Using Structural Equation Modeling to Construct Calibration Equations Relating PM_{2.5} Mass Concentration Samplers to the Federal Reference Method Sampler

Richard A. Bilonick^{a,c,1,*}, Daniel P. Connell^b, Evelyn O. Talbott^c, Judith R. Rager^c, Tao Xue^c

^a*University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA*

^b*CONSOL Energy Inc., Research and Development, Pittsburgh, Pennsylvania, USA*

^c*University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA*

Abstract

The objective of this study was to remove systematic bias among fine particulate matter (PM_{2.5}) mass concentration measurements made by different types of samplers used in the Pittsburgh Aerosol Research and Inhalation Epidemiology Study (PARIES). PARIES is a retrospective epidemiology study that aims to provide a comprehensive analysis of the associations between air quality and human health effects in the Pittsburgh, Pennsylvania, region from 1999 to 2008. Calibration was needed in order to minimize the amount of systematic error in PM_{2.5} exposure estimation as a result of including data from 97 different PM_{2.5} samplers at 47 monitoring sites. Ordinary regression often has been used for calibrating air quality measurements from

*Corresponding Author

Email address: rab45@pitt.edu (Richard A. Bilonick)

¹Tel.: +1 412 647 5756 Fax: +1 412 647 5119

pairs of measurement devices; however, this is only appropriate when one of the two devices (the "independent" variable) is free from random error, which is rarely the case. A group of methods known as "errors-in-variables" (e.g., Deming regression, reduced major axis regression) has been developed to handle calibration between two devices when both are subject to random error, but these methods require information on the relative sizes of the random errors for each device, which typically cannot be obtained from the observed data. When data from more than two devices (or repeats of the same device) are available, the additional information is not used to inform the calibration. A more general approach that often has been overlooked is the use of a measurement error structural equation model (SEM) that allows the simultaneous comparison of three or more devices (or repeats). The theoretical underpinnings of all of these approaches to calibration are described, and the pros and cons of each are discussed. In particular, it is shown that both ordinary regression (when used for calibration) and Deming regression are particular examples of SEMs but with substantial deficiencies. To illustrate the use of SEMs, the 7,865 daily average $\text{PM}_{2.5}$ mass concentration measurements made by seven collocated samplers at an urban monitoring site in Pittsburgh, Pennsylvania, were used. These samplers, which included three federal reference method (FRM) samplers, three speciation samplers, and a tapered element oscillating microbalance (TEOM), operated at various times during the 10-year PARIES study period. Because TEOM measurements are known to depend on temperature, the constructed

SEM provided calibration equations relating the TEOM to the FRM and speciation samplers as a function of ambient temperature. It was shown that TEOM imprecision and TEOM bias (relative to the FRM) both decreased as temperature increased. It also was shown that the temperature dependency for bias was non-linear and followed a sigmoidal (logistic) pattern. The speciation samplers exhibited only small bias relative to the FRM samplers, although the FRM samplers were shown to be substantially more precise than both the TEOM and the speciation samplers. Comparison of the SEM results to pairwise simple linear regression results showed that the regression results can differ substantially from the correctly-derived calibration equations, especially if the less-precise device is used as the independent variable in the regression.

Keywords: Calibration, PM_{2.5}, Federal Reference Method, TEOM

1. Introduction

Several different methods are commonly used to measure ambient mass concentrations of fine particulate matter (PM_{2.5}), and these methods are known to exhibit systematic biases relative to one another. For example, it is well documented that PM_{2.5} mass concentrations measured using a tapered element oscillating microbalance (TEOM) can be biased low relative to those measured using a federal reference method (FRM) monitor or PM_{2.5} speciation sampler because of the loss of semi-volatile material, including ammonium nitrate and semi-volatile organic compounds, from the TEOM (Allen

10 et al. [1], Chung et al. [2], Eatough et al. [3], Cyrus et al. [4], Tortajada-
11 Genaro and Borrs [5]). The magnitude of this bias is dependent upon the
12 chemical composition of the $\text{PM}_{2.5}$, the TEOM operating conditions, and
13 ambient conditions, and therefore is site- and monitor-specific. Hence, for
14 air quality monitoring applications or scientific studies (e.g., air pollution
15 epidemiology studies) that must incorporate or compare data from multiple
16 types of $\text{PM}_{2.5}$ measurement devices, it is necessary to first develop calibra-
17 tion equations to remove systematic biases so that the data can be used
18 interchangeably.

19 Ordinary linear regression analysis commonly has been employed to de-
20 velop calibration equations relating $\text{PM}_{2.5}$ measurement devices. Ordinary
21 regression analysis will provide a correct description of the relationship be-
22 tween two devices when it is applied to data spanning a wide range of the
23 quantity to be measured, when one of the devices is a highly precise refer-
24 ence measurement (that is, one with little or no random error), and when
25 this highly precise reference is used as the "independent" variable in the re-
26 gression. In the case of calibrating $\text{PM}_{2.5}$ samplers, though, as with many
27 applications, both devices are subject to substantial random error, even if
28 one is termed a "reference" method. In these cases, ordinary regression pro-
29 duces two contradictory curves, depending on which device is used as the
30 independent variable in the model, both of which are biased away from the
31 correct calibration curve (Bollen [6]). Nevertheless, numerous $\text{PM}_{2.5}$ or PM_{10}
32 method comparison studies (e.g., Allen et al. [1], Ayers et al. [7], Chung et al.

[2], Charron et al. [8], Schwab et al. [9], Zhu et al. [10]) have been based on ordinary regression, and a number of studies (e.g., Bortnick et al. [11], Vega et al. [12], Kashuba and Scheff [13]) comparing TEOM and filter-based (e.g., FRM) measurements have used the less-precise TEOM as the independent variable, exacerbating the problem. In fact, the U.S. Environmental Protection Agency (EPA) guidelines (Eberly et al. [14]) for relating FRM and continuous $\text{PM}_{2.5}$ measurements specify the use of ordinary linear regression with the presumably less-precise continuous measurement device used as the independent variable.

In the late nineteenth century, Adcock [15], aware of the problems with using ordinary regression to calibrate two types of measurement that are both affected by random error, described a method that today is often referred to as Deming regression. Deming regression (Kummell [16]; Deming [17]) is a particular example of an errors-in-variables approach often applied when there are only two devices being compared, or applied in a pairwise fashion when there are more devices and/or repeats available. Several $\text{PM}_{2.5}$ method comparison studies (e.g., Hauck et al. [18], Lee et al. [19], Hains et al. [20]) have used this approach. In practice, however, Deming regression can be impractical to use for two reasons. Firstly, Deming regression assumes that all the error is random error and that the measurement scales are the same. Secondly, it assumes that the ratio of the imprecision standard deviations (that is, the random errors for each method) is known. Even if this ratio were known, it would be impossible to know whether the measurement scales

are the same based only on the paired measurements. Pairwise application of Deming regression and other error-in-variables methods will also fail to incorporate all the available information when measurements from more than two devices are available.

Another errors-in-variable approach that has been used is reduced major axis (RMA) regression (Kermack and Haldane [21]), which is sometimes referred to as orthogonal regression. Ayers [22] in a technical note suggested the use of RMA regression in place of ordinary regression for air quality data, and Charron et al. [8] and Green and Fuller [23] applied RMA regression to compare PM samplers. However, when the ratio of the device imprecision standard deviations is equal to one, Deming regression reduces to RMA regression. Thus, RMA regression suffers from the same defects as Deming regression.

Two issues should be further noted. Firstly, of the PM measurement comparison studies cited above, only Bortnick et al. [11] transformed the PM measurements. Transformations are often required to help insure that the measurement errors are normally distributed and the relationship is linear, as assumed by most methods of analysis. Secondly, almost all of the papers used correlation coefficients as a measure of agreement, regardless of whether they used ordinary regression or not. In an ordinary regression with a single independent variable, the correlation coefficient is merely a linear transformation of the regression slope and vice versa. Importantly, the (Pearson product-moment) correlation coefficient is only a measure of linear associa-

tion and does not measure agreement between devices. For example, there could be a perfect correlation between temperatures measured in Fahrenheit degrees and Celsius degrees yet, in general, there would be no agreement (in the sense that you could not substitute one for the other). Two devices can be perfectly correlated yet have no agreement and, conversely, can have a correlation of zero and yet be in good agreement. The latter can occur when the quantity being measured has a very narrow range. The correlation coefficient is highly dependent on the range of measurements - measurements made by two devices will tend to be highly correlated if the range of the true values is very large compared to the typical size of the random error in each measurement, regardless of the amount and type of bias.

