

PALPEBRAL FISSURE LENGTH MEASUREMENT: ACCURACY OF THE FAS FACIAL PHOTOGRAPHIC ANALYSIS SOFTWARE AND INACCURACY OF THE RULER

Susan J. Astley

Professor of Epidemiology and Pediatrics, University of Washington, Seattle WA

ABSTRACT

Background

Accurate fetal alcohol spectrum disorder diagnoses require accurate facial measurement. The Fetal Alcohol Syndrome (FAS) Facial Photographic Analysis Software was developed to overcome measurement error known to occur with ruler measurement of the PFL. Recent publications have queried the Software's accuracy.

Objectives

1) Demonstrate the Software's ability to accurately measure a PFL from a 2-dimensional digital facial photograph. 2) Demonstrate the frequency and magnitude of error when the PFL is measured directly by clinicians using a ruler.

Methods

Objective 1: PFLs of mannequins were measured using the Software and a sliding digital caliper, with the latter serving as the gold-standard accurate measure. Mannequins allowed the caliper prongs to be placed directly on the landmarks that define the PFL. Objective 2: PFLs of 1,027 patients evaluated at the University of Washington FAS Diagnostic & Prevention Network were measured with the Software and directly by one or two clinicians using a ruler.

Results

Objective 1: The Software derived PFLs that were identical to or within 0.2 mm of the caliper measures. Objective 2: There was tremendous inter-rater variability in PFLs measured by clinicians using a hand held ruler. Seventy-seven percent of patients had their PFLs measured incorrectly (greater than 1 mm error) by at least one of the two clinicians using a ruler.

Conclusion

The FAS Facial Photographic Analysis Software measures the PFL with the same accuracy as a sliding digital caliper, as it was programmed to do. Direct measurement of the PFL with a ruler is very prone to error.

Key Words: *Fetal alcohol spectrum disorders, palpebral fissure length, FASD 4-Digit Diagnostic Code, WA State Fetal Alcohol Syndrome Diagnostic & Prevention Network*

FAS is a birth defect syndrome caused by maternal use of alcohol during pregnancy. FAS is characterized by growth deficiency, a unique cluster of minor facial anomalies and central nervous system (CNS) structural, neurological and/or functional abnormalities.¹

The three diagnostic facial features of FAS as defined by the FASD 4-Digit Diagnostic

Code² are: small palpebral fissure lengths (2 or more standard deviations below the mean), a smooth philtrum (Rank 4 or 5 on the University of Washington Lip-Philtrum Guide), and thin upper lip philtrum (Rank 4 or 5 on the University of Washington Lip-Philtrum Guide).

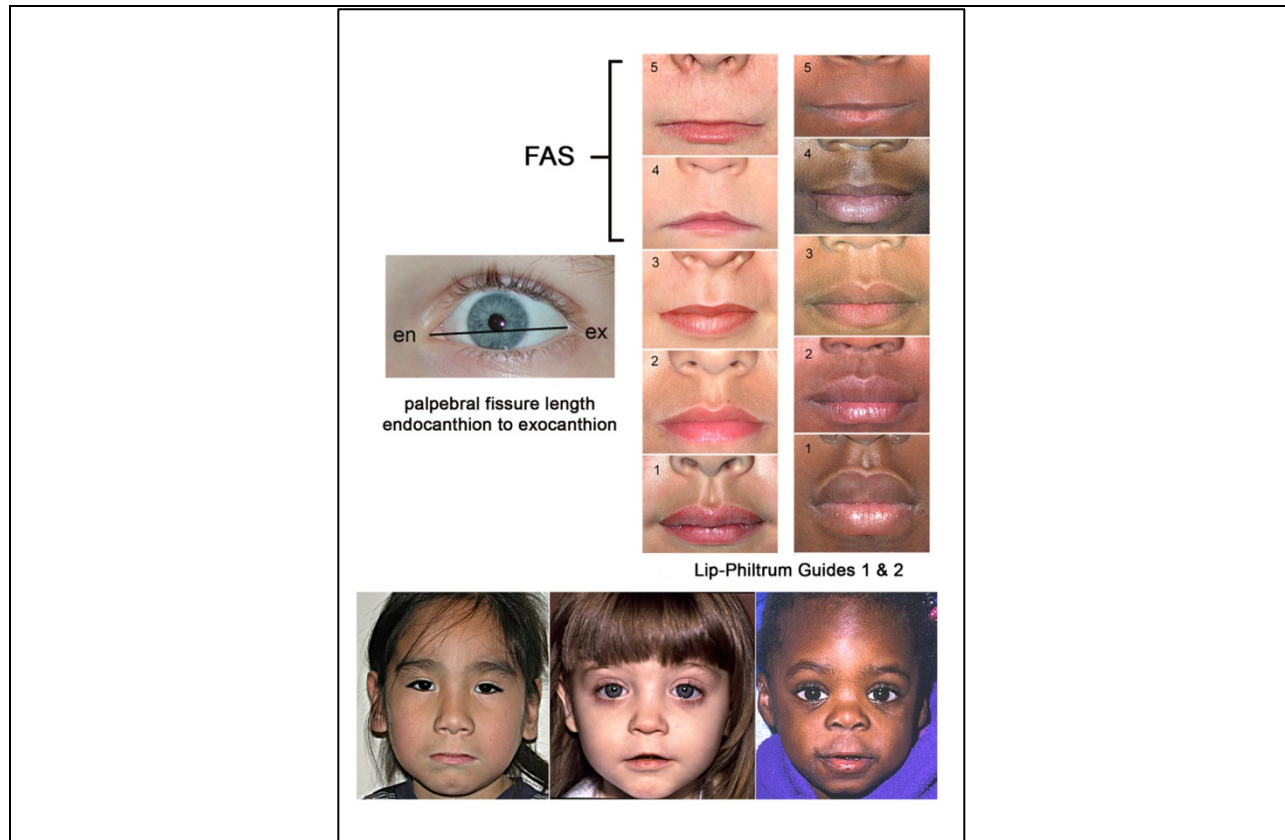
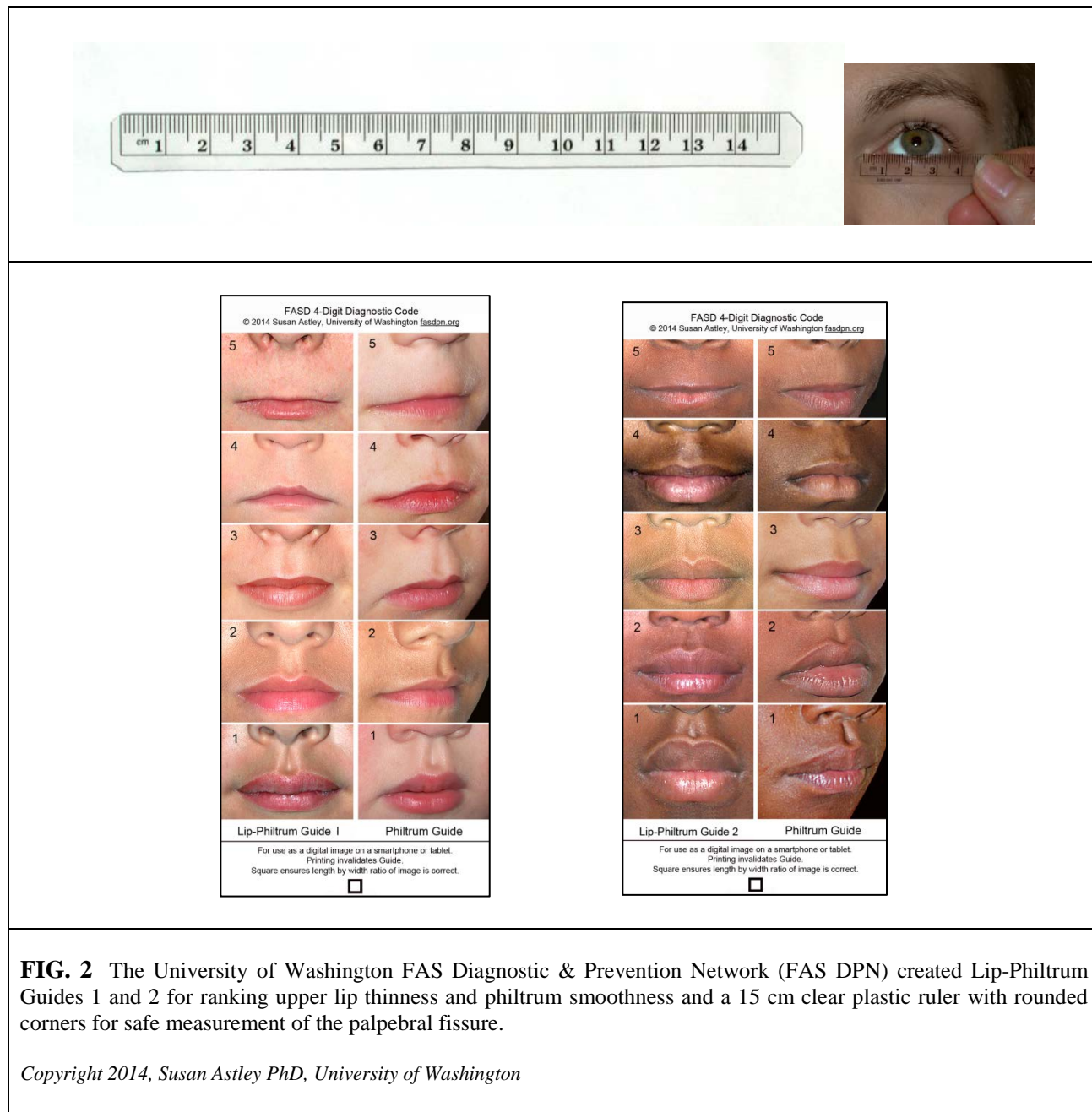


