Review of Psychology, 2016, Vol. 23, No. 1-2, 39-49
International Journal of Croatian Psychological Association published by Naklada Slap

UDC 004.9:159.922.1
Original scientific paper

DOI: 10.21465/rp0023.0004

Analysis of different sources of measurement error in determining second-to-fourth digit ratio, a potential indicator of perinatal sex hormones exposure

UNA MIKAC, VESNA BUŠKO, WERNER SOMMER and ANDREA HILDEBRANDT

Brain structures change in a consistent way due to perinatal and pubertal sex hormones exposure. Data on these changes and their influence on human behaviour, called organisational effects of sex hormones, are difficult to obtain. One potential indicator is the second-to-fourth digit ratio (2D:4D). Measurement of 2D:4D has recently gathered support based on evidence concerning its validity. Available research on its reliability indicated that software methods of determining finger length are more reliable than direct measurements and that inter-rater reliability varied .60–.90. The present research focused on both of these aspects simultaneously, the retest reliability of software measures, and the comparison of software methods using a balanced design. Two scans, each of left and right hands of 213 participants, were collected following methodological recommendations from previous research. Length estimates followed a balanced design with at least six ratings per participant, varying with regard to scan, software method (GIMP or AutoMetric) and the number of raters. At least two raters were involved for each participant. Reliability analysis mostly indicated strong concordance between raters, methods, and scans. We discuss implications for assessing multiple sources of measurement error on continuous rating measures in general, and for measuring 2D:4D in special.

Key words: 2D:4D measurement, reliability, rater effects, software

Human behaviour can be studied on multiple levels, one of them being the interaction of biochemical processes in the body in relation with observable behaviour. Hormones are assumed to moderate those processes and there is accumulated evidence on their linkage with multiple behaviours. The relation of hormones and behaviour is usually described by organisational and activational effects (Neave, 2008). So called organisational effects of sex hormones are mostly irreversible changes in cortical and subcortical structures. Most of them take place during prenatal development, although further developmental phases of intensive hormone secretion seem to be of importance as well, especially during puberty (Sisk & Zehr, 2005). Another way how hormones operate has been referred to as activational effects. They are due to hormonal fluctuations and describe the relationship of the behaviour with sex-hormone blood concentration levels and their natural fluctuations. While there are multiple approaches for studying activational effects (e.g., taking blood samples or observing the behavior during natural hormonal fluctuations like menstrual cycle), the investigation or organisational effects is more challenging and accordingly less developed. Some researcher consider hormone levels in amniotic fluid as one of the most direct indicators of organisational effects (Lutchmaya, Baron-Cohen, Raggatt, Knickmeyer, & Manning, 2004). However, their measurement can hardly be obtained, because it is expensive and require a research design spanning over decades. Except methodological issues with using these measurements, Cohen-Bendahan, van de Beek, & Berenbaum (2005) pointed also to a theoretical challenge due to the fact that the relation of hormone levels in amniotic fluid to hormone levels in foetus blood, which are the variables of interest, is still not determined.

Una Mikac, Department of Psychology, Faculty of Humanities and Social Sciences, Ivana Lučića 3, 10000 Zagreb. E-mail: umikac@ffzg.hr (the address for correspondence);

Vesna Buško, Department of Psychology, Faculty of Humanities and Social Sciences Zagreb;

Werner Sommer, Department of Psychology, Humboldt-Universität zu Berlin;

Andrea Hildebrandt, Department of Psychology, Ernst-Moritz-Arndt-Universität Greifswald.

What is the second-to-fourth-digit ratio?

Second-to-fourth-digit ratio (2D:4D), that is, the ratio of the length of the second to forth finger, has been suggested as one of the rare non-invasive and easily accessible indica-

tors of organisational effects of sex hormones during prenatal development (Hönekopp, Bartholdt, Beier, & Liebert, 2007). Other indicators, such as otoacoustic emissions and dermatoglyphics (Cohen-Bendahan et al., 2005) have been also considered, but they did not gather much data on validity to support their usefulness. Evidence for the use of 2D:4D as indicator of organisational effects of hormones is various and equivocal (McIntyre, 2006). The rationale for use of this indicator comes from the observation that males show a lower 2D:4D than females. However, this difference is not large (d = .28-.35) and smaller than the difference in prenatal testosterone levels (d = 1.4; Hönekopp & Watson, 2010). It has been additionally observed that a significant part of prenatal digit development occurs when testosterone is at one of its peak levels (Vaillancourt, Dinsdale, & Hurd, 2012), and the relation of sex hormones and bone growth has already been established in research on mammals (Kondo, Zákány, Innis, & Duboule, 1997). A stronger evidence for the use of 2D:4D as indicator of organisational effects is the reported negative correlation of right-hand 2D:4D to the ratio of testosterone and oestrogen in the amniotic fluid (r^2 ~ .20; Lutchmaya et al., 2004). However, this finding should be interpreted with caution, since the relation of sex-hormone levels in amniotic fluid to levels in foetus blood are not well established (Cohen-Bendahan et al., 2005). This finding further alerts to caution with respect to the possible interpretation of 2D:4D as indicator of prenatal sex-hormone levels (e.g., Cohen-Bendahan et al., 2005) as opposed to its interpretation as an indicator of prenatal testosterone levels (e.g., McIntyre, 2006).

Research from the field of genetics supports the use of 2D:4D measurement. Individuals with congenital adrenal hyperplasia, a syndrome involving exceptionally high prenatal androgens exposure, have lower 2D:4D than same-sex individuals not suffering from the syndrome (Brown, Hines, Fane, & Breedlove, 2002). This stresses the interpretation related to the importance of testosterone for 2D:4D development. Individuals with androgen insensitivity syndrome who are genetically male and have androgen levels typical for males, but do not react to androgens due to dysfunction of androgen receptors, show higher 2D:4D which is characteristic for females (Berenbaum, Bryk, Nowak, Quigley, & Moffat, 2009). Additive genetic effects on 2D:4D explain about 60% of variance, while non-shared environmental effects, which include different prenatal environment, explain 20-50% of 2D:4D variation across persons (Gobrogge, Breedlove, & Klump, 2008; Voracek & Dressler, 2009). Different gene polymorphisms have been suggested and considered as candidate gene for 2D:4D, namely polymorphisms in or close to the SMOC1, TA polymorphism in ESR1, and CAG/GGN repeat polymorphisms in AR (Lawrance-Owen et al., 2013; Vaillancourt et al., 2012; Zhang et al., 2013, respectively). Research on mice has shown that sex-organs and digits development are influenced by same genes: HoxA and HoxD (Kondo et al., 1997).