The work presented here was performed as part of a larger effort to calibrate fine particulate matter ($\text{PM}_{2.5}$) mass concentration measurements made by 97 samplers (including more than ten different types of samplers) at 47 monitoring sites used in the Pittsburgh Aerosol Research and Inhalation Epidemiology Study (PARIES). PARIES is a retrospective epidemiology study that aims to provide a comprehensive analysis of the associations between air quality and human health effects in the Pittsburgh, Pennsylvania, region from 1999 to 2008. In this paper, we focus on calibrating measurements from seven samplers, including three $\text{PM}_{2.5}$ FRM samplers, three $\text{PM}_{2.5}$ speciation samplers, and one TEOM sampler, that were collocated at a single site in the Lawrenceville neighborhood of the city of Pittsburgh. This site was chosen because it had the largest number of collocated samplers and

102 included a TEOM sampler. (A second paper will describe how this analysis
 103 was expanded to all 47 monitoring sites). As a result of investigating various
 104 potential approaches, a structural equation model (SEM), specifically con-
 105 sisting of a set of linked sub-models, is proposed to appropriately describe
 106 the measurement errors associated with these various samplers. Structural
 107 equation modeling has a long history but has been primarily used in psy-
 108 chology, genetics, and economics to accommodate complicated measurement
 109 error problems. As will be illustrated, both errors-in-variables regression and
 110 ordinary regression are specific, relatively simple, examples of SEMs. We in-
 111 tend to show that SEMs are ideally suited to the problem of PM_{2.5} mass
 112 concentration sampler calibration and similar calibration problems. We first
 113 describe the classic measurement error model and the basic SEM methodol-
 114 ogy and then describe ordinary regression as a SEM. Next, we discuss the
 115 problems inherent in the use of ordinary regression for calibration and the
 116 general impracticality of using errors-in-variables regression in most cases.
 117 We develop a comprehensive SEM that describes the relationships among
 118 the sampling device measurements and in particular handles TEOM calibra-
 119 tion appropriately as a nonlinear function of temperature.

120 **2. General Approach to Calibration**

121 *2.1. Classic Measurement Error Model*

122 A measurement error model describes how the quantity that we attempt
 123 to measure is distorted by the measurement process (Jaech [24]). The path

124 diagram for a one-factor measurement error model for three devices is illus-
 125 trated in Figure 1a. The sources of distortion involve both systematic error
 126 (bias, lack of accuracy, lack of trueness) and random error (imprecision, lack
 127 of repeatability, lack of reliability). Formally, each measurement Y_{ij} from
 128 device i for item j is described as:

$$Y_{ij} = \alpha_i + \beta_i \mu_j + \epsilon_{ij} \quad (1)$$

129 where α_i and β_i jointly describe the bias for device i , μ_j is the true value
 130 for item j that we attempt to measure, and ϵ_{ij} is the random error which
 131 is usually assumed to be normally distributed with mean zero and standard
 132 deviation σ_i . The true values μ have standard deviation σ and can be treated
 133 as either fixed or random values depending on the application. This model
 134 would be an ordinary regression model except that each μ_j is in practice
 135 forever unknown, making the application of ordinary regression impossible.
 136 The slope bias parameter β characterizes the measurement scale - the size of
 137 the measurement unit - for device i relative to the true scale. The intercept
 138 bias parameter α is an arbitrary but necessary reference point and is difficult,
 139 in general, to interpret in isolation from the scale bias β . When, however,
 140 $\beta_i/\beta_{i'} = 1$, the intercept represents a constant bias between devices i and i'
 141 equal to $\alpha_i - \alpha_{i'}$ - that is, a bias that is independent of μ . When $\beta_i/\beta_{i'} \neq 1$,
 142 the bias depends on μ and the measurement units differ.

143 The true values represent the assumption at the heart of any calibra-

Figure 1: Path diagrams comparing a) measurement error model for measurements Y_1 , Y_2 , and Y_3 from three devices 1, 2, and 3, and b) regression models for predicting measurements from device 1 (Y_1) from measurements from device 2 (Y_2), and vice versa. A circle indicates a latent variable and a square indicates a measured (observed) variable. Responses are denoted by the color yellow while independent variables are indicated by the color blue. The measurement error model treats the measurements as responses while the regression models must treat one of the measurements as an independent variable that has no random error.

tion/measurement error analysis. Each instrument must attempt to measure the same quantity at the same time and same location. If the instruments are attempting to measure different quantities there can be no basis for calibration. For air quality studies, measurements are made on small parcels of air. Ideally, the same parcel of air would be analyzed by all of the instruments at the same exact point in time. However, attempting to analyze the same air parcel at the same time could lead to interactions between the instruments and changes to the air parcel, so in practice slightly different air parcels are sampled with the goal of having these air parcels be as similar as possible. This usually means taking air parcels that are physically very close together and making measurements at the same instant of time. To the extent that the air parcels have different compositions, the resulting error will be misidentified as measurement error. Because the collocated samplers typically are located within a few meters of each other, the error due to misidentification is assumed to be negligible. One other point should be emphasized. Although μ is a parameter in the model, there are no unique estimates for the true values, and all determinations of bias are relative. Only highly precise reference samples, which are accurate based on theoretical understanding, can provide information on absolute accuracy.

For three samplers $i = 1, 2$, and 3 (simplifying by ignoring j), measure-

164 ments Y_1 , Y_2 , and Y_3 are described as:

$$\begin{aligned} Y_1 &= \alpha_1 + \beta_1\mu + \epsilon_1 \\ Y_2 &= \alpha_2 + \beta_2\mu + \epsilon_2 \\ Y_3 &= \alpha_3 + \beta_3\mu + \epsilon_3 \end{aligned} \tag{2}$$

165 with ϵ_1 having standard deviation σ_1 , ϵ_2 having standard deviation σ_2 , and ϵ_3
 166 having standard deviation σ_3 . A scale must be provided for the unobserved
 167 μ . This can be accomplished either by fixing σ to 1 and the mean of μ , $\bar{\mu}$,
 168 to zero or by arbitrarily selecting the i^{th} device and setting β_i to one and
 169 α_i to zero. Arbitrarily setting the scale for μ to equal the scale of one of
 170 the devices is typically more convenient. The fact that μ is unknown is of
 171 little concern as the real interest is the relationship between any two devices
 172 i and i' . This can be derived easily because the two equations both contain
 173 μ , yielding the unique relationship between Y_i and $Y_{i'}$:

$$Y_i = \left[\alpha_i - \frac{\beta_i}{\beta_{i'}} \alpha_{i'} \right] + \frac{\beta_i}{\beta_{i'}} Y_{i'} + \left[\epsilon_i - \frac{\beta_i}{\beta_{i'}} \epsilon_{i'} \right] \tag{3}$$

174 or equivalently, rewritten for $Y_{i'}$ in terms of Y_i :

$$Y_{i'} = \left[\alpha_{i'} - \frac{\beta_{i'}}{\beta_i} \alpha_i \right] + \frac{\beta_{i'}}{\beta_i} Y_i + \left[\epsilon_{i'} - \frac{\beta_{i'}}{\beta_i} \epsilon_i \right] \tag{4}$$

175 with

$$\epsilon_i - \frac{\beta_i}{\beta_{i'}} \epsilon_{i'} \sim N \left(0, \sigma_i^2 + \frac{\beta_i^2}{\beta_{i'}^2} \sigma_{i'}^2 \right) \tag{5}$$

176 and

$$\epsilon_{i'} - \frac{\beta_{i'}}{\beta_i} \epsilon_i \sim N \left(0, \sigma_{i'}^2 + \frac{\beta_{i'}^2}{\beta_i^2} \sigma_i^2 \right) \quad (6)$$

177 Thus, in principle, because the random errors vanish via expectation, it
178 is possible to remove bias between devices - that is, to calibrate one device in
179 terms of the other. When $\beta_i/\beta_{i'} \neq 1$, the devices have different measurement
180 scales, and this is referred to as a scale bias or equivalently as a non-constant
181 bias, because the bias is a function of the true quantity μ . When $\beta_i/\beta_{i'} = 1$
182 and $\alpha_i - \alpha_{i'} \neq 0$, there is a constant bias. As noted by Dunn [25], the
183 random error is gauged by comparing the imprecision standard deviations
184 adjusted for differences in scale, namely: σ_i/β_i and $\sigma_{i'}/\beta_{i'}$. Bias can be
185 removed by calibration, but differences in imprecision can only be diminished
186 by averaging two or more independent repeats.

187 Although the values for β_i and $\beta_{i'}$ are dependent on the choice of a referent
188 device, the ratio $\beta_i/\beta_{i'}$ is invariant to this choice. Similarly, the difference $\alpha_i -$
189 $\frac{\beta_i}{\beta_{i'}} \alpha_{i'}$ is invariant, but note that this difference can only be easily interpreted
190 when $\beta_i/\beta_{i'} = 1$. Given that either $\sigma \equiv 1$ and $\bar{\mu} \equiv 0$, or for one arbitrarily
191 selected device i , $\beta_i \equiv 1$ and $\alpha_i \equiv 0$, then for $m \geq 1$ devices, there are $3m$
192 parameters that need to be estimated. From the sample data, there would be
193 $\frac{m(m+3)}{2}$ sample variances, covariances, and means. (Under the assumption of
194 normally distributed true values and errors, these statistics would represent
195 the "sufficient" statistics.)