FIG. 1 The three diagnostic facial features of FAS as defined by the FASD 4-Digit Diagnostic Code² include: 1) short palpebral fissure lengths (2 or more standard deviations below the mean), 2) A smooth philtrum (Rank 4 or 5 on the Lip-Philtrum Guide), and 3) a thin upper lip (Rank 4 or 5 on the Lip-Philtrum Guide). Lip-Philtrum Guides 1 and 2 are used to rank upper lip thinness and philtrum smoothness. The philtrum is the vertical groove between the nose and upper lip. The guides reflect the full range of lip and philtrum shapes with Rank 3 representing the population mean. Ranks 4 and 5 reflect the thin lip and smooth philtrum that characterize the FAS facial phenotype. Guide 1 is used for Caucasians and all other races with lips like Caucasians. Guide 2 is used for African Americans and all other races with lips as full as African Americans. Examples of the FAS facial phenotype across three races: Native American, Caucasian, and African American.

Copyright 2014, Susan Astley PhD, University of Washington



In 1997, the University of Washington Fetal Alcohol Syndrome Diagnostic & Prevention Network (FAS DPN)³ developed tools to more accurately measure the facial features of FAS. These tools included Lip-Philtrum Guides 1 and 2 to measure lip thinness and philtrum smoothness and a 15 cm clear plastic ruler with rounded corners to measure the PFL (Figure 2). The Lip-

Philtrum Guides improved the accuracy and reproducibility of lip and philtrum measures by introducing pictorial scales that case-defined and rank-ordered lip thinness and philtrum smoothness.⁴ The 15 mm clear plastic ruler was an improvement over the foot-long wooden ruler or soft tape measure observed to be used by some clinicians. Despite these improvements, facial

measures were still prone to error. Selecting the correct 5-point rank for lip thinness and philtrum smoothness could prove difficult if the subject's features fell close to the transition between two ranks. And measuring a PFL with a ruler poses many challenges. An accurate measure requires the clinician to place the ruler very close to the open eye and align themselves first with one corner of the patient's eye and then the other corner of the eye to avoid parallax errors. Young patients are often reluctant to allow a clinician to do this and may be unable to sit still enough to allow this measure to be obtained safely and

accurately. A 1 mm error in measurement translates into almost 1 SD of error on the PFL growth chart⁵, thus there is little room for error. Animations depicting these errors are presented on the FAS DPN website. With the advent of digital photography and computerized image analysis, the University of Washington developed the FAS Facial Photographic Analysis Software in 2003⁶ (upgraded in 2012⁷) that allowed the clinician to more accurately measure facial features from 2 dimensional (2D) digital facial photographs (Fig. 3).

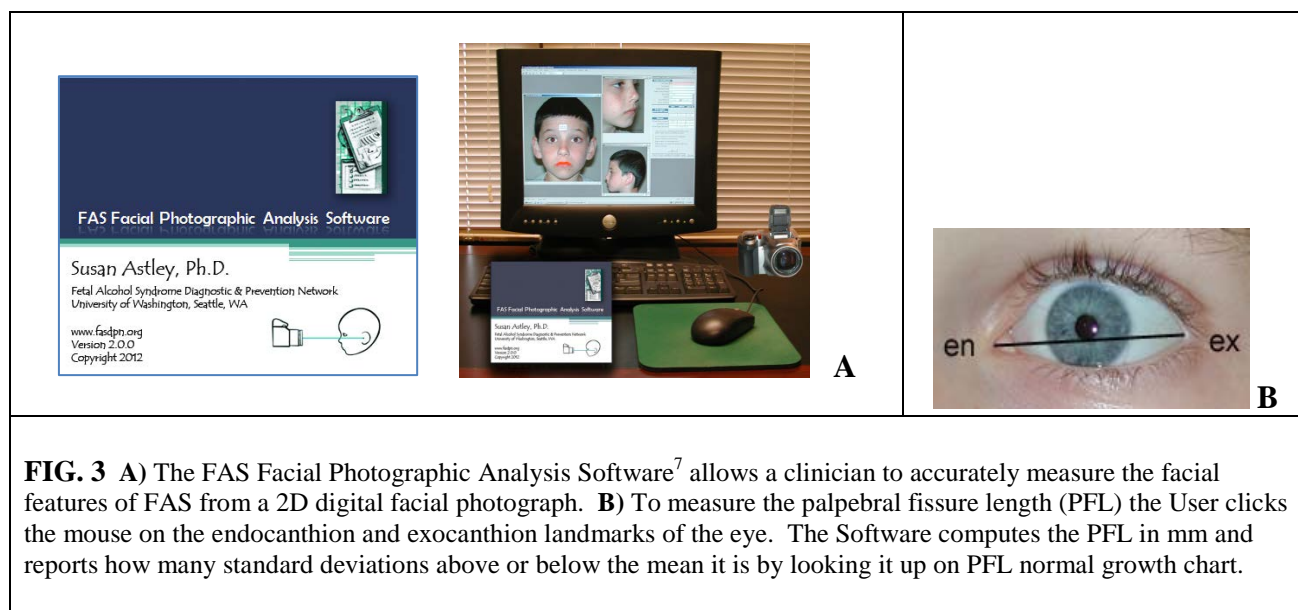


FIG. 3 A) The FAS Facial Photographic Analysis Software⁷ allows a clinician to accurately measure the facial features of FAS from a 2D digital facial photograph. B) To measure the palpebral fissure length (PFL) the User clicks the mouse on the endocanthion and exocanthion landmarks of the eye. The Software computes the PFL in mm and reports how many standard deviations above or below the mean it is by looking it up on PFL normal growth chart.

The Software was specifically designed to overcome the error often observed when PFLs were measured with a handheld ruler.^{8,9} The Software was programmed to measure a PFL with the accuracy of a sliding digital caliper (the gold standard). When used in accordance with the *Software's Manual of Instructions*, the Software generates accurate measures. The Software is programmed to generate accurate measures of the PFL, innercanthal distance, and lip circularity from a 2D image of a 3D object when the digital images meet the specifications (resolution, alignment, facial expression, etc) specified in the Instruction Manual. Throughout this publication,

all reference to the "Software" refers to Version 1.0⁶ or Version 2.0⁷, as both use the same methods for measuring the FAS facial features. The upgrade in Version 2.0 simply provided Users with access to additional PFL normal growth charts.