Although 2D:4D is considered an indicator of prenatal environment, in the light of the present literature it seems to be more precise to consider it an indicator of perinatal environment. There are three stages during development in boys when testosterone reaches levels similar to those in adult men: (a) during 10. to 18. week of prenatal development, (b) one to two weeks after birth, and (c) from eight weeks until four to six months of age (McIntyre, 2006). Sex differences in 2D:4D are noticeable already at the end of the first trimester of prenatal development (Malas et al., 2006, cited in Hönekopp et al., 2007), but become relatively stable after five years of age and do not change during puberty. These observations endorse the view of 2D:4D as indicator of perinatal testosterone level. In accordance, we use the term *perinatal* in the present manuscript. Also, 2D:4D is not related to adult testosterone levels, which further supports the view on 2D:4D as being an indicator of perinatal testosterone level (Hönekopp et al., 2007). Manning et al. (2000) explored 2D:4D in different populations and ethnic groups, and sex differences were evident in all of the inspected populations. However, population and ethnic differences explained more of 2D:4D variance than sex differences themselves, which is a finding that has not yet been fully understood.

The above summarized findings demonstrate that although the association between 2D:4D and perinatal hormone levels is not strong, there is consistent evidence on its empirical verity and as such it can be used to investigate the strength of the relationship between behavioural manifestations of personality and perinatal hormone levels (Hönekopp & Watson, 2010). Thus, 2D:4D has shown significant correlations with cognitive abilities, sexual orientation, aggressive and assertive behaviour, and sport skills, although effect sizes are generally small (Kemper & Schwerdtfeger, 2009). Wacker, Mueller, and Stemmler (2013) have shown that this small and in some studies non-significant effect might be rather a product of low specificity of the personality measures used. The authors show that using measures of high trait specificity would allow new conclusions on the relation of 2D:4D to certain personality traits. Besides the specificity of the behavioural measurement, the precision and reliability of the 2D:4D measurement is a crucial prerequisite for investigating its validity.

The measurement of 2D:4D

When measuring the 2D:4D, there are multiple challenges to be taken into consideration. First, a choice between direct measurements of finger length or indirect measurements of scanned hands needs to be made. Some researchers also used x rays (e.g., Robertson et al., 2008); however, this third method is neither well studied nor easily accessible and will not be discussed here in more details. Length of the finger is defined as the distance from ventral proximal crease at the bottom of the finger to the finger tip and is determined by use of callipers or rulers. Direct meas-

DOI: 10.21465/rp0023.0004

urements by callipers and indirect measurements of scanned hands correlate r = .65-.75, although the indirect measurements are usually more reliable and result in lower 2D:4D and stronger sex differences (difference in effect size Δd = .13; Hönekopp & Watson, 2010; Manning, Fink, Neave, & Caswell, 2005; Manning, Fink, Neave, & Szwed, 2006).

However, the main challenge of indirect measurements is the unknown amount of pressure of the fingers' soft tissue on the area being scanned. This pressure results in finger length extension and is most probably the factor that leads to both lower 2D:4D and stronger sex differences. Length extension seems to be disproportional for the second and fourth finger and thus, it leads to a systematic decrease of digit ratios when using indirect measurements. The stronger sex differences that emerge when using indirect measurements as compared to those emerging when using x rays and direct methods may be the result of sex differences either in the amount of soft tissue or of differences in pressure applied due to compliance with the instruction, strength or other factors (Berenbaum et al., 2009). The crucial issue on the quality of indirect measurements is the question of where these differences in soft tissue or pressure stem from. If factors, other than perinatal sex hormone levels, are related to them, it could mean that using indirect measurement will result in capturing more construct irrelevant variance than using direct measurements (Hönekopp & Watson, 2010). It is therefore important to even the pressure participants dispense during hand scans. In order to achieve this, the arm position must be controlled since certain positions can result in uneven elongation of the second and the fourth digit. Mayhew, Gillam, McDonald, and Ebling (2007) recommend scanning participants with the hand being at the same levels with the arm. Further, the wrist should be unbended in such a way that the line drawn through the third digit passes between radius and ulna, two main bones of the forearm.

Once scanned, the picture of the hand can be processed in multiple ways. For example the image can be printed and the length can be measured with callipers. However, the use of digital callipers results in highest inter-rater reliability (Allaway, Bloski, Pierson, & Lujan, 2009). This digital method also allows the raters to adjust picture contrast to better determine the start and end point of the finger and it requires shorter measurement time (Allaway et al., 2009). To control for possible differences in picture resolution, Kemper and Schwerdtfeger (2009) recommended the use of a mark of a standardized length on the scanner during the scan, i.e., next to the hand, and by adjusting the contrast and the resolution in such a way that enables transformation of pixels to desired length unit easily. Voracek, Manning, and Dressler (2007) recommended marking the points used for determining length on the participant's hand with a marker before scanning. The authors further suggested covering the hand with foil in order to achieve a higher contrast, while Hiraishi, Sasaki, Shikishima, and Ando (2012) recommended a white cloth to the same purpose.

Finally, there are observable differences in ratios of the right and the left hand, but there is no clear consensus on which hand to prefer for research. The ratios on two hands correlate across persons about r = .65 (Manning et al., 2006). Right hand ratio shows larger sex differences (Δd = .13), larger differences due to congenital adrenal hyperplasia, as well as significant correlation with amniotic fluid hormone levels (Hönekopp & Watson, 2010). These findings suggest right hand can be preferred in research. Similar inference stems from research by Mayhew et al. (2007) who demonstrated that left hand 2D:4D changes due to menstrual cycle, although to a very small amount (3-4%). However, this result should be interpreted in line with methodological and theoretical limitations described in more detail in Mayhew et al. (2007). However, right 2D:4D shows lower correlations with different behaviours (Wacker et al., 2013), which might suggest using the left hand ratio as the perinatal hormone level indicator. Importantly, Putz, Gaulin, Sporter, and McBurney (2004) claimed for caution against deciding post-hoc on the hand to be used for studying relevant relationships, because of obviously higher chance of Type I error in case of post-hoc decisions.