196 This model can be easily estimated using either the Method of Moments

197 or the Method of Maximum Likelihood, provided there are at least $m \geq 3$
 198 devices. Typically, maximum likelihood estimates would be preferred, espe-
 199 cially when there are missing values, using full information maximum like-
 200 lihood. In the Method of Moments estimation, the sample moments are
 201 equated to the theoretical moments (which are derived from the model).
 202 With three devices there would be nine equations and nine unknowns. Solv-
 203 ing for the nine unknowns provides the parameter estimates.

204 *2.2. Structural Equation Model*

205 The measurement error model for three samplers is an example of a simple
 206 structural equation model (SEM) which has a path diagram shown in Fig-
 207 ure 1a (Bollen [6]; Kline [26]; Marcoulides and Moustaki [27]). When there
 208 are fewer than three devices, the measurement error model has no unique
 209 solutions, and the SEM is said to be non-identifiable. It can easily be seen
 210 that when there are only two samplers, if one assumes that the ratio of the
 211 device random error standard deviations $\frac{\sigma_1}{\sigma_2}$ is known, then this effectively
 212 eliminates one parameter and a unique solution can be found. Some errors-
 213 in-variables models for the comparison of exactly two devices depend on this
 214 constraint to achieve a unique solution; however, a unique solution also can
 215 be achieved by making other assumptions. For example, one could assume
 216 that $\beta_1 = \beta_2$, or more generally that $\beta_1 = k\beta_2$ with $k \geq 0$. Arbitrary assump-
 217 tions should be avoided and all parameters estimated from the available data
 218 from an appropriately designed study. If sufficient information is not avail-

able within the measurement error study, then the random error standard deviations could be estimated via an external study in which repeat measurements are made with both devices under identical conditions. To the extent that the externally estimated random error parameters differ from what they should be for the calibration study, this will distort estimation of the bias parameters.

2.3. Regression as a SEM

Ordinary regression is one of the simplest examples of a structural equation model, consisting of only a single equation, but it requires declaring one of the measurements a dependent (response) variable and the other an independent (explanatory) variable. If Y_1 is (arbitrarily) chosen as the independent variable and Y_2 is chosen as the dependent variable, then the path diagram is shown in the bottom of Figure 1b and:

$$Y_2 = \alpha_2^* + \beta_2^* Y_1 + \epsilon_2^* \quad (7)$$

232

$$\epsilon_2^* \sim N(0, \sigma_2^{*2}) \quad (8)$$

where the asterisks emphasize that these parameters are not necessarily the same (and in fact, generally are different) than in the classical measurement error model described previously. Compare equation 7 with the calibration

236 equation (derived from Equation 3 with $i' = 1$, $i = 2$, $\alpha_1 = 0$, and $\beta_1 = 1$):

$$Y_2 = \alpha_2 + \beta_2 Y_1 + [\epsilon_2 - \beta_2 \epsilon_1] \quad (9)$$

237 It is clear that the error terms differ and the calibration equation has two
238 sources of error rather than the one source of error for regression. If instead,
239 Y_2 is chosen as the independent variable and Y_1 is chosen as the dependent
240 variable, then the path diagram is shown in the top of Figure 1b and results
241 in an entirely different equation:

$$Y_1 = \alpha_1^* + \beta_1^* Y_2 + \epsilon_1^* \quad (10)$$

242

$$\epsilon_1^* \sim N(0, \sigma_1^{*2}) \quad (11)$$

243 Compare equation 10 with the calibration equation written for Y_1 as a func-
244 tion of Y_2 :

$$Y_1 = -\frac{\alpha_2}{\beta_2} + \frac{1}{\beta_2} Y_2 + \left[\epsilon_1 - \frac{\epsilon_2}{\beta_2} \right] \quad (12)$$

245 Note that, unlike the two distinct regression equations, equation 9 and equa-
246 tion 12 are re-expressions of the same calibration equation. Note also that
247 the slopes for the two ways to express the calibration are reciprocals of each
248 other – this is not generally the case for the two different regression equations.

249 2.4. Why and How Ordinary Regression Fails for Calibration

250 A fundamental assumption of ordinary regression is that the indepen-
 251 dent variable is free from random measurement error. The regression path
 252 diagrams shown in Figure 1b emphasize that only the response variable is
 253 subject to random error – the independent variable has no random error.
 254 When this is not the case, as mentioned previously, it is well-known that the
 255 regression slope is biased toward zero, such that if β_1 is constrained to equal
 256 1, then:

$$\beta_2^* < \frac{\beta_2}{\beta_1} = \frac{\beta_2}{1} = \beta_2 \quad (13)$$

257 and

$$\beta_1^* < \frac{\beta_1}{\beta_2} = \frac{1}{\beta_2}. \quad (14)$$

258 For example, if there were no bias such that $\alpha_1 = \alpha_2 = 0$ and $\beta_1 = \beta_2 = 1$ so
 259 that $Y_1 = Y_2$, then in general $\beta_1^* \neq \beta_2^* \neq 1$ and $\alpha_1^* \neq \alpha_2^* \neq 0$, and regression
 260 will wrongly indicate a bias. This bias increases as the magnitude of the
 261 correlation between Y_1 and Y_2 decreases (Bollen [6]).

262 In general, the calibration line will fall between the two regression lines.
 263 If $\frac{\sigma_1}{\beta_1} = \frac{\sigma_2}{\beta_2}$, then the calibration curve will fall equidistantly from the two
 264 regression lines. For a fixed $\sigma_1 \neq 0$, as σ_2 approaches zero, the calibration
 265 line will approach the regression line for Y_1 as a function of Y_2 . Similarly, for
 266 a fixed $\sigma_2 \neq 0$, as σ_1 approaches zero, the calibration line will approach the
 267 regression line for Y_2 as a function of Y_1 .

268 Therefore, in practice, ordinary regression is only suitable for calibration

when one of the two measurement devices being compared can be considered to have essentially no random measurement error compared with the other device, and this first device is used as the independent variable in the regression.

3. Methods

3.1. $PM_{2.5}$ Mass Concentration Measurements

Structural equation modeling was applied to calibrate collocated $PM_{2.5}$ mass concentration measurements that were made between 1999 and 2008 at a monitoring site in Pittsburgh, Pennsylvania. The site (40.466N, 79.961W), which is maintained by the Allegheny County Health Department (ACHD), is located in Pittsburgh's Lawrenceville neighborhood, an urban area about two miles northeast of downtown. Seven $PM_{2.5}$ samplers, run by three different monitoring organizations, operated at the Lawrenceville site during some or all of the 1999-2008 period. These samplers included three $PM_{2.5}$ FRM samplers, three $PM_{2.5}$ speciation samplers, and one TEOM.

The primary FRM sampler (FRM 1) at the Lawrenceville site was a Rupprecht and Patashnick (R&P) Partisol-Plus Model 2025 Sequential Air Sampler operated by the ACHD. This sampler measured $PM_{2.5}$ mass concentrations on an approximately daily basis during February 1999-December 2008. The ACHD operated a second R&P FRM sampler (FRM 2) at the site on a 1-in-6-day frequency between January 2005 and December 2008 to provide collocated data. In addition, a third R&P FRM sampler (FRM 3) was

291 operated approximately every sixth day between February 1999 and August
292 2001 by the Upper Ohio River Valley Project (UORVP), which stationed
293 some of its air monitoring equipment at the Lawrenceville site.

294 The three PM_{2.5} speciation samplers at the Lawrenceville site included a
295 Met One Spiral Aerosol Speciation Sampler (SASS) operated by the ACHD,
296 a Desert Research Institute (DRI) Sequential Filter Sampler (SFS) operated
297 by the UORVP, and an Interagency Monitoring of Protected Visual Envi-
298 ronments (IMPROVE) sampler. The SASS sampler measured PM_{2.5} mass
299 concentrations on an approximately 1-in-3-day schedule between June 2001
300 and December 2008, and the IMPROVE sampler measured PM_{2.5} mass con-
301 centrations on a 1-in-3-day schedule between April 2004 and June 2005. The
302 SFS sampler generally operated every sixth day between February 1999 and
303 August 2001, except that it operated on a daily basis during several sampling
304 intensives within this period and during additional intensives in January 2002
305 and October 2002-February 2003.