There has been some confusion lately in the published literature as to the Software's ability to accurately measure a PFL from a 2D photograph. Two studies have reported on the discordance of ruler, caliper, and Software measures of the PFL. One study concluded that their Software measures were often smaller than their ruler measures.¹⁰⁻¹¹ The other study concluded

their Software measures were comparable to their ruler measures, but shorter than their caliper measures.¹² Since neither study included a gold-standard measurement of the PFL, neither study could comment on the accuracy of any of the three methods of measurement. If the Software measures tended to be smaller than the ruler and caliper measures, were the Software measures correct and the ruler and caliper measures overestimated the true PFL? Or were the Software measures incorrect and the Software underestimated the true PFL? The purpose of this study was to answer these questions.

OBJECTIVES

The objectives of this study were:

1. To demonstrate the ability of the FAS Facial Photographic Analysis Software's to accurately measure a PFL from a 2-dimensional digital facial photograph. The PFL is the distance between the endocanthion and exocanthion landmarks (Fig 2).

2. To demonstrate the frequency and magnitude of error when PFLs are measured directly by clinicians using a ruler.

It is important to clarify that this study is not assessing the accuracy of the Software for the first time, but rather demonstrating the accuracy of the Software. The accuracy of the Software was assessed and confirmed prior to its release in 2003. The Software was developed to overcome the high frequency of error known to occur when the PFL is measured directly using a ruler (Fig. 4).

Although a demonstration of the Software's PFL measurement accuracy is posted on the FAS DPN website and the website cautions clinicians about the risk of error associated with ruler measurement of the PFL¹, neither of these topics have been comprehensively addressed by the FAS DPN in the published literature. With recent inquiry into which method of PFL measurement is most accurate (ruler, caliper, or Software)^{10,11} a clear demonstration of the Software's accuracy and the ruler's inaccuracy is warranted to help guide clinicians in their choice of method.

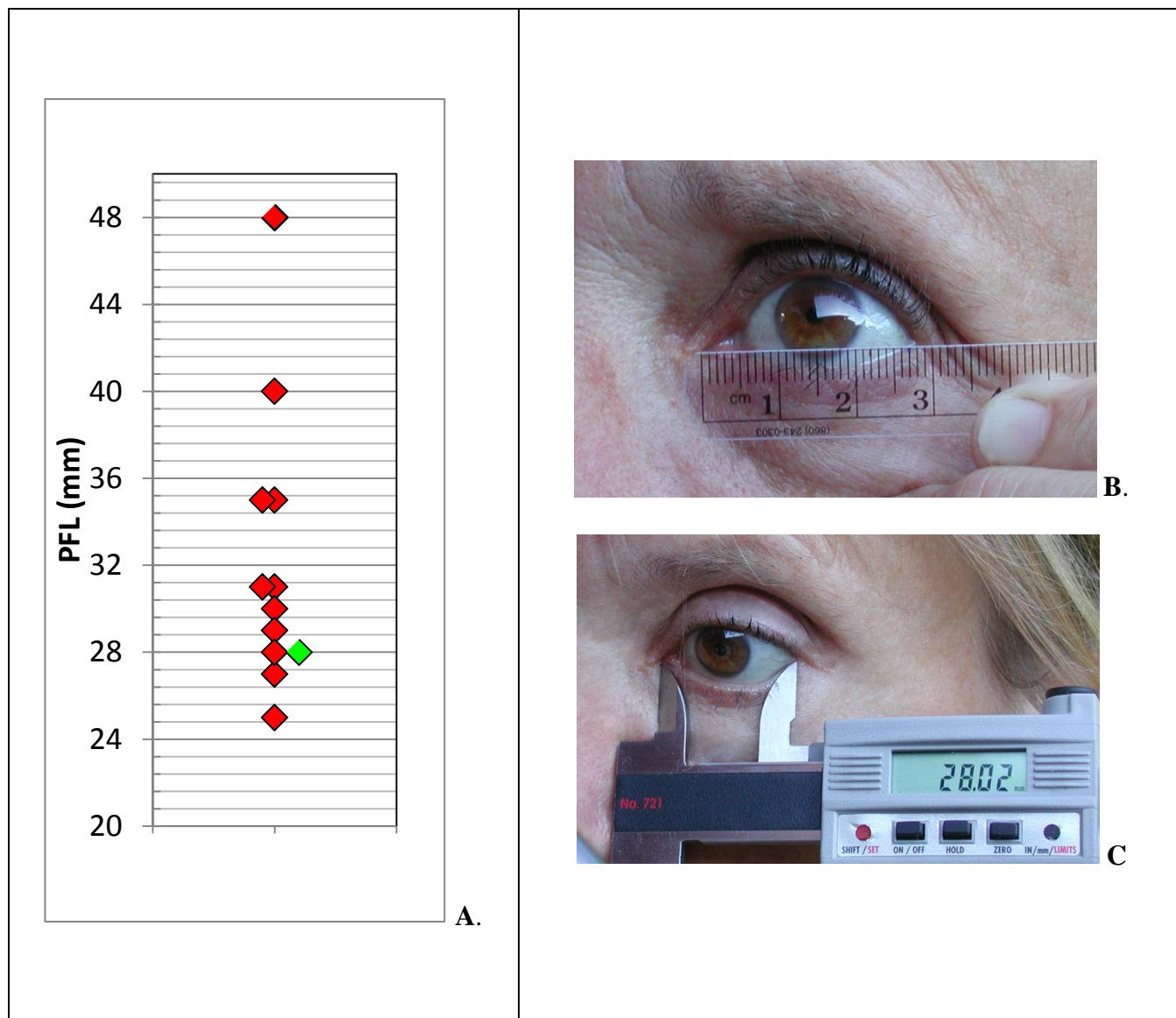

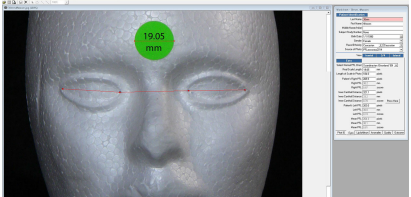


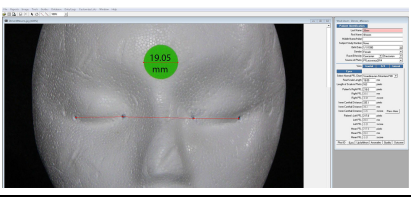


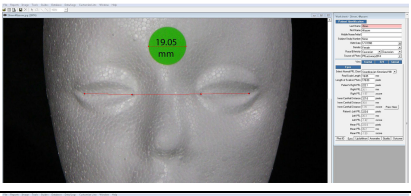


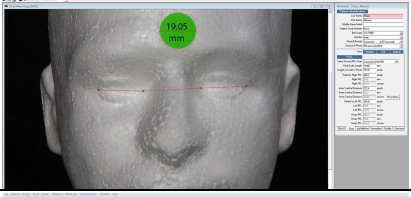


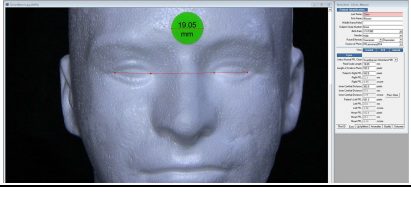


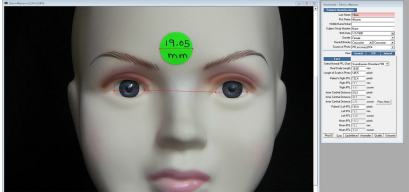


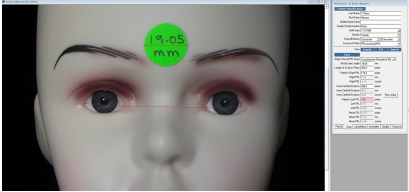


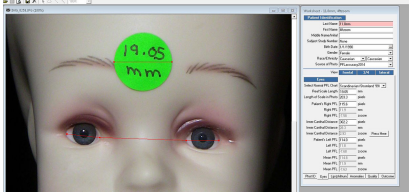


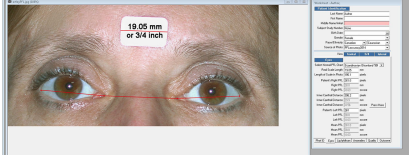



FIG. 4 A. The inherent inaccuracy of measuring a PFL with a ruler was demonstrated in an exercise in which the author's left PFL was measured by 11 clinicians using a ruler. Although measurement of a PFL with a ruler may seem straightforward (B), the clinicians' measures ranged from 25 mm to 48 mm (red diamonds). C. The true PFL was 28.02 mm (green diamond) as demonstrated with a sliding digital caliper placed directly on the endocanthion and exocanthion landmarks. USE OF A CALIPER IN THIS WAY IS VERY DANGEROUS AND SHOULD NEVER BE DONE WITH A PATIENT.