Comparison of multiple raters and multiple measures of one rater mostly indicated reliability above .90 (e.g., Lujan, Podolski, Chizen, Lehotay, & Pierson, 2010; Van der Bergh & Dewitte, 2006). However, Voracek et al. (2007) and Manning et al. (2006) observed lower values (.60 - .70) when comparing different research groups and recommended to calculate 2D:4D based on multiple length measurements. Van Dongen (2009) has compared multiple scans on a small sample and reported a correlation above .97. In their methods comparison, Kemper and Schwerdtfeger (2009) used two softwares, Adobe Photoshop and AutoMetric, and reported higher inter-rater agreement than after use of plastic ruler and vernier caliper. However, due to a non-balanced design implemented by Kemper and Schwerdtfeger (2009) a firm conclusion on the least demanding and most effective method cannot be made.

The present study

We aimed to take into consideration the possible sources of measurement error that have been separately encountered in previous research: raters, software methods, and pressure. We decided to study rater effects because, although research on raters' effects mostly showed high agreement, there was some inconsistency identified in the literature (Manning et al., 2006; Voracek et al., 2007). Following the research by Kemper and Schwerdtfeger (2009), we used indirect measurement because they showed highest inter-rater agreement. We combined the factors in a balanced design in order to provide more precise recommendations on economical utility of different measurement conditions. We collected two scans of each hand, in between which the hands were shortly lifted, in order to assess how much difference in pressure

DOI: 10.21465/rp0023.0004

exist when the instruction is kept constant. Finally, because at present there are no firm conclusions if left and right hand measures differ due to factual, systematic differences, we aimed to estimate the difference between hands and study how hand side interacts with other sources of measurement error.

METHOD

Sample

The data was collected as a part of a larger study conducted at the Humboldt University in Berlin (Kaltwasser, Hildebrandt, Wilhelm, & Sommer, 2016). There were 213 participants, 50% female, with the mean age of 27.8 years (ranging 17 to 40) of heterogeneous levels of education (24% elementary & high school, 42% students, 34% university graduates).

Measurement of 2D:4D

The measurement procedure followed methodological recommendations from previous research described above[1]. A see-through foil was placed on the scanner for each participant. This served a hygienic purpose and more importantly, it assured that a standard length was present on all scans because a ruler of standard length was printed on it (Kemper & Schwerdtfeger, 2009). Before scanning, the proximal crease was marked with a water-soluble marker as to ease the determination of ventral proximal crease (Voracek et al., 2007). Participants were instructed to press lightly with both hands at the same time. The experimenter controlled participants followed this instruction and checked that their hand position was in accordance to the guidelines provided by Mayhew et al. (2007). As suggested by Hiraishi et al. (2012), white cloth was put on the hands by the experimenter in order to achieve more contrast and an easy determination of points on the scanned pictures. After scanning participants lifted their hands and then put them on the scanner again to repeat the procedure. The resolution was kept standard for the collected 426 scans of both hands.

Six raters estimated the length of the fingers on the two scans using two open source software, AutoMetric (DeBruine, 2004) and GIMP (Kimball, Mattis, & GIMP Development Team, 2008). The second software is similar to Adobe Photoshop. Raters were psychology students with no previous experience with 2D:4D measurement. They first participated on a training conducted in the group. Subsequently, every individual rater practiced on ten scans using both softwares. Results were then checked for consistency and the encountered issues have been discussed individually. The raters were instructed to work in batches and rest when needed in order to avoid biased measurement due to fatigue. Their task, besides estimating finger length as described below, was also to mark scans where it was evident that too much pressure leads to white fingertips. Additionally, they noted the time required for measuring a batch of scans and any observation they might have on given scans.

Most papers using measurements of 2D:4D declare to have been estimating the finger length from the tip to the proximal crease. However, to our knowledge, there is no standardized procedure how to determine the exact points from which to measure. While the tip of the finger can easily be defined as a point, the proximal crease is a line. Choosing different points along this line can lead to somewhat different estimations of the finger length, which in turn, due to rather small differences in length between the second and fourth finger can lead to different 2D:4D estimations. In AutoMetric the rater can only choose the middle point of the proximal crease based on his/her own estimation. For GIMP we developed a detailed method for establishing the middle point of the proximal crease[2]. This point was determined as the point splitting in half the straight line connecting the intersections of proximal crease and finger edges on the scan. After determining the tip and the crease point, rater measured the length of the line between them. In AutoMetric the length was automatically measured and saved by the software. In GIMP raters manually inserted them in an electronic protocol after reading it from the software window.

Rating design

We conceived a rating design which allows comparing how much error can be attributed to different sources of measurement error. We planned our rating design by following three goals. The first essential goal was to make comparisons possible while one of the factors varied and all others were held constant. Second, we required that each parameter was calculated on at least about 50 participants. Third, we limited individual rater's workload to about 12.5 h in order to avoid biased measurement due to fatigue. To achieve these goals, we randomly divided the sample into four equally sized subsamples. Four randomly chosen raters estimated finger length of both hands using the two scans and both methods. One rater was allocated to one of the subsamples (see Figure 1). This allowed for comparison of methods and scans with raters being kept constant. Other two raters measured the fingers in all four subsamples, but each with a different method and using only one of the two scans (see Figure 1). This allowed for comparison of raters while keeping method and scan constant. Each cell in Fig-

1 The protocol can be obtained from the first author on request.

2 The protocol can be obtained from the first author on request.

DOI: 10.21465/rp0023.0004

	Subsample of participants			
	1	2	3	4
Ratings	$R_1M_1S_1$	$R_2M_1S_1$	$R_3M_1S_1$	$R_4M_1S_1$
	$R_1M_1S_2$	$R_2M_1S_2$	$R_3M_1S_2$	$R_4M_1S_2$
	$R_1M_2S_1$	$R_2M_2S_1$	$R_3M_2S_1$	$R_4M_2S_1$
	$R_1M_2S_2$	$R_2M_2S_2$	$R_3M_2S_2$	$R_4M_2S_2$
	$R_5M_1S_1$	$R_5M_2S_1$	$R_5M_1S_2$	$R_5M_2S_2$
	$R_6M_2S_1$	$R_6M_1S_1$	$R_6M_2S_2$	$R_6M_1S_2$

Figure 1. Rating design showing which raters measured a given scan by using a given method on four equally sized subsamples. $R_{1\text{-}6}$ = one of the six raters; $M_{1\text{-}2}$ = method used (GIMP or AutoMetric); $S_{1\text{-}2}$ = one of the two available scans.

ure 1 represents about 212 estimations of finger lengths, i.e., two fingers on each of two hands for about 53 participants.