306 In all of the FRM and speciation samplers, integrated PM_{2.5} samples were
307 collected on Teflon filters at ambient temperature and pressure. Each filter
308 was conditioned in the laboratory before and after exposure under controlled
309 conditions of temperature and humidity, and the mass of sampled PM_{2.5}
310 was determined gravimetrically. The average PM_{2.5} concentration during
311 the sampling period was then determined based upon the mass of sample
312 collected and the volume of air sampled. Although the FRM and speciation
313 samplers employ a common measurement principle, slight differences in per-

314 formance are possible because of differences in sampler design. For example,
315 the R&P FRM samplers were equipped with WINS impactors to separate
316 $\text{PM}_{2.5}$ from larger particles and sampled ambient air at a flow rate of 16.7
317 L/min. The SASS, SFS, and IMPROVE samplers were equipped with cy-
318 clones for particle size separation and sampled at flow rates of 6.7, 20, and
319 22.8 L/min, respectively. The FRM and speciation samplers also differed
320 in their flow control schemes. These variations in sampler design result in
321 different filter face velocities, size cut points, and particle collection efficien-
322 cies, which can lead to discrepancies in measured $\text{PM}_{2.5}$ mass concentrations
323 (Hains et al. [20], Chow et al. [28], U.S. Environmental Protection Agency
324 [29]).

325 The R&P TEOM sampler at the Lawrenceville site was operated by
326 ACHD and measured $\text{PM}_{2.5}$ mass concentrations on a continuous basis dur-
327 ing May 2000-December 2008. In the TEOM, 16.7 L/min of ambient air
328 is drawn through a sharp-cut cyclone to selectively sample $\text{PM}_{2.5}$, and a 3
329 L/min stream of this sampled air is then drawn through a Teflon-coated
330 glass fiber filter that sits atop an oscillating hollow tube. The frequency of
331 oscillation changes as a function of the mass of $\text{PM}_{2.5}$ that has accumulated
332 on the filter, enabling a near real-time determination of the concentration
333 of $\text{PM}_{2.5}$ in the sampled air. Because the TEOM is a continuous monitor,
334 the filter cannot be conditioned in the laboratory to control moisture, as is
335 done in the manual sampling methods described above. Rather, the TEOM
336 filter and sample stream were heated to 50°C to remove moisture from the

337 sampled air and prevent condensation on the filter. At the Lawrenceville
338 site, the TEOM was configured to automatically adjust $\text{PM}_{2.5}$ mass concen-
339 trations, using a linear transformation with slope = 1.03 and intercept = 3.0,
340 in order to partially account for moisture equilibration differences relative to
341 the manual sampling methods (Thermo Scientific [30]). However, even with
342 this correction, additional bias can arise from the loss of semi-volatile par-
343 ticulate matter (e.g., ammonium nitrate, semi-volatile organic compounds)
344 at the elevated TEOM operating temperature, causing an underestimation
345 of total $\text{PM}_{2.5}$ mass relative to the manual samplers (Allen et al. [1], Ayers
346 et al. [7]). This bias is explored in depth in this paper.

347 Data from the ACHD FRM samplers, SASS sampler, and TEOM sam-
348 pler were obtained from the U.S. Environmental Protection Agency's (EPA)
349 Air Quality System (Technology Transfer Network Air Quality System [31]).
350 Data from the UORVP FRM sampler and SFS sampler were obtained from
351 the UORVP project team, and data from the IMPROVE sampler were ob-
352 tained from the Visibility Information Exchange Web System (Views Data
353 Wizard [32]). The quality assurance and quality control (QA/QC) proce-
354 dures applied to each dataset were reviewed, and a common set of data vali-
355 dation flags were applied to all of the data to provide for consistency. Invalid
356 data identified in the native datasets were excluded from the analysis, as were
357 27 outliers identified in the native datasets and one outlier identified upon
358 subsequent QA/QC review. In general, the FRM and speciation samplers
359 measured 24-hour average $\text{PM}_{2.5}$ mass concentrations from midnight to mid-

night. However, the SFS sampler measured 6-hour average concentrations during the sampling intensives identified above. These 6-hour data, as well as the 1-hour concentrations reported by the TEOM, were averaged to compute 24-hour, midnight-to-midnight averages for comparison with the other data. These 24-hour averages were considered to be valid only if they included at least 19 hours of valid measurement time. Table 1 presents descriptive statistics for the 24-hr average $PM_{2.5}$ mass concentrations determined by each monitor at the Lawrenceville site.

In addition to the $PM_{2.5}$ mass concentration data, daily mean temperature data from the Pittsburgh International Airport weather station (40.501N, 80.231W - 23.16 km from the Lawrenceville site) were used to explore the effect of ambient temperature on the calibration results. These data were obtained from the National Climatic Data Center [33].

3.2. Statistical Modeling

The R language and environment for statistical computations (version 2.13.1) was used for all computations and graphics (R Core Team [34]). The OpenMx R package (version 1.1.2-1818) was used to describe the SEM and estimate the model parameters using full information maximum likelihood (FIML) (Boker et al. [35], Boker et al. [36]). The OpenMx package also provided the ability to compute likelihood-based confidence intervals for any and all desired functions of the model parameters.

Preliminary analyses showed that the $PM_{2.5}$ mass concentration measure-

Sampler Type	Monitoring Program	Abbr.	Sampling Period	N	$\mu\text{g}/\text{m}^3$			
					Mean	SD	Min	Median Max
R&P FRM	ACHD	FRM 1	2/1999 - 12/2008	3125	15.2	8.5	2.0	13.4 63.9
R&P FRM	ACHD	FRM 2	1/2005 - 12/2008	228	14.5	8.1	2.4	13.3 58.2
R&P FRM	UORVP	FRM 3	2/1999 - 8/2001	129	17.1	10.8	3.7	14.8 65.2
DRI SFS ^a	UORVP	SFS	2/1999 - 2/2003	439	14.7	8.5	2.6	13.2 74.1
IMPROVE	IMPROVE	IMPROVE	4/2004 - 6/2005	122	14.8	8.5	3.0	13.0 48.0
Met One SASS	ACHD	SASS	6/2001 - 12/2008	766	16.9	8.9	3.1	14.7 62.1
R&P TEOM 50°C ^b	ACHD	TEOM	5/2000 - 12/2008	3056	13.8	8.8	0.0	11.8 57.5

^aCertain 24-hour average concentrations were derived by averaging 6-hour data.

^b24-hour average concentrations were derived by averaging 1-hour data. R&P = Rupperecht & Patashnick, FRM = Federal Reference Method, ACHD = Allegheny County Health Department, UORVP = Upper Ohio River Valley Project, SASS = Spiral Aerosol Speciation Sampler, DRI = Desert Research Institute, SFS = Sequential Filter Sampler, IMPROVE = Interagency Monitoring of Protected Visual Environments, TEOM = Tapered Element Oscillating Microbalance

Table 1: Summary of 24-hour average $\text{PM}_{2.5}$ mass concentrations determined by the collocated samplers at the Lawrenceville site.

382 ments tended to have distributions that were positively skewed. Further in-
383 vestigation showed that the log transformation tended to produce negatively-
384 skewed distributions while the square root transformation produced more
385 normally distributed values. All SEM computations were therefore applied
386 to square root transformed values, and the results, where needed, were trans-
387 formed back to the original scale.

388 Because the bias between PM_{2.5} measurement devices has been shown to
389 vary with temperature and/or season, the SEM developed for the Lawrenceville
390 samplers allowed for an effect of temperature on the calibration results. Fig-
391 ure 2 shows the path diagram for a temperature-stratified SEM that includes
392 all seven samplers at the Lawrenceville site. There are eight sub-models: one
393 for each temperature stratum (only the lowest and highest temperature strata
394 are explicitly shown). In each sub-model, the FRMs are (arbitrarily) used
395 as the reference. (As explained previously, the relevant parameter functions
396 are invariant to this selection.) Each sub-model could be estimated sepa-
397 rately. However, extensive analyses using these types of models indicated,
398 as would be expected, that only the TEOM bias and imprecision parameters
399 depended on temperature. Because these parameters for the other samplers
400 were not temperature-dependent, the model was simplified by equating their
401 parameters across temperature strata. Only the TEOM bias and impreci-
402 sion parameters were allowed to vary with temperature in this model, and
403 thus the overall model is composed of eight linked sub-models. Not all seven
404 samplers operated simultaneously, and in fact, there were never more than

Figure 2: Path diagram for structural equation model for measurement error relating all seven samplers and stratified by temperature. This model is composed of eight sub-models with some parameters constrained to be equal across the temperature strata.

405 a total of five collocated samplers operating on any given day. Therefore,
 406 the dataset of collocated measurements includes numerous missing values.
 407 However, there was enough overlap in the dataset to allow for estimation of
 408 the agreement of every sampler with all six other samplers. Because of the
 409 missing values, the SEM parameters cannot be estimated by simply using the
 410 information contained in the covariance matrix and means vectors. Instead,
 411 the solution was computed using FIML estimation using the raw data. In
 412 this method, the likelihood function was computed for each daily set of col-
 413 located measurements (including those with missing values). They provide
 414 less independent information.