TABLE 1 Software and caliper measures of the palpebral fissure lengths (PFL) from eight mannequins and one human subject.

Mannequin	OFC (cm)	Assigned Age (yrs)	Left PFL (mm)			Right PFL (mm)			Mean PFL (mm)		Facial Image	Software PFL	Caliper PFL
			Caliper	Software	Caliper – Software	Caliper	Software	Caliper – Software	Software	Caliper - Software			
Adult female	51.5	20	30	30	0	30	30.2	-0.2	30.1	-0.1			
Adolescent female 2	53	14	28	28	0	28	28	0	28	0			
Adolescent female	51.5	15	26	26.3	-0.3	26	26.1	-0.1	26.2	-0.2			
Adolescent male	51.5	15	25	24.9	0.1	25	24.9	0.1	24.9	+0.1			
Adolescent male	54.5	15	22	22.0	0	22	22.3	-0.3	22.2	-0.2			

Palpebral fissure length measurement: accuracy of the FAS facial photographic analysis software and inaccuracy of the ruler

Child female	53.5	8	19	19.3	-0.3	19.0	19.1	-0.1	19.2	-0.2			
Toddler female	47.5	2	17.8	17.9	-0.1	17.8	17.8	0	17.9	-0.1			
Infant female	39.0	0.16	11.8	11.8	0	11.8	11.9	-0.1	11.9	-0.1			
Author	54.8		28.02	28	+0.02								

OFC: occipital frontal circumference. mm: millimeters

METHODS

Objective 1: Software Accuracy

Mannequin Heads: Eight life-size mannequin heads representing males and females across the lifespan (infant, child, adult) were used (Table 1). The head circumferences of the mannequins ranged from 39 to 54.5 mm, reflective of normal head circumferences for individuals infant through adult. Three mannequin heads were made from plastic and had eyes clearly painted on the face. Their PFLs were measured with calipers to be 11.8 mm, 17.8 mm, and 19.0 mm. Five mannequins were made from white Styrofoam and had facial features formed into the Styrofoam, but did not have eyes painted on the face. To accommodate this, a black, fine-tip felt marker was used to place small dots on each of the five mannequins to represent the endocanthion and exocanthion landmarks for the right and left palpebral fissure lengths. The five Styrofoam mannequins came with a variety of eye sizes ranging from 22 to 30 mm, thus the landmarks were placed to create PFLs of exactly 22, 25, 26, 28 and 30 mm, purposely spanning a range of PFL that would be observed in infants, children and adults.⁵

Caliper Measures of PFLs: The right and left PFLs were measured with a sliding digital caliper (Fig 4, Table 1). The prongs of the caliper were placed directly on the endocanthion and exocanthion landmarks of the palpebral fissure to obtain an exact (gold standard) measure of the

PFL. A caliper can only serve as a gold standard of accurate measurement if the prongs of the caliper can be placed directly in contact with the 2 points in space being measured. A photograph was taken of each caliper measure to document the reading on the caliper. Human subjects were not used in this study because the gold standard measure of the PFL requires placing the prongs of the caliper directly on the endocanthion and exocanthion landmarks; a procedure that cannot be safely conducted with a human.

Facial Photograph: A frontal facial photograph of each mannequin head was obtained in accordance with the FAS Facial Photographic Analysis Software Version 2.0 Instruction Manual.⁷ Briefly, a ¾ inch (19.05 mm) round paper sticker was placed between the eyebrows to serve an internal unit of measure in the digital image (Fig. 5, Table 1). A 5-megapixel digital camera was held 4 feet from the mannequin. The zoom feature was used to zoom in on the face until the head filled the camera frame (Table 1). Care was taken to obtain a frontal photograph with no vertical (tipped up or down) or horizontal (turned right or left) rotation.

To achieve this, the camera was held in the mannequin's Frankfort horizontal plane (the plane that runs through the external ear canals and the lower borders of the bony orbital rims) (Fig. 5). The camera was also aligned such that the right and left ears were equally visible. This camera alignment procedure is demonstrated in an animation posted on the FASDPN website.

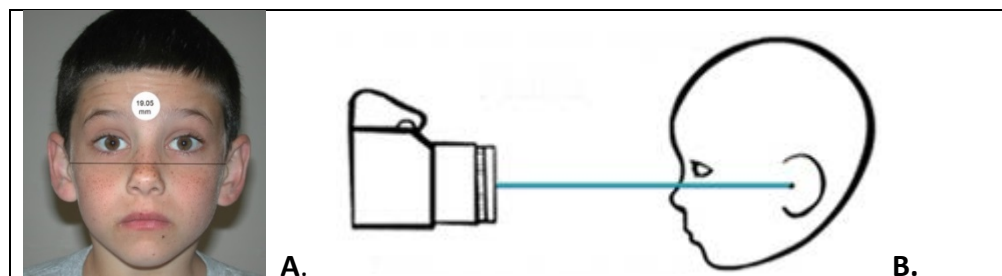


FIG. 5 A) Example of a properly aligned facial photograph. The camera is aligned in the subject's Frankfort horizontal plane (B. the plane (blue line) that runs through the external auditory canals and the lower borders of the bony orbital rims). When the camera is held in the subject's Frankfort horizontal plane, a line drawn between the right and left auditory canals on the facial photograph would fall along the lower bony orbital rims.

Copyright 2014, Susan Astley PhD, University of Washington

Software Measures of PFLs: Each facial image was imported into the Software for analysis. All analyses were conducted by the author. In accordance with the Software Instruction Manual, the Software's zoom feature was used to enlarge the image so the eyes and sticker covered the full width of the computer monitor (Table 1). This affords more accurate measures. To establish the measurement scale in the photo, the diameter of the paper sticker was measured. Distance in a digital image is measured in units called pixels. Pixels are the little dots of light that make up the image displayed on a computer screen. The diameter of the sticker in pixels was measured using the "single distance" tool. With the mouse the User selects the tool and clicks on the right edge of the sticker and then the left edge, bisecting the circle. This procedure establishes how many pixels is equivalent to 19.05 mm in the digital image. Next, the 3-distance tool was used to measure the right PFL, inner canthal distance (the distance between the eyes), and left PFL. With the mouse, the User selects the tool and clicks on the right exocanthion, right endocanthion, left endocanthion, and left exocanthion, in that order. In so doing, the Software records the right PFL, inner canthal distance, and left PFL in pixels. The Software then converts these three measures from pixels to mm using the pixel- to-mm conversion ratio established from measuring the diameter of the paper sticker.

Data Analysis: All data was entered into SPSS¹³ for analysis. The caliper measure of the PFL served as the gold standard in this study. If the Software was generating an accurate measure of the PFL, the Software measure should match the caliper measure. The Software measure of each PFL was subtracted from the caliper measure of the PFL to assess the Software's measurement accuracy. If the Software was generating an accurate measure of the PFL, the Software measure should match the caliper measure (the difference between the two measures would be zero). A negative outcome would document the Software underestimated the PFL. A positive outcome would document the Software overestimated the PFL.

Objective 2: Ruler Variability

The frequency and magnitude of error observed when clinicians use a ruler to measure the PFL was demonstrated by comparing: A) Ruler measures to caliper measures; B) Ruler measures to Software measures; and C) Ruler measures obtained by two clinicians. All data collection and analysis had Human Subjects Review Board approval.

A. Ruler versus Caliper

In 2000, the author invited 11 clinicians at the University of Washington to participate in an exercise in which they were asked to measure her left PFL to the best of their ability with a 15 cm plastic ruler (Fig. 4). These clinicians were selected because measuring PFLs was a routine part of their clinical practice. The only instruction they received was to view the image in Fig. 4B documenting the endocanthion and exocanthion landmarks that define the PFL. They were asked to write the PFL measure on a piece of paper and insert it into an envelope. The identity of the clinicians was not recorded. The true length of the palpebral fissure (obtained with a caliper) was shared with the clinician after they placed their measure in the envelope. They were asked not to reveal what measure they obtained, to maintain the anonymity of the exercise.

