIS data include materiality calculation explanations, final materiality values, tolerable error, posting thresholds and issuer identifiers. Because inspectors might enter data for both the overall engagement and the engagement team they are inspecting, Preeti manually reviewed the 43 engagements with multiple observations in the IIS system to determine which materiality values represent the overall audit. Preeti checked for errors in the number of zeros by searching for final materiality values less than 100,000; she found 26 such observations and corrected 9 observations by reviewing the EP data. Preeti reviewed the materiality calculation explanations for each observation to confirm the materiality base and the percentage applied. If either the base or the percentage was missing from the IIS system and sufficient notes were not provided in the calculation explanation, Preeti referred to the EP data to fill in missing information. Finally, Preeti filled in the prior year materiality and tolerable error for each observation from the EP data to ensure consistency with prior-years' data. During her review of 2015 source documents, if Preeti found differences between the IIS source and the EP source for final materiality values, Preeti made corrections using a process similar to the process described in section 4b.

7. Categorizing materiality base data

IIS materiality base data are categorized in the IIS system because the IIS system uses drop boxes to enter this data. When IIS data are not available, Preeti searched the EP source documents for materiality descriptions using a key word search for the first instance of words similar to net income, pretax income, assets, revenue, equity, cash flow and permutations of these key words in two fields where auditors might have recorded materiality descriptions. The materiality base "net assets" is categorized as equity. Because auditors may have described in words the materiality calculation in different places, the data checks previously described may result in corrections or updates to the key word searches. Preeti grouped similar materiality bases as follows:

1. Set materiality base "Cash Flow - Operations" as "CFO" (4 changes)
2. Set materiality base "Annualized Net Income" as "normalized net income" (1 change)
3. Set materiality base Gross Margin to Gross Profit (1 change)
4. Set materiality base "N/A" or "see memo/workpaper" to "No base provided) (10 changes)
5. Set materiality base operating income to Adjusted NIBT (1 change)

8. Final materiality checks

8a. *Comparison of reported materiality amounts to estimates obtained from Compustat data and auditor-reported materiality percentages*

Preeti obtained financial statement subtotal data from Compustat to compute line items used as materiality bases (e.g., EBITDA, normalized pretax income as average of past 3 years of pretax income). The PCAOB data analyst merged the Compustat data with the corrected materiality data for all years in the sample. Preeti calculated estimated materiality by multiplying the materiality percentage provided by the auditor times each relevant base obtained from Compustat. If the percentage was missing, Preeti used the mean percentage value to estimate materiality (0.82% for assets, 2.2% for EBITDA, 1% for equity, 1.2% for gross profit, 5% for pretax income and net income, 5.5% for normalized pretax income, 5.1% for normalized net income, 0.75% for normalized revenue). Preeti computed the following difference values: (final materiality from PCAOB data – estimated materiality using PCAOB percentage and Compustat data) /final materiality. Preeti investigated values exceeding +/- 75% of final materiality, a total of 446 observations. These checks resulted in:

1. 4 corrections to cik
2. 3 corrections to fiscal year end
3. 22 corrections to final materiality values
4. 25 corrections to planning materiality
5. 27 changes to tolerable error (many tolerable error values are not reported; we do not use tolerable error data in our analysis)
6. 11, 8 and 12 changes to prior year materiality, planning materiality and tolerable error, respectively
7. 276 corrections to materiality percentage. Many of these amounts are from 2015, which was not subject to review for percentage errors or missing values that Preeti filled in for other years.
8. 234 corrections to materiality base. Some of these corrections are minor, such as adding the modifier normalized or adjusted and some involved filling in missing and updated amounts with values obtained from comparing IIS and EP documents for 2014 and 2015, when both are available.

To summarize, based on a review of 446 of 2853 observations (approximately 15% of the sample) with the largest differences between materiality values extracted and estimated materiality, the error rate in values of most importance to our study, CIK, FYE, and final materiality values, was 6.5%.

8b. *Manual review of final materiality amounts*

To ensure that updates to final materiality based on the corrections made using the process described in section 8a flow through to tolerable error[2] and planning materiality values and to ensure the accuracy of

[2] Tolerable error data are not analyzed in this project.

all 2015 data which were hand-reviewed, Preeti performed additional manual review of 83 observations identified using the following five checks:

1. planning materiality/final materiality values less than 0.5 or greater than 2
2. final and planning materiality and tolerable error values less than 100,000
3. prior year tolerable error/prior year final materiality greater than 0.9 or less than 0.4
4. materiality percentages less than 0.05 or greater than 12
5. difference between Compustat-based materiality estimate and final materiality reported in EP source document or IIS system greater than 2 or less than -2

83 observations were identified by these five checks and 36 observations were corrected after reviewing the EP source documents and IIS if available:

1. 2 updates to correct cik
2. 1 update to final materiality
3. 16 updates to tolerable error (some were filled in using EP data because they were missing; we do not use tolerable error data in our analysis)
4. 6 updates to planning materiality
5. 2 updates to prior year final materiality
6. 7 updates to prior year tolerable error (we do not use tolerable error data in our analysis)
7. 1 update to prior year planning materiality
8. 3 updates to percentages
9. 10 updates to materiality base

8c. *Review of randomly selected observations*

Preeti used stata to randomly select 56 observations (2% of total observations) with both a Compustat link and materiality data. For these observations, Preeti tied the following values: CIK, FYE, inspection year, audit firm, percentage, and materiality base and current/prior year: final materiality, tolerable error and planning materiality to amounts in EP source documents to confirm the accuracy of cleaned data. The only error identified pertains to PwC from 2008-2011 where planning materiality and tolerable error values were sometimes swapped. This occurred because PwC used "planning materiality" to mean tolerable error or performance materiality before AS 11 standardized the terminology. To correct these values Preeti reviewed 120 observations for PwC from 2008-2011 when tolerable error was missing (meaning this error was not previously corrected) and when planning materiality was less than final materiality (it is only these cases where planning materiality could be mis-recorded as tolerable error because tolerable error cannot exceed final materiality). Most error corrections involved swapping planning materiality and tolerable error and entering planning materiality when available.

8d. *Errors corrected in merging data.* In merging materiality data with Compustat data Preeti identified 41 duplicate observations, ten of which had the same values for all fields of interest; the five duplicates were removed. The remaining duplicate observations were investigated. Preeti found 17 observations (one of which had two duplicates) where the numerical fields were the same, but the materiality base

and or percentage was filled in for only one of the observations; the more complete observation was retained. Four observations were duplicated because two had incorrect CIKs; these were corrected. The remaining observations were duplicates because both a parent and a subsidiary were inspected and noted in separate engagement profile documents. In these cases the parent data were retained.