415 **4. Results and Discussion**

416 *4.1. Description of the Pattern of Measurement Collocation*

417 Table 2 shows the number of measurements for each type of sampler and
 418 the number of collocated measurements for each possible pair of samplers.
 419 For example, there were 3,125 FRM 1 measurements, with 220 collocated
 420 pairs with FRM 2 and 87 pairs with FRM 3. FRM 2 and FRM 3 mea-
 421 surements were never collocated (so the number of pairs was zero). For the
 422 speciation samplers, SFS and SASS had 96 pairs as did IMPROVE and SASS.
 423 SFS was never collocated with IMPROVE. For TEOM, there were a total
 424 of 3,056 measurements. Of these, 2,907 were collocated with FRM 1, 223
 425 with FRM 2, and 66 with FRM 3. The TEOM measurements were split into
 426 eight temperature strata having means that ranged from -5.8°C to 24.4°C.

427 The number of measurements in each stratum ranged from 359 (for -5.8°C)
428 to 407 (18.1°C). Of the 372 TEOM measurements in the 14.4°C stratum, 347
429 were collocated with FRM and 97 were collocated with SASS.

430 *4.2. SEM Parameter Estimates*

431 The first SEM included separate bias and imprecision estimates for the
432 FRM samplers (1, 2, and 3). These results are shown in Table 3. For FRM
433 1 and 2, there was essentially no scale bias and no evidence of any constant
434 bias, and the bias estimates are not significantly different ($\alpha = 0.05$) from
435 a no-bias model. FRM 3 consisted of only a small number of measurements
436 (87) collocated with FRM 1, and a small but statistically significant bias
437 was detected relative to FRM 1 and 2. The imprecision (on the square
438 root scale) was similar among the three FRM samplers in comparison to
439 the non-FRM samplers as will be shown. There was enough information,
440 however, to detect small absolute differences in imprecision (i.e., the ratios
441 were statistically significantly different from 1). FRM 2 had an imprecision
442 standard deviation (after correction for the tiny scale bias) that was at least
443 9% lower than FRM 1. The imprecision for FRM 3 (after correction for the
444 somewhat larger scale bias) was at least 47% higher than that for FRM 1.
445 These small biases and small differences in absolute imprecision were ignored
446 for the final analysis so that the three FRMs were used interchangeably.

447 The SEM parameter estimates (relative to the FRM) for the speciation
448 samplers and TEOM are shown in Table 4. The included corresponding 95%

	FRM			Speciation			TEOM ($^{\circ}\text{C}$)									
	1	2	3	SFS	IMP	SASS	All	-5.8	0.7	5.3	10.0	14.4	18.1	21.1	24.4	
1	3125	220	87	338	114	737	2907	343	379	358	352	347	394	369	365	
FRM 2	228	0	0	17	192	223	20	29	27	32	33	26	27	29		
3	129	123	0	6	66	7	9	5	5	8	11	14	7			
SFS				439	0	96	292	65	55	42	18	19	24	38	31	
Spec. IMP					122	96	120	12	9	18	13	16	22	16	14	
SASS						766	749	85	86	94	85	97	88	104	110	
TEOM							3056	359	400	380	369	372	407	387	382	

Table 2: Sample sizes for each sampler and pairs of sampler for $\text{PM}_{2.5}$ mass concentration measurements used in the structural equation model for measurement error at the Lawrenceville site. For the TEOM, sample sizes for each of the eight temperature strata are also shown.

FRM	Parameter	Lower	Estimate	Upper
Scale-Adjusted Intercept Differences				
1	α_1 (fixed)		0.00	
2 to 1	$\alpha_2 - \frac{\beta_2}{\beta_1}\alpha_1$	-0.04	0.01	0.06
3 to 1	$\alpha_3 - \frac{\beta_3}{\beta_1}\alpha_1$	-0.48	-0.32	-0.15
3 to 2	$\alpha_3 - \frac{\beta_3}{\beta_2}\alpha_2$	-0.50	-0.33	-0.15
Scale Bias				
1	β_{1P} (fixed)		1.00	
2 to 1	β_2/β_1	0.98	0.99	1.01
3 to 1	β_3/β_1	1.02	1.06	1.10
3 to 2	β_3/β_2	1.03	1.07	1.11
Scale-Adjusted Imprecision SDs				
1	σ_1/β_1	0.08	0.10	0.11
2	σ_2/β_2	0.02	0.05	0.08
3	σ_3/β_3	0.15	0.18	0.23
Imprecision SD Ratios				
2 to 1	σ_2/σ_1	0.20	0.56	0.91
3 to 1	σ_3/σ_1	1.47	1.91	2.48
3 to 2	σ_3/σ_2	2.24	3.40	18.99

Table 3: Comparison of measurement error among FRM 1, FRM 2, and FRM 3 on the square root transformed scale. Point estimates and corresponding 95% confidence interval bounds are shown. For scale bias and imprecision SD ratios, estimates shown in bold are statistically significantly different from one.

449 confidence intervals indicate that the parameter estimates are reliable (the
 450 interval widths are reasonably narrow). The results in Table 4 clearly show
 451 that the precision of the FRM sampler is superior to that of the other sam-
 452 plers. Among the speciation samplers, the IMPROVE and SASS samplers
 453 had comparable imprecision standard deviations (1.92 and 1.96, respectively,
 454 times that of the FRM), while the SFS sampler was considerably less precise
 455 (imprecision was 4.90 times that of the FRM). The TEOM measurements
 456 made near -5.8°C (the lowest temperature stratum) had the highest imprec-
 457 ision of all samplers (about 7 times that of the FRM). The imprecision of
 458 the TEOM improves with increasing temperature, but even at the highest
 459 temperatures (about 24.4°C) the TEOM is much less precise than the FRM
 460 (imprecision ratio = 2.94). For every comparison presented in Table 4, the
 461 imprecision ratios were statistically significant as shown by the corresponding
 462 confidence intervals.

463 For systematic bias compared to the FRM, Table 4 shows that the IM-
 464 PROVE and SASS samplers may have no scale bias, as the confidence inter-
 465 vals include 1, and the point estimates are near 1. Assuming that there is
 466 no scale bias ($\beta = 1$), the intercepts for these two samplers are near zero al-
 467 though the interval for the SASS sampler is narrow enough to exclude zero.
 468 The SFS scale bias relative to the FRM is estimated as 0.89, with a 95%
 469 confidence interval of 0.85 to 0.93.

470 The results in Table 4 can be used to write the calibration equations to
 471 remove the relative bias between the SFS and FRM samplers. To emphasize

472 that the parameters are relative to the FRM, let

$$\beta_{i/FRM} = \frac{\beta_i}{\beta_{FRM}} = \beta_i \quad (15)$$

473 and

$$\alpha_{i/FRM} = \alpha_i - \frac{\beta_i}{\beta_{FRM}} \alpha_{FRM} = \alpha_i. \quad (16)$$

In this model, the FRM was arbitrarily used as the reference (α_{FRM} fixed to 0 and β_{FRM} fixed to 1) and measurements were stratified into eight temperature strata (having roughly equal sample sizes) with the stratum mean temperatures indicated. Because the FRM was used as the reference, the relationship between each sampler and the FRM is simply:

$$\sqrt{Sampler_i} = \alpha_{i/FRM} + \beta_{i/FRM} \sqrt{FRM} + \epsilon_i - \beta_{i/FRM} \epsilon_{FRM} \quad (17)$$

$$\sqrt{FRM} = -\frac{\alpha_{i/FRM}}{\beta_{i/FRM}} + \frac{1}{\beta_{i/FRM}} \sqrt{Sampler_i} + \epsilon_{FRM} - \frac{\epsilon_i}{\beta_{i/FRM}} \quad (18)$$

474 where ϵ_i is the normally distributed random error with mean zero and stan-
 475 dard deviation σ_i and ϵ_{FRM} is normally distributed with mean zero and
 476 standard deviation σ_{FRM} . In this model, σ_{FRM} is measured on the same
 477 scale as μ and σ_i is measured on the scale $\beta_i \mu$. When the sampler measure-
 478 ment scales differ from the FRM (i.e., $\beta_i/\beta_{FRM} \neq 1$), then to compare the
 479 imprecision of the two devices, it is necessary to compare σ_i/β_i to σ_{FRM} .
 480 The calibration equation linking SFS and FRM after taking expectations