Data Analysis: The caliper measure of the author's left PFL and ruler measures obtained by 11 clinicians were plotted to illustrate inter-rater variability (Fig. 4).

B. Ruler versus Software

Over the past 20 years in the WA FAS DPN clinics, 1,027 patients had their PFLs measured by one of 21 different medical doctors using a 15 cm ruler. All 1,027 patients also had a digital facial photograph taken that was measured by the author using the FAS Facial Photographic Analysis Software. The 1,027 patients were 43% female and were on average 8.5 (5.5 SD) years of age. They ranged in age from 2 months to 48 years of age, with 92% under the age of 15 years.

Data Analysis: To compare the ruler versus Software measures of each patient's PFLs, the mean of the right and left PFLs derived by the Software was subtracted from the mean of the right and left PFLs obtained by the clinician using a

ruler (ruler mean PFL minus Software mean PFL). A negative difference reflected the Software measure of the mean PFL was longer than the ruler measure. A positive difference reflected the Software measure of the mean PFL was shorter than the ruler measure. Since the smallest unit of measure on the ruler is 1 mm (Fig. 2) and the smallest unit of measure using the Software is 0.1 mm, the ruler measure was considered a match to the Software measure if the ruler measure was within -0.9 to +0.9 mm of the Software measure. The outcomes were plotted by documenting what proportion of subjects had PFL measures that were < 1, 1, 2, or 3 or more mm different between the ruler and Software measures. Differences were categorized into 1 mm bins because 1 mm is the smallest unit of measurement demarcated on the ruler. An error of 1 mm is equivalent to an error of roughly three quarters of a SD on a normal growth chart for PFL.⁵ A 1 mm error could result in an incorrect diagnosis under the umbrella of FASD. For example, if a 13 year old girl had a PFL that was truly 27 mm, her PFL would fall 0.6 standard deviations below the mean for a girl her age using the Stromland normal PFL growth charts⁵. The FASD 4-Digit Code would classify this PFL as being in the normal range (PFL ABC-score = A). If the clinician incorrectly measured her PFL by -1 mm, her 26 mm PFL would appear to be 1.3 SDs below the mean. The FASD 4-Digit Code would classify this PFL as being in the moderately short range (PFL ABC-score = B). If the clinician incorrectly measured her PFL by 2 mm, her 25 mm PFL would appear to be 2.0 SDs below the mean. The FASD 4-Digit Code would classify this PFL as being in the significantly short range (PFL ABC-score = C). Thus, a 1 mm error can incorrectly classify the PFL in the PFAS range and a 2mm error can incorrectly classify the PFL into the FAS range.

It is important to note that in the absence of a gold-standard measure (e.g., a caliper measure of the PFL), if a difference is observed between the ruler and Software measures, one of three conclusions can be drawn: 1) the ruler measure is incorrect; 2) the Software measure is incorrect; or 3) both measures are incorrect. The outcome of Objective 1 will help determine which of these three outcomes is supported.

C. Ruler versus Ruler

241 patients that had their PFLs measured with a ruler by both the medical doctor and the author. The 241 patients were 44% female and were on average 8.9 (5.1 SD) years of age. They ranged in age from 1.3 to 40 years of age, with 93% under the age of 15 years.

Data Analysis: To compare the two ruler measures of a patient's PFLs, the medical doctor's measure of the patient's left PFL was subtracted from the author's measure of the patient's left PFL (author's measure of PFL minus doctor's measure of PFL). A negative difference reflected the doctor's measure of the left PFL was longer than the author's measure. A positive difference reflected the doctor's measure of the left PFL was shorter than the author's measure. Since the smallest unit of measure on the ruler was 1 mm, the two clinicians' measures of the PFL were considered a match if they were within plus or minus 0.9 mm of one another. The outcomes were plotted by documenting what proportion of subjects had PFL measures that were < 1, 1, 2, or 3 or more mm different between the two ruler measures.

RESULTS

Objective 1. Software Accuracy:

The Software produced right and left PFL's that either matched or were no more than 0.2 mm different from the right and left PFLs measured with the calipers. These outcomes did not vary by gender, age (infant, child, adult), OFC (39-54.5 cm), or PFL (11.8 -30.0 mm) (Table 1).

Objective 2A. Ruler Variability:

Ruler versus Caliper

The PFL measures recorded by the 11 clinicians ranged from 25 mm to 48 mm (Fig. 4). The true PFL measured with a caliper was 28.02 mm. Eight of the 11 physician measures were incorrect by 2 to 16 mm. .

Objective 2B. Ruler Variability:

Ruler versus Software

One thousand twenty-seven patients across the full age span (infant to adult) had their PFLs measured by both the medical doctor using a ruler and from a photograph using the Software (Fig.

6). These measures involved 21 different doctors over 21 years. The doctor's measure with the ruler was within 1 mm of the Software measure 44.7% of the time. The two measures differed by 2 or more mm 21.1% of the time. When the ruler and Software PFL measures were compared among the subset of 166 patients that were under the age of 4 years, the distribution of error was near identical to that of the entire age spectrum. Since

Objective 1 demonstrated the Software accurately derives a PFL from a 2D facial photo when the Software is used properly (high quality photos with proper alignment) and Objective 2A demonstrated high variability in ruler measures, the discordance between the ruler and Software measures presented here were most likely the result of incorrect ruler measures.

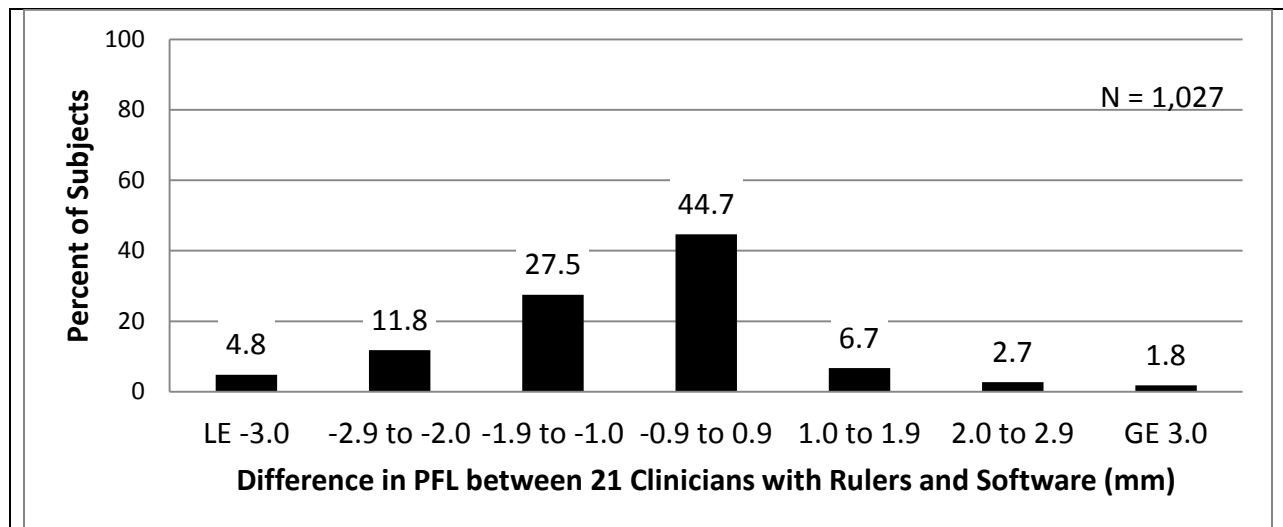


FIG. 6 Objective 2B The prevalence and magnitude of discordance is illustrated when the PFLs of 1,027 patients were measured from 2D facial photos using the FAS Facial Photographic Analysis Software and measured by one of 21 medical doctors using a 15 cm ruler. The columns represent how often the ruler measure minus the Software measure was 1 or more mm different. A negative difference reflects the ruler measure of the PFL was smaller than the Software measure of the PFL. A positive difference reflects the ruler measure was larger than the Software measure. Differences less than 1 mm were considered a match since 1 mm was the finest level of measurement marked on the ruler.