Half of the raters in each subgroup of raters (4+2) used GIMP first, and the other half started with AutoMetric. The order of scans in each subsample was randomized mixing the order of the participants and of their two scans, thus the two scans of the same person were not rated one after another. Scans were renamed in order to keep raters blind to the number and order of scans belonging to the same participant. The left and the right hand were scanned together and as described above they appeared on the same scan. Finger length of the two hands depicted on the same scan were always measured consequently following the same order, first right fourth and second finger and then left second and fourth finger.

RESULTS

All statistical analyses were conducted using SPSS 19.0 and Microsoft Excel 2007. Finger length of second digit was divided with the length of the fourth digit to calculate 2D:4D for (a) each hand, (b) for each scan, (c) for each method, and (d) for each rater, leading to 2543 ratios. All together, 5086 finger lengths were estimated on 1275 scans of right hands and 1268 of left hands. This number of estimations does not correspond to the number of planned estimation. Two scans were corrupted which resulted in two estimation less for the right hand (intended for two raters) and two estimations less for the left hand (originally intended for estimation by two different raters). All four were intended for estimation with AutoMetric. Two left hands were accidentally skipped by the same rater when using AutoMetric. One participant had her left hand in cast which resulted in five ratios less for the left hand.

As a preliminary analysis, we inspected skewness and kurtosis. Together with a visual inspection of the distributions we can conclude that 2D:4Ds on all subsamples follow an approximately normal distribution.

Subsequently, two sets of analysis were performed on the estimated ratios. First, Pearson correlations were calculated on each of the subsamples. These gave information on reliability, i.e., the amount of error that is present when the only difference between two variables is the source of error of interest for that given analysis. The correlations are presented in a way that allows us to discuss reliability estimates when a factor is held constant (see Table 1). Left and right hands showed similar trends, thus correlations calculated on both hands are presented together.

Second, average 2D:4D levels were compared across different factors (Table 2). The significance level of .05 was adjusted according to Bonferroni correction and $p < .0006$ was considered significant. In all analyses, sex was included because sex differences are expected for the 2D:4D. For four raters that measured both scans with both methods (R_1 to R_4 in Figure 1) we performed a five-way mixed type 4 (raters) × 2 (sex) × 2 (methods) × 2 (scans) × 2 (hand side) ANOVA. The size of the subgroups varied from 17 to 34.

Table 1
Pearson correlations of 2D:4D measured by different raters using different methods on two scans of left and right hand (N = 213)

Constant factor	Varying factor	M_r	k	Min	Max	n
M_1 (G)	S	.89	8[a]	.57	.96	48-54
S_1	M	.91	8[a]	.88	.95	46-54
M_2 (AM)	S	.93	8[a]	.89	.96	46-54
S_2	M	.87	8[a]	.57	.95	46-54
R_1	R	.92	4[b]	.90	.94	49
R_2	R	.94	4[b]	.92	.96	43-45
R_3	R	.92	4[b]	.89	.94	47
R_4	R	.89	4[b]	.85	.95	51
R_5	R	.92	8[c]	.88	.96	43-51
R_6	R	.91	8[c]	.85	.95	44-51

Note. All correlations are significant at $p < .001$. S = scan of hand; M = software method used; G = GIMP; AM = AutoMetric; R = raters; M_r = average correlation; k = number of correlations calculated; Min/Max = largest/smallest correlation; n = size of subsamples correlations were calculated on; [a] calculated on four subsamples for left and right hand (with raters constant on each of the subsamples); [b] relation with other raters on one of the subsamples for left and right hand (with the method and scan kept constant); [c] relation with other raters on one of the subsamples calculated on four subsamples for left and right hand (with the method and scan kept constant).

DOI: 10.21465/rp0023.0004

Table 2
Means (and standard deviations) of 2D:4D of female (f) and male (m) participants in the four subsamples for the left and the right hand (N = 213)

	Subsample							
	1		2		3		4	
Rating factor	f [a]	m [b]	f [c]	m [d]	f [e]	m [f]	f [g]	m [h]
	Left hand							
M_1S_1	0.982 (0.039)	0.984 (0.039)	0.987 (0.041)	0.974 (0.033)	0.983 (0.031)	0.983 (0.04)	0.974 (0.031)	0.973 (0.033)
M_1S_2	0.984 (0.032)	0.98 (0.039)	0.986 (0.036)	0.975 (0.031)	0.982 (0.029)	0.979 (0.04)	0.977 (0.031)	0.975 (0.029)
M_2S_1	0.986 (0.04)	0.987 (0.043)	0.986 (0.037)	0.976 (0.032)	0.975 (0.028)	0.977 (0.039)	0.988 (0.028)	0.982 (0.029)
M_2S_2	0.984 (0.031)	0.98 (0.043)	0.985 (0.037)	0.979 (0.029)	0.976 (0.029)	0.975 (0.04)	0.989 (0.028)	0.981 (0.027)
R_5	0.994 (0.036)	0.989 (0.043)	0.983 (0.039)	0.971 (0.03)	0.974 (0.03)	0.978 (0.04)	0.991 (0.03)	0.974 (0.025)
R_6	0.987 (0.038)	0.981 (0.041)	0.996 (0.041)	0.976 (0.027)	0.97 (0.031)	0.974 (0.039)	0.993 (0.03)	0.987 (0.024)
	Right hand							
M_1S_1	0.99 (0.028)	0.981 (0.039)	0.983 (0.037)	0.968 (0.026)	0.997 (0.031)	0.988 (0.042)	0.966 (0.029)	0.979 (0.04)
M_1S_2	0.989 (0.03)	0.968 (0.031)	0.983 (0.031)	0.966 (0.025)	0.993 (0.03)	0.986 (0.042)	0.968 (0.024)	0.983 (0.042)
M_2S_1	0.99 (0.028)	0.98 (0.038)	0.976 (0.032)	0.963 (0.022)	0.985 (0.03)	0.973 (0.041)	0.979 (0.024)	0.986 (0.045)
M_2S_2	0.991 (0.025)	0.978 (0.041)	0.979 (0.03)	0.963 (0.023)	0.984 (0.029)	0.973 (0.041)	0.979 (0.023)	0.986 (0.044)
R_5	0.995 (0.031)	0.981 (0.043)	0.981 (0.033)	0.972 (0.026)	0.982 (0.031)	0.977 (0.044)	0.977 (0.022)	0.981 (0.052)
R_6	0.996 (0.032)	0.983 (0.043)	0.99 (0.031)	0.976 (0.023)	0.981 (0.029)	0.981 (0.044)	0.984 (0.022)	0.993 (0.045)