Sampler	Temp. °C	Bias (Systematic Error)				Random Error				Ratio of Adj. Imprecision SD to FRM			
		Slope (Scale)		Intercept (Location)		Scale Adjusted Imprecision SD							
		$\frac{\beta_i}{\beta_{FRM}}$		$\alpha_i - \frac{\beta_i}{\beta_{FRM}}\alpha_{FRM}$		$\frac{\sigma_i}{\beta_i}$				$\frac{\sigma_i/\beta_i}{\sigma_{FRM}}$			
		Lower Est.	Upper Est.	Lower Est.	Upper Est.	Lower Est.	Upper Est.	Lower Est.	Upper Est.	Lower Est.	Upper Est.		
FRM 1-3		1.00		0.00		0.10	0.10	0.11		1.00			
SFS		0.85	0.89	0.93	0.08	0.25	0.42	0.46	0.50	0.55	4.37	4.90	5.52
IMPROVE		0.98	1.02	1.06	-0.24	-0.09	0.07	0.17	0.20	0.23	1.59	1.92	2.31
SASS		0.97	0.98	1.00	0.18	0.24	0.30	0.19	0.20	0.21	1.76	1.96	2.17
TEOM	-5.8	0.63	0.70	0.77	0.19	0.45	0.72	0.63	0.72	0.83	6.04	7.00	8.21
	0.7	0.67	0.72	0.77	0.34	0.52	0.70	0.51	0.57	0.63	4.86	5.52	6.28
	5.3	0.72	0.78	0.83	0.28	0.46	0.64	0.49	0.54	0.60	4.65	5.27	6.02
	10.0	0.79	0.86	0.93	0.09	0.31	0.53	0.49	0.54	0.61	4.63	5.29	6.07
	14.4	0.89	0.94	1.00	-0.07	0.12	0.31	0.41	0.45	0.49	3.84	4.35	4.94
	18.1	0.96	1.01	1.06	-0.24	-0.05	0.15	0.38	0.42	0.46	3.60	4.06	4.58
	21.1	1.01	1.05	1.09	-0.38	-0.28	-0.09	0.34	0.38	0.41	3.25	3.66	4.12
	24.4	0.99	1.02	1.05	-0.35	-0.18	-0.01	0.28	0.30	0.33	2.61	2.94	3.31

Table 4: Parameter estimates and confidence intervals computed using the structural equation model for PM_{2.5} mass concentration measurement error at the Lawrenceville site. The measurements were square root transformed to achieve approximately Normal distribution, and the estimates are shown on the square root transformed scale.

481 and solving for $i = SFS$ is then given by:

$$SFS = \left(\alpha_{SFS/FRM} + \beta_{SFS/FRM} \sqrt{FRM} \right)^2 = \left(0.25 + 0.89 \sqrt{FRM} \right)^2. \quad (19)$$

482 Equivalently, solving the same equation for FRM :

$$FRM = \left(-\frac{\alpha_{SFS/FRM}}{\beta_{SFS/FRM}} + \frac{1}{\beta_{SFS/FRM}} \sqrt{SFS} \right)^2 = \left(-0.28 + 1.12 \sqrt{SFS} \right)^2. \quad (20)$$

483 The calibration equations relating SASS and IMPROVE with FRM can
 484 be similarly derived and along with SFS are illustrated in Figure 3. The
 485 scale bias for the TEOM as shown in Table 4 and Figure 4 diminishes
 486 (β_i/β_{FRM} approaches 1) as the temperature increases, while the intercept
 487 ($\alpha_i - \beta_i/\beta_{FRM}\alpha_{FRM}$) trends from positive to zero and then to negative.
 488 Four-parameter logistic functions were fitted to the intercepts (α_T) and to the
 489 slopes (β_T) for $\sqrt{TEOM}$ as a function of $\sqrt{FRM}$ across temperature T us-
 490 ing weighted non-linear regression. (Modeling was performed using the **SSfpl**
 491 self-starting R function with the **nls** R function for nonlinear regression.) A
 492 weighted linear regression was used to fit the $\sqrt{TEOM}$ scale-adjusted impre-
 493 cision variances (σ_T/β_T) as a linear function of temperature. The regression
 494 models provided a way to pool information across the eight temperature
 495 strata and to interpolate between the observed temperature stratum means.
 496 The weights were equal to the reciprocal of confidence interval widths for
 497 each dependent variable (α_T , β_T , or σ_T/β_T). Figure 5 shows that the rela-

Figure 3: Calibration curves (solid red lines) for the FRM as a function of the speciation samplers SFS, IMPROVE, and SASS. The no bias model is denoted by the dashed green line. IMPROVE shows essentially no bias while the SFS shows a scale bias.

Meaning	Symbol	Estimate	Std. Error	t Value	Pr ($> t $)
Intercepts α_T					
Asymptote $t \rightarrow -\infty$	A_α	0.492	0.054	9.17	0.0008
Asymptote $t \rightarrow \infty$	B_α	-0.266	0.089	-3.00	0.0401
Location ($^{\circ}\text{C}$)	T_{mid_α}	14.271	1.297	11.00	0.0004
Scale ($^{\circ}\text{C}$)	S_α	3.025	1.244	2.43	0.0719
Slopes β_T					
Asymptote $t \rightarrow -\infty$	A_β	0.699	0.022	32.08	< 0.0001
Asymptote $t \rightarrow \infty$	B_β	1.045	0.018	58.03	< 0.0001
Location ($^{\circ}\text{C}$)	T_{mid_β}	10.464	0.934	11.20	0.0004
Scale ($^{\circ}\text{C}$)	S_β	3.773	0.888	4.25	0.0132
Scale-Adjusted Imprecision Variances σ_T/β_T					
Intercept	a	0.378	0.020	18.55	< 0.0001
Slope	b	-0.012	0.001	-9.71	< 0.0001

Table 5: Maximum likelihood parameter estimates for modeling the $\sqrt{TEOM}$ calibration relative to $\sqrt{FRM}$ as a function of temperature (T).

498 tionship for both the scale bias and the intercept are non-linear and follow
 499 sigmoidal (logistic) patterns, whereas the imprecision variance decreases in
 500 a linear fashion. The corresponding parameter estimates are summarized in
 501 Table 5. The calibration equation for relating the TEOM and the FRM can
 502 be expressed as:

$$\sqrt{TEOM} = \alpha(T) + \beta(T)\sqrt{FRM} + [\epsilon_{TEOM}(T) - \beta(T)\epsilon_{FRM}(T)] \quad (21)$$

503 with expectation:

$$TEOM = \left(\alpha(T) + \beta(T)\sqrt{FRM} \right)^2 \quad (22)$$

Figure 4: Calibration curves (solid red lines) for FRM as a function of TEOM stratified by temperature in °C. The mean temperature for each stratum is shown above the corresponding plot. The no bias model is denoted by the dashed green line.

Figure 5: Model fits for the TEOM/FRM calibration as a function of temperature, T (FRM used arbitrarily as the reference): a) intercepts $\alpha(T)$, b) slopes $\beta(T)$, and c) scale-adjusted imprecision variances $(\sigma/\beta)^2(T)$. The calibration curve is then given by: $\sqrt{TEOM} = \alpha(T) + \beta(T)\sqrt{FRM}$ and the scale-adjusted imprecision standard deviation $(\sigma/\beta)(T) = \sqrt{0.378 + 0.012t}$. Alternatively, based on these functions: $FRM = \frac{-\alpha(T)}{\beta(T)} + \frac{1}{\beta(T)}TEOM$ and the imprecision standard deviation is $\sigma_{FRM}(T) = 0.103$.

504 or alternatively as:

$$\sqrt{FRM} = -\frac{\alpha(T)}{\beta(T)} + \frac{1}{\beta(T)}\sqrt{TEOM} + \left[\epsilon_{FRM}(T) - \frac{\epsilon_{TEOM}(T)}{\beta(T)} \right] \quad (23)$$

505 with expectation:

$$FRM = \left(-\frac{\alpha(T)}{\beta(T)} + \frac{1}{\beta(T)}\sqrt{TEOM} \right)^2 \quad (24)$$

506 where

$$\alpha(T) = A_\alpha + \frac{B_\alpha - A_\alpha}{1 + e^{\frac{T_{mid_\alpha} - T}{S_\alpha}}} = 0.492 + \frac{-0.266 - 0.492}{1 + e^{\frac{14.271 - T}{3.025}}}, \quad (25)$$

507

$$\beta(T) = A_\beta + \frac{B_\beta - A_\beta}{1 + e^{\frac{T_{mid_\beta} - T}{S_\beta}}} = 0.699 + \frac{1.045 - 0.699}{1 + e^{\frac{10.464 - T}{3.773}}}, \quad (26)$$

508

$$\epsilon_{TEOM}(T) \sim N(0, (a + bT)^2) = N(0, (0.378 - 0.012T)^2) \quad (27)$$

509 and

$$\epsilon_{FRM}(T) \sim N(0, 0.103^2). \quad (28)$$

510 The calibration curves relating TEOM and FRM as a function of tempera-
 511 ture are shown in Figure 4. As temperature increases, the calibration curve
 512 approaches the diagonal line (no bias model).

513 4.3. Calibration Predictions and Bias for FRM Based on TEOM

514 The calibration equations previously developed for FRM based on TEOM
515 measurements (and vice versa) (equations 21 - 26) are a nonlinear function
516 of the observed TEOM measurement and temperature.