Of the 1,027 patients who had their PFL measured by a clinician using a ruler, 297 also had their PFL measured by the author. The 297 patients were 42% female and were on average 8.9 (5.0 SD) years of age. They ranged in age from 8 months to 40 years of age, with 94% under

the age of 15 years. When the author's ruler measures were compared to the Software measures, 84.9 % of the ruler measures were less than 1 mm different from the Software measures (Fig. 7). Less than 2% were 2 or more mm different from the Software.

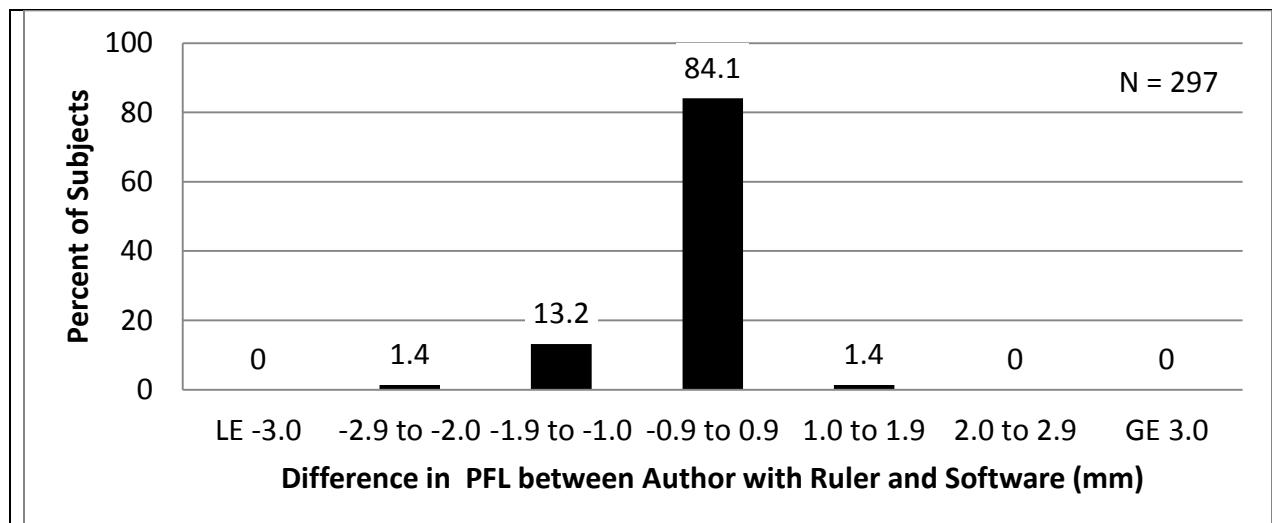


FIG. 7. Objective 2B: The prevalence and magnitude of discordance is illustrated when the PFLs of 297 patients were measured from 2D facial photos using the FAS Facial Photographic Analysis Software and measured by the author using a 15 cm ruler. The columns represent how often the ruler measure minus the Software measure was 1 or more mm different. A negative difference reflects the ruler measure of the PFL was smaller than the Software measure of the PFL. A positive difference reflects the ruler measure was larger than the Software measure. Differences less than 1 mm were considered a match since 1 mm was the finest level of measurement marked on the ruler.

As documented above, it is standard procedure in the University of Washington FAS DPN clinic to measure a patient’s PFL directly with a ruler and from a photo using the Software. Due to the confirmed accuracy of the Software, we use the Software measure for the diagnosis. The ruler measure simply serves as an opportunity for the clinician to hone their skills with a ruler in the event they have to measure a patient without access to the Software. To demonstrate the value of this training, Fig. 8 documents the improvement in skill across three clinicians. The mean difference between the ruler and Software measures across all patients measured each year are presented from 2004 through 2011. These 320 patients were 42% female and were on average 8.1 (5.1 SD) years of age. They ranged in age from 2 months to 40

years of age, with 93% under the age of 15 years. When the Software was first introduced in the clinic in 2003, PFLs measured with a ruler were on average 2 mm discordant from the Software measures. Over time, the ruler measures moved into the green zone, documenting the ruler measures on average were more concordant (within 1 mm) of the Software measures. But it is important to point out that the 1 SD error bars (Fig. 8A) and the plot of individual patient measures (Fig. 8B) document the clinicians’ individual measures still had an unacceptable level of variation from the Software measures, even if on average their measures were improving over time. Over half of the individual measures fell outside the green lines (were more than 1 mm discordant from the Software measure).

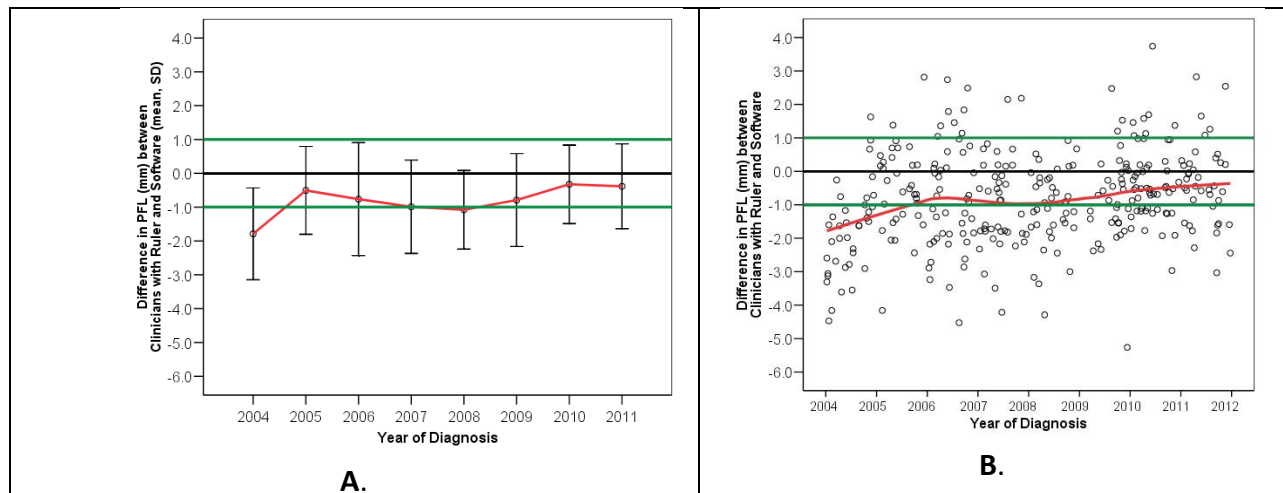


FIG. 8 Objective 2B: The difference in the ruler minus Software measure of the PFL is plotted annually across three clinicians and 320 patients. Ruler measures were collected at the beginning of the FASD diagnostic evaluation. Software measures were not available to compare to until the end of the evaluation. A. On average, the accuracy of PFL measures obtained with a ruler improved over time when the three clinicians had Software measures to compare to their ruler measures. The mean difference in the ruler minus Software measure of all PFLs collected each year are plotted with 1 SD error bars. The zone between the green lines reflects the preferred level of accuracy (less than 1 mm difference between the ruler and Software measures). The Software was introduced into the Clinic in 2003. Even though the mean difference tends to fall between the green lines (Fig. 8A), the magnitude of difference between the ruler and Software measures for each individual patient (Fig 8B) continued to show far too much variability with over half the individual measures falling outside the green zone.