Note. The table follows the rating design presented in Figure 1. Rating factor column represents the factor(s) constant for the whole row. Raters 1 to 4 are constant for each sample for first four rows of each hand subsection, and method and scan vary across samples for the last two rows of each hand subsection in accordance with Figure 1 and *Rating design* section. M_1 = GIMP; M_2 = AutoMetric; S_{1-2} = one of two scans; R_{5-6} = Raters 5 and 6.

[a] n = 34. [b] n = 15-16. [c] n = 25-28. [d] n = 18-21. [e] n = 18-20. [f] n = 29-30. [g] n = 20-21. [h] = 31-33.

Only the interaction of methods and raters was significant, $F(3, 205) = 23.543$, $p < .001$, $\eta^2 = .255$, with two of the raters having same ratios when using different methods, one having a higher ratio when using GIMP, and one when using AutoMetric.

The remaining two raters were included in order to allow comparisons of rater effects, therefore they were compared with the rater who rated the same scan with the same method on the same sample (e.g., on Subsample 1 we compared $R_1M_1S_1$ to $R_5M_1S_1$and $R_1M_2S_1$ to $R_6M_2S_1$). Due to data missing per design, this resulted in eight three-way mixed-type 2 (sex) × 2 (raters) × 2 (hand side) ANOVAs. The analysis showed that in two of eight analyses there were significant differences between raters, $F(1, 42) = 23.617$, $p < .001$, $\eta^2 = .360$ (GIMP, Scan 1, Subsample 2), $F(1, 49) = 63.797$, $p < .001$, $\eta^2 = .547$ (GIMP, Scan 2, Subsample 4). In both of these analyses the same rater's estimations led to larger ratios, in both cases when using GIMP. The main effects of hands and sex were not significant. In one of the analyses significant interaction showed that for one rater the ratio of the two hands was equal and for the other the right ratio was smaller, $F(1, 41) = 17.467$, $p < .001$, $\eta^2 = .299$ (AutoMetric, Scan 1, Subsample 2).

Although there were no significant effects of sex or hand side, we calculated the average effect size of the sex difference in order to be compared with existing data. This was calculated on each of the four subsamples for each hand for each combination of rater, method, and scan (altogether 48 comparisons, see Figure 1). Average effect size of the sex difference was $d = 0.182$ (females: $M = 0.984$, $SE = 0.0052$;

DOI: 10.21465/rp0023.0004

males: $M = 0.978$, $SE = 0.0065$), ranging from $d = -0.447$ to $d = 0.673$, with 25% of the comparison showing the unexpected trend of larger 2D:4D for men. Average difference in effect size of the sex difference between left and right hand was $\Delta d = -0.046$, ranging from $\Delta d = -0.553$ to $\Delta d = 0.737$.

As for other information collected by raters, the pressure evident in white fingertips occurred in 15.26% of the scans. Pressure was present on only one of the scans for 13.15% of the participants, and on both scans for 4.22% of the participants. Performing t-tests revealed no significant differences between ratios where pressure was present and those where it was not present. This was true for ratios estimated by any combination of method, scan, and rater.

Average time spent was three and a half minutes per hand (two digits) when using GIMP, and half a minute per hand when using AutoMetric. Raters encountered issues that might have influenced the precision, although still made the estimation possible, on 151 out of 426 scans. A high number of scans, 26.06% challenged the estimations due to moisture. This made the scans unclear and the identification of points more difficult. Other challenges were due to invalid scanning mistakes, such as non-removed hand jewellery (2.4%), hidden creases by markers (1.6%), markers on wrong fingers (1.6%), and wrong hand position (1.2%); and anatomical issues including unclear or very curved creases (1.4%) and very curved fingers (1.2%).

DISCUSSION

Summary of findings

The analysis of sources of error in 2D:4D measurement revealed that it has the potential to be a highly reliable measure if strict recommendations for data collection are adhered to. All of the correlations indicated strong concordance, i.e., all of the sources taken into account do not contribute much to error. Only one of raters showed lower estimations of reliability, but only while using one of the methods, as demonstrated by minimums in Table 1. The method in question was GIMP. Although we gave our best to make the procedure more precise with introducing an exact way to pinpoint the middle of the crease, this complex procedure seemed to be more prone to mistakes, contributing to more variable estimates. Another factor that might have contributed to additional error is the fact that raters had to write down the estimations manually from GIMP, while in AutoMetric both the length estimations and ratio calculation was automatically done by the software. More error is also evidenced by significant differences between raters that appeared only when GIMP was used.

Inter-rater reliability was high and in accordance with previous research. It should be noted that all reliability estimates less than .80 were due to the same rater (both $r(48) = .57$, $p < .001$). This suggests that raters contribute differently to error and in studies involving only one rater there is risk of providing non-reliable estimations. Significant differences between raters in the average ratios also suggest that involving more than one rater is advisable in studies using 2D:4D measurement. Importantly, in most cases differences were evident when GIMP was used. According to our data, using AutoMetric can alleviate the differences between raters. However, this findings needs to be replicated with larger sample size. Additionally, it should be taken into account that the use of AutoMetric in our research also revealed significant interactions of raters with hand side and with method.