517 To obtain a clearer understanding of how the bias changes, Table 6 shows
518 the predicted FRM for observed TEOM measurements from 0 to 50 $\mu\text{g}/\text{m}^3$
519 in steps of 10 $\mu\text{g}/\text{m}^3$ with temperature ranging from -10 °C to 40 °C in steps
520 of 5 °C. Figure 6 shows the predicted bias (FRM - TEOM) for the same con-
521 centration and temperature ranges. The FRM and TEOM exhibit excellent
522 agreement at certain combinations of PM_{2.5} concentration and ambient tem-
523 perature, as indicated by solid thick zero bias contour lines. However, it is
524 clear that for observed TEOM values from about 6 to 34 $\mu\text{g}/\text{m}^3$, the TEOM
525 is always biased low relative to the FRM regardless of temperature. This
526 bias is substantial at low temperatures (e.g., 20 $\mu\text{g}/\text{m}^3$ for a TEOM value
527 of 30 $\mu\text{g}/\text{m}^3$ and an ambient temperature of -10 °C) but decreases in mag-
528 nitude as temperature increases and ambient PM_{2.5} concentration decreases.
529 For TEOM values above 34 $\mu\text{g}/\text{m}^3$, the bias follows a similar trend with
530 temperature, except that it eventually reverses sign (i.e., such that TEOM
531 values are biased slightly high relative to FRM values at sufficiently high
532 temperatures). For TEOM values below 6 $\mu\text{g}/\text{m}^3$, the bias tends to be small
533 regardless of temperature, and it vanishes for certain combinations of TEOM
534 and temperature in a nonlinear fashion. In all cases, the bias between the
535 TEOM and FRM tends to be small (i.e., $< 1 \mu\text{g}/\text{m}^3$) for ambient tempera-

tures above 17.65 °C , which is denoted by the vertical dashed line in Figure 6.

4.4. Comparison of the SEM Calibration with Ordinary Regression

As the correlation between the samplers drops from one and approaches zero, the regression lines increasingly diverge. The calibration curve always falls between the two regression lines and is closer to the regression line that has the independent variable with the lowest imprecision standard deviation. For example, because the FRM has a much smaller imprecision than the TEOM at low temperatures, the regression line for TEOM as a function of FRM is very close to the actual calibration line at these temperatures, whereas the regression line for FRM as a function of TEOM is farther from the actual calibration line. This effect diminishes as the TEOM imprecision standard deviation decreases (i.e., with increasing temperature) and approaches the constant imprecision standard deviation for the FRM. If the two samplers had equal imprecision, the calibration line would be equally distant from both regression lines.

For illustrative purposes, the TEOM and FRM collocated measurements at -5.8 °C were used to compute the ordinary regression curves (on the square root scale and transformed back to the original scale) for FRM as a function of TEOM and vice versa. The deficiencies in doing this have been previously described. As shown in Figure 7, the regression of TEOM on FRM produces a curve that is not too far from the actual calibration curve, be-

TEOM Value	Temperature (°C)										
	-10	-5	0	5	10	15	20	25	30	35	40
0	0.49	0.48	0.46	0.36	0.16	0.00	0.03	0.06	0.06	0.06	0.06
10	14.53	14.37	13.85	12.50	10.71	10.28	10.67	10.78	10.78	10.77	10.76
20	32.29	31.93	30.72	27.54	22.97	20.83	20.71	20.66	20.60	20.57	20.56
30	50.65	50.08	48.16	43.06	35.52	31.42	30.65	30.40	30.27	30.23	30.21
40	69.33	68.54	65.89	58.83	48.21	42.04	40.55	40.07	39.87	39.80	39.78
50	88.21	87.21	83.82	74.76	61.00	52.67	50.41	49.69	49.42	49.33	49.31

Table 6: Predicted FRM measurement based on observed TEOM measurement as a function of temperature.

Figure 6: Bias between observed TEOM measurement and equivalent FRM measurement as a function of temperature. The vertical grey line at 17.65 °C indicates the temperature at which the bias is a constant bias.

558 cause the FRM (used as the independent variable) has substantially smaller
 559 imprecision than the TEOM at this temperature. The typical approach used
 560 by the EPA and others (Eberly et al. [14], Bortnick et al. [11], Vega et al.
 561 [12], Kashuba and Scheff [13]), however, is to use the regression of the FRM
 562 on the TEOM. This regression curve is substantially different from the actual
 563 calibration curve (and the other regression curve). For example, a TEOM
 564 measurement of $15 \mu\text{g}/\text{m}^3$ corresponds to an FRM measurement of about 24
 565 $\mu\text{g}/\text{m}^3$ at -5.8°C as shown by the calibration curve, giving a relative bias of
 566 about $9 \mu\text{g}/\text{m}^3$. The regression curve for FRM as a function of TEOM pre-
 567 dicts an FRM mass concentration of only about $17.5 \mu\text{g}/\text{m}^3$, resulting in an
 568 underestimation in the bias of about $6.5 \mu\text{g}/\text{m}^3$. In general, this regression
 569 curve is a very poor estimate of the actual calibration curve. The fact that
 570 the TEOM as a function of the FRM was a better choice in this situation
 571 unfortunately cannot be generalized. When the imprecision standard devia-
 572 tions of both samplers are equal, both regression lines will be poor estimates
 573 of the calibration curve if there is sufficient random error in the samplers.

574 **5. Conclusions**

575 Although ordinary linear regression analysis commonly has been applied
 576 for calibrating $\text{PM}_{2.5}$ measurement devices, this approach generally is in-
 577 appropriate when both measurement devices are subject to random error.
 578 Ordinary regression will produce a nearly correct calibration curve when one
 579 of the measurement devices is much more precise than the other device, pro-

Figure 7: Comparison of the SEM-derived calibration equation with naive regression equations for TEOM versus FRM and FRM versus TEOM at -5.8 °C.

580 vided that the more precise device is specified as the independent variable in
581 the regression. Using an imprecise measurement device as the independent
582 variable can produce a regression line that is far from the correct calibration
583 curve. Moreover, while ordinary regression can produce an approximately
584 correct calibration line when one device is much more precise than the other,
585 it requires a priori knowledge of the difference in imprecision standard de-
586 viations (adjusted for any scale bias) between the two devices. In practice,
587 this is difficult to apply, because if the scale bias and imprecision were al-
588 ready known, the calibration equation would also in effect already be known.
589 Errors-in-variables methods such as Deming regression and RMA regression
590 require similar upfront knowledge of the imprecision standard deviations and
591 are similarly impractical to apply. In addition, all of these regression methods
592 under the best circumstances only use information from two measurement
593 devices and therefore fail to incorporate all the available information in cal-
594 ibrations involving more than two collocated devices.

595 Structural equation modeling as described in this paper provides a general
596 and flexible methodology to appropriately estimate the calibration curves re-
597 lating $\text{PM}_{2.5}$ mass concentration samplers. This approach uses all available
598 information to simultaneously estimate bias and imprecision and does not
599 require making uninformed assumptions or obtaining external information
600 about the random errors associated with those samplers. In addition, it is
601 possible to include other factors that influence the calibration even if the
602 effect of those factors is non-linear. Finally, the SEM software makes it pos-

603 sible to easily compute calibration equations for most practical applications,
604 to place confidence interval bounds on the calibration curve, and to produce
605 prediction intervals for converting measurements from one device scale to
606 another. It should be emphasized, however, that datasets appropriate for
607 constructing calibration curves must either include collocated measurements
608 from more than two devices or include independent replicates of at least one
609 of the devices when there are only two devices. This assumes that the ran-
610 dom errors in each device are statistically independent and uncorrelated. In
611 some circumstances, for example, when one measurement is in part derived
612 from another measurement, the errors will likely be correlated and this will
613 require a larger number of devices and/or independent replicates.

614 The results presented here demonstrate the successful application of SEM
615 to characterize the relative bias and imprecision among seven collocated
616 PM_{2.5} samplers at an urban monitoring site in Pittsburgh, Pennsylvania.
617 The modeling shows that: (1) the three FRM samplers were considerably
618 more precise than the three speciation samplers and the TEOM sampler,
619 (2) the TEOM sampler showed a significant negative bias relative to the
620 FRM samplers, particularly at low ambient temperature, and the magnitude
621 of this bias decreased non-linearly with increasing temperature, and (3) the
622 speciation samplers showed only small or statistically insignificant bias rel-
623 ative to the FRM samplers, and this bias was not temperature-dependent.
624 This analysis is expanded in a future paper to calibrate data from 97 PM_{2.5}
625 measurement devices at 47 monitoring sites so that these data can be used

626 to provide an unbiased representation of spatiotemporal variations in $\text{PM}_{2.5}$
627 concentration for epidemiologic modeling.

628 **6. Acknowledgments**

629 This study was funded by the Electric Power Research Institute con-
630 tracts EP-P29581/C13936 and EP-P46301/C19907. The feasibility study
631 for PARIES was funded by the U.S. Department of Energy contract DE-
632 PS26-04NT42066. Additional funding has been provided by the Allegheny
633 County Health Department (Pittsburgh, Pennsylvania).