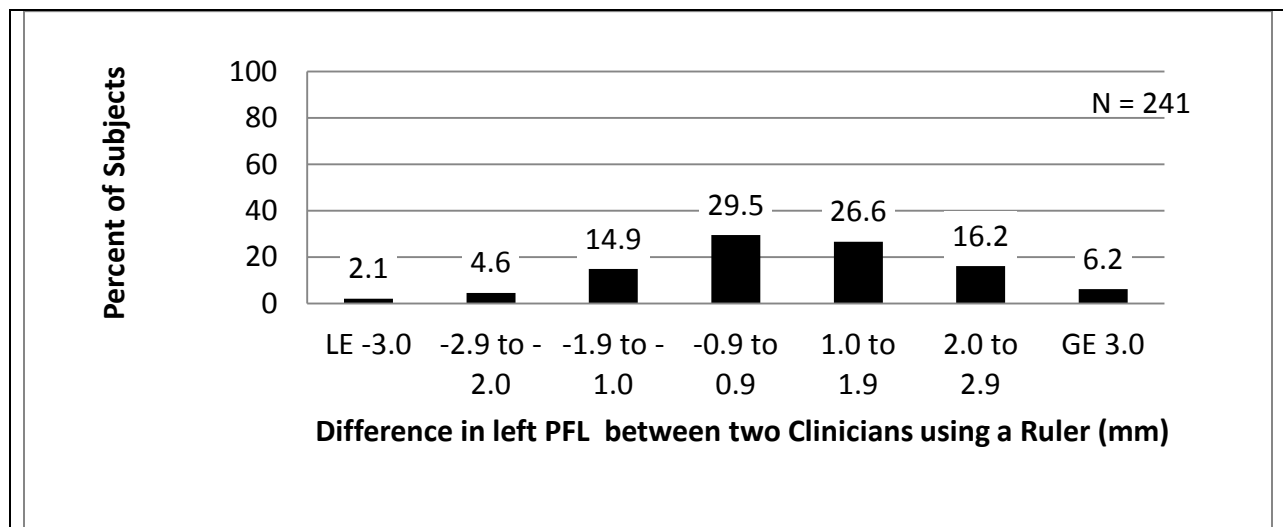


FIG. 9 Objective 2C: Ruler vs Ruler Measure. The prevalence and magnitude of discordance is illustrated when the left PFLs of 241 patients were measured by both the author and the medical doctor using a 15 cm ruler. The columns represent how often the two measures were 1 or more mm different. A negative difference reflected the doctor's measure of the left PFL was longer than the author's measure. A positive difference reflected the doctor's measure of the left PFL was shorter than the author's measure. Since the smallest unit of measure on the ruler was 1 mm, the two clinicians' measures of the PFL were considered a match if they were within plus or minus 0.9 mm of one another.

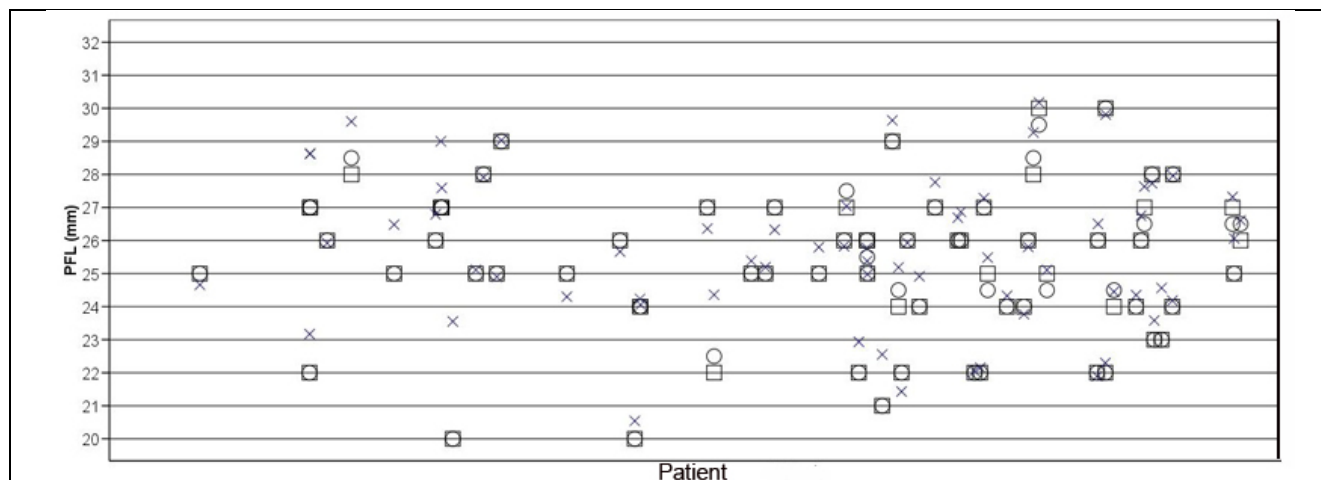


FIG. 10 Objective 2C: Ruler vs Ruler vs Software Measures of the 71 (29.5%) patients in Fig. 9 with concordant ruler measures (measures within 0.9 mm of one another). It is important to note that when the two PFL measures obtained with a ruler were concordant (within 1 mm of one another), this does not confirm the PFL measures were correct. The two measures could be concordant and both incorrect. This figure documents that 15 of the 71 patients with concordant PFL ruler measures have one or more ruler measures that are discordant (greater than 1 mm different) from the Software measure Key: Circle = Clinician 1. Square = Clinician 2. X = Software measure.

Objective 2C. Ruler Variability:

Ruler versus Ruler

When patients ($n = 241$) had their PFLs measured by two different individuals (the medical doctor and the author), the PFL measures were within 1 mm of one another 29.5% of the time (Fig. 9). One mm is the smallest unit of measure marked on the ruler. The PFL measures were 2 or more mm different 29.1% of the time. Most importantly, the measures were discordant (1 or more mm different) 70.5% of the time. Despite the absence of a gold standard (caliper) measure of the patients' PFLs, one can still conclude that at least 70.5% of the patients had their PFL measured incorrectly by at least one of the two clinicians. It is also important to note that when the two PFL measures obtained with a ruler were concordant (within 1 mm of one another), this does not confirm the PFL measures were correct. The two measures could be concordant and both incorrect. The majority of these 241 patients also had their PFLs measured from a photo using the Software. Figure 10 documents how often the two ruler measurements

matched the Software measurement among the 71 (29.5%) patients with concordant ruler measurements. If the Software measures are assumed accurate (as demonstrated in Objective 1), then 15 of the 71 patients with concordant PFL ruler measures have one or more ruler measures that are discordant (greater than 1 mm different) from the Software measure. This would suggest that 77% of the patients had their PFLs measured incorrectly (greater than 1 mm error) with a ruler by at least one of the two clinicians.

DISCUSSION

The data presented in this report confirm that the FAS Facial Photographic Analysis Software generates accurate measures of the PFL from a 2D digital facial photograph when compared to the gold standard. The gold standard was a sliding digital caliper with the caliper prongs placed directly in contact with the endocanthion and exocanthion landmarks that define the PFL. This measurement accuracy was confirmed prior to the

release of the Software. The release of the Software was contingent on its ability to generate accurate measures of the FAS facial features.

The data presented in this report also demonstrate the prevalence and magnitude of error when a patient's PFL is measured directly with a ruler. The prevalence and magnitude of this error was well known prior to the release of the Software.^{4,8,9} One of the primary reasons the Software was developed was to overcome this error.

In 2004, a document was posted on the FAS DPN website demonstrating the Software's ability to accurately measure a PFL. The author's PFLs were measured with a ruler, with a caliper, and with the Software (Table 1). Photos of each measure were taken to document the outcomes. These data are included in this report (Fig. 4) to demonstrate the PFL is measured accurately whether obtained from a mannequin or human.

Two other investigative teams¹⁰⁻¹² have reported on the discordance of ruler, caliper, and Software measures of the PFL. Since neither study included a gold-standard measurement of the PFL, neither study could comment on the accuracy of any of the three methods of measurement. Inclusion of a gold standard of measurement in our study allowed us to confirm the technical accuracy of the FAS Facial Photographic Analysis Software. This confirmation provides some helpful context for comparing our outcomes to those reported in these two previous studies.

In the first study, the PFLs of 40 children (2 months to 15 years old) referred for a FASD evaluation were measured using both a ruler and the Software.^{10,11} The number of clinicians involved was not reported. Avner et al^{10,11} reported their Software measures of the PFLs were on average 2 mm shorter than their ruler measures across all 40 children; and 3 mm shorter among the subset of 21 children under 4 years of age. These contrasts are 3 to 4-fold greater than observed in our study. Our Software measures were on average 0.7 mm longer (not shorter) than the ruler measures across 1,027 patients measured by one of 21 clinicians, and 0.9 mm longer among the subset of 166 patients under 4 years of age. Since our study demonstrated the Software measures a PFL within 0.2 mm of a gold standard

caliper measure in a properly standardized photograph, the 2.0 mm to 3.0 mm contrast between their ruler and Software measures cannot be explained by Software error. Contrasts that are 2.0 mm to 3.0 mm in magnitude are the result of User error. The investigators discuss the types of User error that may occur when measuring a PFL with a ruler (e.g., patient cooperation, examiner's skill), but there are also opportunities for User error when measuring a PFL with the Software (e.g. poor photo quality, eyes not fully open, inaccurate identification of landmarks by User, etc). These sources of error are discussed more fully below. The investigators went on to compare the number of children identified with PFLs 2 or more SDs below the mean using the ruler and Software methods of measurement. The ruler method identified 9 children; the Software method identified 14 children.