There was high concordance and no difference in ratios between two scans. This suggests that the scans include the same information and that the pressure could be successfully controlled by instruction. We did not control the pressure by precise measurement; we only noted the cases in which the pressure was so strong that it led to white fingertips. This is obviously not a very precise measure, since white fingertips could also depend on other factors, such as blood flow and outside temperature. However, we are confident that white fingertips indicate that the participant did not just lightly laid hands on the scanner. Although for a small amount of participants white fingertips indicating pressure were observed on one scan and not the other, it seems that this did not influence the ratios themselves. We conducted some detailed analyses in this regard, which did not show any differences in ratios where white fingers were present and where they were not. Due however to small samples where pressure was obviously visible ($n = 15$-24), these results should be treated with caution, because the low number of observations may have lead to low statistical power of the difference test. While the findings demonstrate that pressure is mostly constant for a person, available measurements do not allow us to determine the relation of the pressure to perinatal sex hormone levels. A design in which pressure is manipulated experimentally and measured with more precision might be of use, but identifying a proper control variable to test the relationship of the different ratios still remains an issue.

Our data suggest no differences in the amount of error when estimating the left or the right hand ratio. We did not confirm the previously reported difference in effect size, which in our study was close to zero as compared to Δd of .13, reported by Hönekopp and Watson's (2010). The difference between left and right hand still remains predominantly a validity issues requiring further research of both hands' relationships with relevant variables and longitudinal research exploring their relationship to characteristics of perinatal environment.

The workload analysis showed that time requirements when using GIMP are seven times larger. This is an average time, based on time needed for unequal batches. Thus, it is a rough estimation which does not allow further analy-

sis, e.g., on the learning curve. For more experienced raters this difference in times for different softwares might vary in size, but due to more complex procedure GIMP will surely require more time. Although in small samples this might not be of importance since using both software methods requires a few minutes, in large samples this factor will need to be taken into account.

The notes made during the estimation pointed to practical challenges that can be encountered. One of the biggest issues was the moisture which made the scans unclear (26.06%). Probably high outside temperatures led to more intensive sweat production. Another set of problems was caused by marking the creases on the hands (3.2%). Although this procedure made finding the proximal crease easier, it also caused some of the problems by obscuring the midpoint of the crease. In some instances jewellery was not removed (2.4%) which might have led to wrong hand position and thus wrong length estimation. Curved fingers were also a problem because the straight line whose length was estimated cannot be considered a good represent of the finger length (1.2%). Similar problem was caused by curved creases when using GIMP because midpoint was defined as a centre of a straight line connecting intersections of crease and fingers (1.4%). Because of low incidence of these issues, we included them in the final sample and we did not test if estimations where these issues were encountered differed from other estimations. After the rating process was completed, raters had a group discussion that pointed to some options for dealing with these issues, which we present in the following section.

We tested the sex differences in the 2D:4D in order to compare our study to others in the field. There were no significant differences, and the effect size was relatively small, $d = 0.182$. In their meta-analysis, Hönekopp and Watson (2010, p. 626) suggest that due to considerable heterogeneity, values above $d = 0.22$ for the right hand and above $d = 0.15$ for the left hand can be considered empirically valid. The average difference that we observed can be considered as one of the smaller differences when compared to other in the literature, but as fitting to these previous reports. Some of our calculated values quantifying sex differences were in the opposite direction than expected. This finding was also encountered in the meta-analysis by Hönekopp and Watson (2010). Considering population and ethnic differences in 2D:4D (Manning et al., 2000), we may consider if this small average sex difference is characteristic for German population. However, our sample differs from the German sample in Manning et al. (2000) regarding the level of the mean, the variation, and the size of sex difference. The samples also differ regarding the towns where data was collected. Berlin is multicultural town and that might explain this deviation from the hypothesized German population based on results by Manning et al. (2000), but it also stresses the need for further research on stability of population differences in 2D:4D.

Overall, our rating design allowed us to analyze a large number of factors at the same time without confounds between variables. There is a high number of possible sources of error when measuring 2D:4D, but our findings clearly indicate that many of them can be controlled in a well-planed and precise data collection and estimation procedure. We established AutoMetric as the method of choice and showed that more than one scan does not bring new information. We have confirmed raters are a source of variation, but we have also shown this variation can be diminished by using AutoMetric and careful choice of raters. We demonstrated left and right hand are equally prone to measurement error. These findings are supported by both, the correlation and the mean level analyses.

Practical implications and recommendations

Our paper summarizes practical recommendations for measuring 2D:4D, many already outlined in the introduction, as well as giving some new advices that can serve as guidelines for future research. A well designed protocol, detailed instructions, and strict control during measurement, especially of hand pressure, can remove the need for multiple scans. Marking the proximal crease on hands is helpful, but some of the problems it causes could be avoided by marking just the ends of the crease and leaving the mid of the crease unmarked. This would facilitate determining which crease is proximal without hiding the midpoint. Based on our experience we also advise researchers to be cautious of the problem of sweating and prepare additional materials, like tissues for the hands. Putting the white cloth over the hands, as recommended in previous research, made determining the point of the finger easier, but raters suggested that maybe putting a dark cloth would lead to more contrast, at least when the participant is of light skin tone. This is easy to test in future research by making a few test scans before data collection using different types of covers.

To achieve higher consistency, comparable reliability, and measurement economy, AutoMetric should be preferred over GIMP. GIMP can be used for practice on finding the midpoint of the proximal crease in order to develop a sense of how personal estimations coincide with the mathematical midpoint, but the results clearly suggest that the estimations from AutoMetric have higer quality. Multiple raters are advisable since some seem to be more prone to inconsistency and ratios differ between different raters. Special care should be taken when choosing raters. This choice should be based on a test estimation of a sample of pictures larger than used in our research, followed by detailed individual discussion with the rater on the process of estimation after completion. For some of the participants, which in our sample presented only a small percent, the length of the finger as presented by a straight line will probably be a poor representative of the length of the finger due to curved fingers. However, due to standardization of the process and the complexity of deter-

DOI: 10.21465/rp0023.0004

mining the curve of the finger we suggest using the straight line for determining the length. The crease will also probably be curved for some of the participants, but this does not influence the estimation when using AutoMetric.

A note on methods of analyses

When considering our findings and recommendations, limitations of our research should be taken into account. We followed the described design in order to control confounding effects and rater's fatigue. This resulted however in small subsamples, which in turn limited the statistical power of our tests, as well as our choice of data analysis methods. As for statistical power, it was low when calculating ANOVA, which means there may be differences that remained non-identified. Pearson correlations were compared without testing the statistical significance of such comparisons. Similar to Voracek et al. (2007), our focus was not just on statistical significance, but on measurement error in the first place. This may partly limit the generalizability of our conclusion, specifically the idea that GIMP is more prone to inconsistency.