634 [1] G. Allen, C. Sioutas, P. Koutrakis, R. Reiss, F. W. Lurmann, P. T.
635 Roberts, Evaluation of the TEOM method for measurement of ambient
636 particulate mass in urban areas, *Journal of the Air & Waste Manage-*
637 *ment Association* 47 (1997) 682–689. PMID: 9180064.

638 [2] A. Chung, D. P. Chang, M. J. Kleeman, K. D. Perry, T. A. Cahill,
639 D. Dutcher, E. M. McDougall, K. Stroud, Comparison of real-time
640 instruments used to monitor airborne particulate matter, *Journal of*
641 *the Air & Waste Management Association* 51 (2001) 109–120. PMID:
642 11218418.

643 [3] D. J. Eatough, R. W. Long, W. K. Modey, N. L. Eatough, Semi-volatile
644 secondary organic aerosol in urban atmospheres: meeting a measure-
645 ment challenge, *Atmospheric Environment* 37 (2003) 1277–1292.

- 646 [4] J. Cyrys, G. Dietrich, W. Kreyling, T. Tuch, J. Heinrich, PM_{2.5} mea-
 647 surements in ambient aerosol: comparison between Harvard impactor
 648 (HI) and the tapered element oscillating microbalance (TEOM) system,
 649 Science of the Total Environment 278 (2001) 191–197.
- 650 [5] L.-A. Tortajada-Genaro, E. Borrs, Temperature effect of tapered ele-
 651 ment oscillating microbalance (TEOM) system measuring semi-volatile
 652 organic particulate matter, Journal of Environmental Monitoring: JEM
 653 13 (2011) 1017–1026. PMID: 21347460.
- 654 [6] K. A. Bollen, Structural Equations with Latent Variables, John Wiley
 655 & Sons, New York, 1989.
- 656 [7] G. P. Ayers, M. D. Keywood, J. L. Gras, TEOM vs manual gravi-
 657 metric methods for determination of PM_{2.5} aerosol mass concentrations,
 658 Atmospheric Environment 33 (1999) 3717–3721.
- 659 [8] A. Charron, R. M. Harrison, S. Moorcroft, J. Booker, Quantitative in-
 660 terpretation of divergence between PM₁₀ and PM_{2.5} mass measurement
 661 by TEOM and gravimetric (Partisol) instruments, Atmospheric Envi-
 662 ronment 38 (2004) 415–423.
- 663 [9] J. J. Schwab, H. D. Felton, O. V. Rattigan, K. L. Demerjian, New
 664 York State urban and rural measurements of continuous PM_{2.5} mass by
 665 FDMS, TEOM, and BAM, Journal of the Air & Waste Management
 666 Association 56 (2006) 372–383. PMID: 16681203.

- 667 [10] K. Zhu, J. Zhang, P. J. Lioy, Evaluation and comparison of continuous
668 fine particulate matter monitors for measurement of ambient aerosols,
669 Journal of the Air & Waste Management Association 57 (2007) 1499–
670 1506. PMID: 18200935.
- 671 [11] S. M. Bortnick, B. W. Coutant, S. I. Eberly, Using continuous PM_{2.5}
672 monitoring data to report an air quality index, Journal of the Air &
673 Waste Management Association 52 (2002) 104–112. PMID: 15152670.
- 674 [12] E. Vega, E. Reyes, A. Wellens, G. Sanchez, J. Chow, J. Watson, Com-
675 parison of continuous and filter based mass measurements in Mexico
676 City, Atmospheric Environment 37 (2003) 2783–2793.
- 677 [13] R. Kashuba, P. A. Scheff, Nonlinear regression adjustments of multi-
678 ple continuous monitoring methods produce effective characterization of
679 short-term fine particulate matter, Journal of the Air & Waste Manage-
680 ment Association 58 (2008) 812–820. PMID: 18581811.
- 681 [14] S. Eberly, T. Fitz-Simons, T. Hanley, L. Weinstock, T. Tamanini,
682 G. Denniston, B. Lambeth, E. Michel, S. Bortnick, Data Quality Objec-
683 tives (DQOs) for Relating Federal Reference Method (FRM) and Con-
684 tinuous PM_{2.5} Measurements to Report an Air Quality Index (AQI),
685 Technical Report EPA-454/B-02-002, U.S. Environmental Protection
686 Agency, Office of Air and Radiation, Office of Air Quality Planning
687 and Standards, Research Triangle Park, NC 27711, 2002.

- 688 [15] R. J. Adcock, A problem in least squares, *The Analyst* 5 (1878) 53–54.
- 689 [16] C. H. Kummell, Reduction of observation equations which contain more
690 than one observed quantity, *The Analyst* 6 (1879) 97–105.
- 691 [17] W. E. Deming, *Statistical Adjustment of Data*, Wiley, 1943.
- 692 [18] H. Hauck, A. Berner, B. Gomiscek, S. Stopper, H. Puxbaum, M. Kundi,
693 O. Preining, On the equivalence of gravimetric PM data with TEOM and
694 beta-attenuation measurements, *Journal of Aerosol Science* 35 (2004)
695 1135–1149.
- 696 [19] J. H. Lee, P. K. Hopke, T. M. Holsen, A. V. Polissar, Evaluation of
697 continuous and filter-based methods for measuring PM_{2.5} mass concen-
698 tration, *Aerosol Science and Technology* 39 (2005) 290–303.
- 699 [20] J. C. Hains, L.-W. A. Chen, B. F. Taubmann, B. G. Doddridge, R. R.
700 Dickerson, A side-by-side comparison of filter-based PM_{2.5} measure-
701 ments at a suburban site: A closure study, *Atmospheric Environment*
702 41 (2007) 6167–6184.
- 703 [21] K. A. Kermack, J. B. S. Haldane, Organic correlation and allometry,
704 *Biometrika* 37 (1950) 30–41.
- 705 [22] G. Ayers, Comment on regression analysis of air quality data, *Atmo-
706 spheric Environment* 35 (2001) 2423–2425.

- 707 [23] D. Green, G. W. Fuller, The implications of tapered element oscillat-
708 ing microbalance (TEOM) software configuration on particulate matter
709 measurements in the UK and Europe, *Atmospheric Environment* 40
710 (2006) 5608–5616.
- 711 [24] J. L. Jaech, *Statistical Analysis of Measurement Errors*, Wiley & Sons,
712 New York, 1985.
- 713 [25] G. Dunn, *Statistical Evaluaton of Measurement Errors: Design and*
714 *Analysis of Reliability Studies*, Hodder Arnold, London, 2004.
- 715 [26] R. B. Kline, *Structural Equation Modeling*, Guilford Press, New York,
716 1998.
- 717 [27] G. A. Marcoulides, I. Moustaki, *Latent Variable and Latent Structure*
718 *Models*, Lawrence Erlbaum Associates, London, 2002.
- 719 [28] J. C. Chow, J. G. Watson, D. H. Lowenthal, L. W. Antony Chen, R. J.
720 Tropp, K. Park, K. A. Magliano, $PM_{2.5}$ and PM_{10} mass measurements
721 in California’s San Joaquin Valley, *Aerosol Science and Technology* 40
722 (2006) 796–810.
- 723 [29] U.S. Environmental Protection Agency, *Evaluation of $PM_{2.5}$ Chemical*
724 *Speciation Samplers for Use in the EPA National $PM_{2.5}$ Chemical Speci-*
725 *ation Network*, Technical Report EPA-454/R-01-005, U.S. Environmen-
726 tal Protection Agency, 2001.

- 727 [30] Thermo Scientific, Operating Manual TEOM Series 1400a Ambient Par-
728 ticulate (PM₁₀) Monitajada (AB Serial Numbers); Revision B, Thermo
729 Fisher Scientific Inc., Franklin, MA, 2007.
- 730 [31] Technology Transfer Network Air Quality System, 2010.
731 <http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqdata.htm/>.
- 732 [32] Views Data Wizard, 2010.
733 <http://views.cira.colostate.edu/web/Data/Wizard>.
- 734 [33] National Climatic Data Center, 2011.
735 <http://www.ncdc.noaa.gov/oa/ncdc.html>.
- 736 [34] R Core Team, R: A Language and Environment for Statistical Comput-
737 ing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- 738 [35] S. Boker, M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies,
739 R. Estabrook, S. Kenny, T. Bates, P. Mehta, J. Fox, Openmx: An open
740 source extended structural equation modeling framework, *Psychome-*
741 *trika* 76 (2011) 306–317.
- 742 [36] S. M. Boker, M. C. Neale, H. H. Maes, M. J. Wilde, M. Spiegel, T. R.
743 Brick, R. Estabrook, T. C. Bates, P. Mehta, T. von Oertzen, R. J. Gore,
744 M. D. Hunter, D. C. Hackett, J. Karch, A. M. Brandmaier, OpenMx 1.3
745 User Guide, 2012.