The investigators concluded "*The method of computer-assisted measurement tends to underestimate the true length and, hence, over diagnose short palpebral fissure, especially in children under four years old*". The study methodology does not support this conclusion because the study did not have a measure of the "true" length of the palpebral fissure. The 'true' length of the palpebral fissure would require use of a gold standard method of measurement. The study did not incorporate a gold standard measure of the PFL. The investigators also computed sensitivity and specificity and reported "*Since the photographic method tends to overestimate the number of short PFLs (sensitivity=100%, specificity=64%), it is likely to over-diagnosis FAS. Therefore, telediagnosis would be most useful if it were followed by direct measurement, (since the photographic method alone produced false positives, but no false negatives).*" Once again the study methodology does not support this conclusion. Sensitivity and specificity require a gold standard measure of the PFL to represent the "true positive". The study had no gold standard measure of the PFL. Thus the study cannot compute the sensitivity or specificity for the ruler or the Software. Our current study demonstrates that it is highly unlikely that direct measurement of the PFL with a ruler would provide a more

accurate measure than the Software when the software is used properly.

In the second study a single clinician measured the PFLs of 50 children and 50 adults using all three tools (Software, ruler, and caliper).¹² Cranston et al¹² reported their Software measures were concordant with their ruler measures 42% of the time. This is consistent with our findings. We observed concordance between our Software measures and ruler measures 44.7% of the time across 1,027 patients measured by one of 21 clinicians. Cranston et al¹² also reported their Software measures were concordant with their caliper measures only 18% of the time. This is in stark contrast with our results. We observed 100% concordance between our Software and caliper measures of the PFL. The most likely reason their Software and caliper measures were discordant was because the Software was measuring the actual PFL and the caliper was measuring an approximation of the PFL. This is illustrated in their Figures 2 and 1B, respectively. Since their subjects were human, they could not obtain an accurate PFL with a caliper because it was too dangerous to place the prongs of the caliper directly on the individual's endocanthion and exocanthion landmarks that define the PFL. We used mannequins in our study to overcome this limitation.

Although the present study has demonstrated that the FAS Facial Photographic Analysis Software is programmed to accurately measure a PFL from a properly standardized 2D digital facial photograph, this does not mean that every measure of a PFL using the Software is accurate. The PFL measures obtained using the Software are only as accurate as the quality of the photo and the skills of the Software User. To minimize User error, the Software comes with detailed instructions and a practice case "John Doe" (Fig. 5A). The practice case serves two purposes. 1) It provides the User with an example of what a perfect set of standardized digital photographs (frontal, ¾, and lateral views) looks like. John Doe's photos display the following qualities: They are focused, well lit, high resolution, and properly aligned. John has no smile, his lips are gently closed, and his eyes are fully open. These qualities are not only important

for accurate photo analysis of FAS facial features, they are also important when the facial features are being measured directly with a ruler and Lip-Philtrum Guide. 2) The Software also provides the User with an opportunity to practice measuring John Doe's photos to confirm they have the necessary skills to derive accurate measures. The Software comes with John Doe's photo set fully and accurately measured and permanently stored in the Software. When measuring facial features with the Software, whether for clinical or research purposes, it is imperative the User ensure and report the quality of the photos measured. It is also important they confirm they can measure John Doe's' photoset with high inter-rater reliability (i.e., their measures of John Doe's facial features match the gold-standard measures recorded in the Software). They should also confirm they have high test-retest reliability (i.e., they obtain the same PFL and lip circularity measures across multiple photos they have taken of a single individual).

The Software is particularly helpful in obtaining accurate facial measures from small moving targets like toddlers. The photo not only renders the moving target motionless, but a toddler is far more likely to let you approach them with a camera than a PFL ruler. While it may prove challenging at times to take a properly aligned photo of a toddler on the move, there are a number of tricks that will help you achieve this. Conduct the photo session in a small quiet room. Take multiple photos to ensure capture of the eyes and lips in proper repose. Keep in mind the eyes can be measured from one photo and the lips from another photo, if both could not be captured properly in a single photo. And if all else fails, simply set your camera to video mode, record 15-20 seconds of video, and capture the single frame or two where the facial features are in proper repose. Perhaps the greatest advantage of the photo over direct measure is the photo provides a permanent record which will prove invaluable for medical and research purposes.

CONCLUSIONS

In summary, the FAS Facial Photographic Analysis Software measures the PFL with the same accuracy as a sliding digital caliper, as it was programmed to do. Direct measurement of the PFL with a ruler is highly prone to error, even among clinicians who have measured hundreds of PFLs. Direct measurement of the PFL with a caliper is far too dangerous at any age, and should not be used.

Acknowledgements

The WA FASDPN has been supported over the past two decades by the following organizations: Centers for Disease Control and Prevention (1992-1997); Western Washington Chapter of the National March of Dimes Birth Defects Foundation (1995); Washington State Department of Social and Health Services, Division of Alcohol and Substance Abuse through the passage of Senate Bill SB5688 (1997-present); and the Chavez Memorial Fund (2002-present). Research reported in this publication was also supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number U54HD083091. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The creation of the FASDPN clinical dataset would not have been possible without the extensive clinical efforts and support of the interdisciplinary diagnostic teams and community health/social service agencies across WA State. And finally, special thanks are extended to the patients and their families for their benevolent contributions to the WA FAS DPN dataset.

Corresponding Author: astley@uw.edu

REFERENCES

1. Astley SJ. Validation of the fetal alcohol spectrum disorder (FASD) 4-Digit Diagnostic Code. *J Popul Ther Clin Pharmacol* Vol 20(3):e416-467; November 15, 2013.

2. Astley SJ. Diagnostic Guide for Fetal Alcohol Spectrum Disorders: The 4-Digit Diagnostic

Code. 3rd ed. Seattle: University of Washington Publication Services; 2004.

3. Astley SJ, Clarren SK. 1997 Diagnostic Guide for Fetal Alcohol Syndrome and Related Conditions: The 4-Digit Diagnostic Code, 1st edition. University of Washington Publication Services, Seattle, WA.

4. Astley SJ, Clarren SK. Diagnosing the full spectrum of fetal alcohol exposed individuals: Introducing the 4-Digit Diagnostic Code. *Alcohol Alcohol* 2000;35:400-410.

5. Stromland K, Chen Y, Norberg T, Wennerstrom K, Michael G. Reference values of facial features in Scandinavian children measured with a range-camera technique. *Scand. J Plast Reconstr Surg. Hand Surg* 1999;33:59-65.

6. Astley SJ. FAS Facial Photographic Analysis Software [computer program]. Version 1.0. Seattle: University of Washington; 2004.

7. Astley SJ. FAS Facial Photographic Analysis Software [computer program]. Version 2.0. Seattle: University of Washington; 2012.

8. Astley S, Clarren S. A fetal alcohol syndrome screening tool. *Alcohol Clin Exp Res* 1995;19(6):1565-1571.

9. Astley SJ, Clarren SK. A case definition and photographic screening tool for the facial phenotype of fetal alcohol syndrome. *Journal of Pediatrics* 1996;129:33-41.

10. Avner M, Henning P, Koren G, Nulman I. Validation of the facial photographic method in fetal alcohol spectrum disorder screening and diagnosis. *JFAS Int* 2006;4:e20-October 10, 2006.

11. Avner M, Henning P, Koren G, Nulman I. Validation of the facial photographic method in fetal alcohol spectrum disorder screening and diagnosis. *J Popul Ther Clin Pharmacol* Vol 21(1):e106-e113; March 6, 2014.

12. Cranston M, Mhanni A, Marles S, Chudley A. Concordance of three methods for palpebral fissure length measurement in the assessment of fetal alcohol spectrum disorders. *Can J Clin Pharmacol* Vol 16 (1) Winter 2009:e234-e241; April 16, 2009.

13. SPSS 14.0 for Windows. SPSS Inc, 2006.