There are however two aspects of our data that support generalization. First, both, the correlation analyses and the mean level analyses indicated that using GIMP results in inconsistency. Three out of four significant effects in ANOVA were on ratios based on GIMP measurements (one interaction of methods and raters and two main effects of raters). Second, results on research questions we addressed and that were already investigated in previous studies were mostly concordant. These aspects are the overall high reliability, inter-rater reliability, and effect size reported for sex differences. Because of this correspondence, we believe that the new insights and derived recommendations for 2D:4D measurement offered by our research can also be used in future.

Furthermore, our choice of analyses was also determined by sample size, as well as by our rating design, which included multiple factors. Abundant research in this field stems from generalizability theory and uses intraclass correlation to estimate reliability (e.g., Allaway et al., 2009; Kemper & Schwerdtfeger, 2009). This approach allows the partitioning of sources of variability and testing the significance of every of those sources (Zhou, Muellerleile, Ingram, & Wong, 2011). In our research we analyzed three possible sources of error (rater, method, and scans) and two possible sources of systematic variation (hand side and gender), which together with subjects as a source a variation constitutes a six-way model. To the best of our knowledge, this complex model has not yet been researched in detail, although new development is evident in papers considering three-way models in detail (Wong & McGraw, 1999; Zhou et al., 2011). In the case we took into account only the sources of error that were of main interest (raters, method, and scans) and disregard hand side and gender as sources of variation, this still, together with subjects, constitutes in a four-way model. However, now that we have established a preferable method and shown there is no need for multiple scans, future research can focus on two sources of variation less. We believe using intraclass correlation can be especially useful because it would allow the determination of the optimal number of raters (Wong & McGraw, 1999). Although our findings, as well as previous research, indicate multiple raters are needed, to our knowledge a recommended number of raters has not been determined yet.

Another potential approach to decompose sources of variance is structural equation modelling. When using multiple raters, ratios in previous research were mostly expressed as averages, but they could also be expressed as latent variables. In our research we decided to compare multiple sources of variation in a way that allowed us to infer on their interactions, but at the same time resulted in smaller subsamples and multiple missing data due to the design. These factors are hard to encompass with structural equation models. Now that we can recommend using only one method (AutoMetric) and one scan, future research can try to use latent variables composed of ratings of both hands made by multiple raters. That way common variance would be analyzed, which is probably due in part to perinatal hormonal environment. However, researchers should keep in mind estimation problems might occur due to high multicolinearity.

CONCLUSION

Our research has lead to new insights that are relevant and useful when measuring 2D:4D. We showed 2D:4D to be a highly reliable measure and suggested some strategies to diminish the error even more. Our literature review revealed multiple possible sources of error, and our results indicate which of them need to be considered as relevant. It has clearly showed AutoMetric is the method to be used, both because of its high reliability and measurement economy and because it contributes to more reliable estimations by raters. We have also shown that one scan is sufficient, which diminishes the practical costs and simplifies future designs. The left and right hands' ratios do not differ regarding measurement error and the differences between them are probably of a more substantial nature. Therefore, evidence on validity of different hands is needed to develop recommendations concerning the hand side. We consider our research only a step towards the validation of 2D:4D as an indicator of perinatal hormones. There are sources of construct irrelevant variance that mask the relationship of perinatal hormones to variables of interest, and identifying them would aid further validity studies that are needed in this area. We believe that research with models simplified based on our findings can lead to new practical suggestions (e.g., on the

number of raters) and new methodological approaches using latent variables.

REFERENCES

Allaway, H. C., Bloski, T. G., Pierson, R. A., & Lujan, M. E. (2009). Digit ratios (2D:4D) determined by computer-assisted analysis are more reliable than those using physical measurements, photocopies, and printed scans. *American Journal of Human Biology, 21*(3), 365-70. doi:10.1002/ajhb.20892

Berenbaum, S. A., Bryk, K. K., Nowak, N., Quigley, C. A., & Moffat, S. (2009). Fingers as a marker of prenatal androgen exposure. *Endocrinology*, *150*(11), 5119-5124. doi:10.1210/en.2009-0774

Brown, W. M., Hines, M., Fane, B. A., & Breedlove, S. M. (2002). Masculinized finger length patterns in human males and females with congenital adrenal hyperplasia. *Hormones and Behavior, 42*(4), 380-6. doi:10.1006/hbeh.2002.1830

Cohen-Bendahan, C. C., van de Beek, C., & Berenbaum, S. A. (2005). Prenatal sex hormone effects on child and adult sex-typed behavior: Methods and findings. *Neuroscience & Biobehavioral Reviews*, *29*(2), 353-384. doi:10.1016/j.neubiorev.2004.11.004

DeBruine, L. M. (2004). AutoMetric (2.2) [Software for measurement of 2D:4D ratios]. Retrieved from www.facelab.org/debruine/Programs/autometric.php

Gobrogge, K. L., Breedlove, S. M., & Klump, K. L. (2008). Genetic and environmental influences on 2D:4D finger length ratios: A study of monozygotic and dizygotic male and female twins. *Archives of Sexual Behavior, 37*(1), 112-8. doi:10.1007/s10508-007-9272-2

Hiraishi, K., Sasaki, S., Shikishima, C., & Ando, J. (2012). The second to fourth digit ratio (2D:4D) in a Japanese twin sample: Heritability, prenatal hormone transfer, and association with sexual orientation. *Archives of Sexual Behavior, 41*(3), 711-24. doi:10.1007/s10508-011-9889-z

Hönekopp, J., Bartholdt, L., Beier, L., & Liebert, A. (2007) Second to fourth digit length ratio (2D:4D) and adult sex hormone levels: New data and a meta-analytic review. *Psychoneuroendocrinology, 32*, 313–321. doi:10.1016/j.psyneuen.2007.01.007

Hönekopp, J., & Watson S. (2010). Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. *American Journal of Human Biology, 22*(5), 619-30. doi:10.1002/ajhb.21054

Kaltwasser, L., Hildebrandt, A., Wilhelm, O., & Sommer, W. (2016). Behavioral and neuronal determinants of negative reciprocity in the ultimatum game. *Social Cognitive and Affective Neuroscience,* nsw069. doi:10.1093/scan/nsw069

Kemper, C. J., & Schwerdtfeger, A. (2009). Comparing indirect methods of digit ratio (2D:4D) measurement. *American Journal of Human Biology, 21*(2), 188-91. doi:10.1002/ajhb.20843

Kimball, S., Mattis, P., & GIMP Development Team (2008). GIMP 2.6.12 [GNU Image Manipulation Program]. Retrieved from http://www.gimp.org/

Kondo, T., Zákány, J., Innis, J. W., & Duboule, D. (1997). Of fingers, toes and penises. *Nature, 390*(6655), 29. doi:10.1038/36234

Lawrance-Owen, A. J., Bargary, G., Bosten, J. M., Goodbourn, P. T., Hogg, R. E., & Mollon, J. D. (2013). Genetic association suggests that SMOC1 mediates between prenatal sex hormones and digit ratio. *Human Genetics, 132*(4), 415-21. doi:10.1007/s00439-012-1259-y

Lujan, M. E., Podolski, A. J., Chizen, D. R., Lehotay, D. C., & Pierson, R. A. (2010). Digit ratios by computer-assisted analysis confirm lack of anatomical evidence of prenatal androgen exposure in clinical phenotypes of polycystic ovary syndrome. *Reproductive Biology and Endocrinology, 8*, 156. doi:10.1186/1477-7827-8-156

Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., & Manning, J. T. (2004). 2nd to 4th digit ratios, fetal testosterone and estradiol. *Early Human Development, 77*(1-2), 23-8. doi:10.1016/j.earlhumdev.2003.12.002

Manning, J. T., Barley, L., Walton, J., Lewis-Jones, D. I., Trivers, R. L., Singh, D., Thornhill, R., Rohde, P., Bereczkei, T., Henzi, P., Soler, M., & Szwed, A. (2000). The 2nd:4th digit ratio, sexual dimorphism, population differences, and reproductive success. Evidence for sexually antagonistic genes? *Evolution and Human Behavior, 21*(3), 163-183. doi:10.1016/s1090-5138(00)00029-5

Manning, J. T., Fink, B., Neave, N., & Caswell, N. (2005). Photocopies yield lower digit ratios (2D:4D) than direct finger measurements. *Archives of Sexual Behavior, 34*(3), 329-33. doi:10.1007/s10508-005-3121-y

Manning, J. T., Fink, B., Neave, N., & Szwed, A. (2006). The second to fourth digit ratio and asymmetry. *Annals of Human Biology, 33*(4), 480-92. doi:10.1080/03014460600802551

Mayhew, T. M., Gillam, L., McDonald, R., & Ebling, F. J. (2007). Human 2D (index) and 4D (ring) digit lengths: Their variation and relationships during the menstrual cycle. *Journal of Anatomy, 211*(5), 630-8. doi:10.1111/j.1469-7580.2007.00801.x

McIntyre, M. H. (2006). The use of digit ratios as markers for perinatal androgen action. *Reproductive Biology and Endocrinology, 4,* 10. doi:10.1186/1477-7827-4-10

Neave, N. (2008). *Hormones and behaviour: A psychological approach*. Cambridge University Press. doi:10.1017/cbo9780511808203

DOI: 10.21465/rp0023.0004

Putz, D. A., Gaulin, S. J. C., Sporter, R. J., & McBurney, D. H. (2004). Sex hormones and finger length: What does 2D:4D indicate? *Evolution and Human Behavior, 25,* 182–199. doi:10.1016/j.evolhumbehav.2004.03.005

Robertson, J., Zhang, W., Liu, J. J., Muir, K. R., Maciewicz, R. A., & Doherty, M. (2008). Radiographic assessment of the index to ring finger ratio (2D:4D) in adults. *Journal of Anatomy, 212*(1), 42-8. doi:10.1111/j.1469-7580.2007.00830.x

Sisk, C. L., & Zehr, J. L. (2005). Pubertal hormones organize the adolescent brain and behavior. *Frontiers in Neuroendocrinology, 26*, 163–174. doi:10.1016/j.yfrne.2005.10.003

Vaillancourt, K. L., Dinsdale, N. L., & Hurd, P. L. (2012). Estrogen receptor 1 promoter polymorphism and digit ratio in men. *American Journal of Human Biology, 24*(5), 682-9. doi:10.1002/ajhb.22297

Van den Bergh, B., & Dewitte, S. (2006). Digit ratio (2D:4D) moderates the impact of sexual cues on men's decisions in ultimatum games. *Proceedings of the Royal Society B: Biological Sciences, 273*(1597), 2091-5. doi:10.1098/rspb.2006.3550

Van Dongen, S. (2009). Second to fourth digit ratio in relation to age, BMI and life history in a population of young adults: A set of unexpected results. *Journal of Negative Results: Ecology & Evolutionary Biology, 6*, 1–7. Retrieved from http://www.jnr-eeb.org/

Voracek, M., & Dressler, S. G. (2009). Brief communication: Familial resemblance in digit ratio (2D:4D). *American Journal of Physical Anthropology, 140*(2), 376-80. doi:10.1002/ajpa.21105

Voracek, M., Manning, J. T., & Dressler, S. G. (2007). Repeatability and interobserver error of digit ratio (2D:4D) measurements made by experts. *American Journal of Human Biology, 19*(1), 142-6. doi:10.1002/ajhb.20581

Wacker, J., Mueller, E. M., & Stemmler, G. (2013). Prenatal testosterone and personality: Increasing the specificity of trait assessment to detect consistent associations with digit ratio (2D:4D). *Journal of Research in Personality, 47*, 171–177. doi:10.1016/j.jrp.2012.10.007

Wong, S. P., & McGraw, K. O. (1999). Confidence intervals and F tests for intraclass correlation coefficients based on three-way random effects models. *Educational and Psychological Measurement, 59*, 270-288. doi:10.1177/00131649921969848

Zhang, C., Dang, J., Pei, L., Guo, M., Zhu, H., Qu, L., Jia, F., Lu, H., & Huo, Z. (2013). Relationship of 2D:4D finger ratio with androgen receptor CAG and GGN repeat polymorphism. *American Journal of Human Biology, 25*(1), 101-6. doi:10.1002/ajhb.22347

Zhou, H., Muellerleile, P., Ingram, D., & Wong, S. P. (2011). Confidence intervals and F tests for intraclass correlation coefficients based on three-way mixed effects models. *Journal of Educational and Behavioral Statistics*, *36*(5), 638-671. doi:10.3102/1076998610381399