Replicated nil associations of digit ratio (2D:4D) and absolute finger lengths with implicit and explicit measures of aggression

Martin Voracek and Stefan Stieger
University of Vienna (Austria)

Preliminary evidence suggests that within-sex individual variation in the length ratio of the index finger to the ring finger (2D:4D), a putative marker of prenatal androgen levels, may be more strongly or consistently related to implicit measures (Implicit Association Test, IAT) than to corresponding explicit (self-report) measures of candidate personality traits. The underlying logic is that the non-introspective IAT may reflect earlier and inaccessible experiences, whereas introspective self-reports reflect more recent and necessarily accessible experiences. Associations of 2D:4D and absolute finger length (a marker of pubertal-adolescent androgen levels) with implicit versus explicit aggression measures (IAT vs. Buss-Perry Aggression Questionnaire and a feeling thermometer) were examined in two samples with identical procedures (N= 244 and 233). Attesting to procedural validity and data typicality, several experiments of related research were replicated in both samples, including sex differences in 2D:4D and theory compliant interrelations among psychometric measures. However, no theory compliant, reliable, or replicable associations of 2D:4D or finger length with implicit or explicit aggression measures resulted. These nil findings cast doubt on hypothesized advantages of implicitly (over explicitly) measured target traits for 2D:4D research. More generally, they add to a growing number of replication failures in this area.

The rediscovery of sex differences (men < women) in the length ratio of the index finger (2D) to the ring finger (4D) (2D:4D; Manning, Scutt, Wilson, & Lewis-Jones, 1998), along with a hypothesis about the origin of these, has spurred ongoing research interest in psychology and adjacent disciplines. Because this sex effect is determined in utero and 2D:4D shows little plasticity thereafter, 2D:4D has been suggested as an indirect marker for the organizational (permanent) masculinizing effects of prenatal androgens on the brain, behavior, and physique (Manning, 2002).

Given this hypothesis is correct, it would straightforwardly allow for comfortable (non-invasive) tests of the biological bases of behavior, cognition, personality, and other individual difference variables that show sex differences and therefore conceivably are partly influenced by prenatal sex-hormone action. Unsurprisingly, this idea has attracted wide interest, resulting in a sizeable literature now approaching 250 reports (partial reviews: Manning, 2002; Putz, Gaulin, Sporter, & McBurney, 2004).

2D:4D research also has a number of problems, including usually modest effects, preponderantly unreplicated findings, statistically underpowered (and thus type I error-prone) studies reporting positive findings, a spate of recently published nonreplications, and single-sample studies (without replication attempts) still being the norm.
Building on these combined facts, concerns of possible publication bias in 2D:4D research have been expressed (Putz et al., 2004). This possibility seems not remote. Publication bias was clearly demonstrated for fluctuating asymmetry research (Palmer, 2000), a field sharing several features with 2D:4D research (e.g., small effects, measurement and methodology issues, underpowered studies, replication failures, rapid growth of the literature, a few productive labs, and competitive publishing).

To give one example, an initial report (Csathó, Osváth, Bicsák, Kárdi, Manning, & Kállai, 2003), based on merely 46 female Hungarian students, suggested associations of 2D:4D with sex-role orientation. According to ISI Web of Knowledge, this report ranks among the top 15 of most frequently cited papers of 2D:4D research. However, five subsequent reports, including large-scale and multi-sample evidence (Hamson, Ellis, & Tenk, 2008; Lippa, 2006; Rammayer & Troche, 2007; Schmukle, Liesenfeld, Back, & Egloff, 2007; Troche, Weber, Hennigs, Andresen, & Rammayer, 2007), and at least six unpublished theses (Evardone, 2006; Fawcett, 2003; Meingäldner, 2003; Robinson, 2005; Schicker, 2005; Vonnamme, 2005) have failed to replicate such associations. Two further studies with positive findings (Beech & Mackintosh, 2005; Scarbrough & Johnston, 2005) were based on similarly small samples (N= 60 and 41). On the whole, there is more evidence against associations of 2D:4D with sex-role orientation than support for these.

In this context, one study (Schmukle et al., 2007) is of particular importance, because it was hypothesized that implicit (automatic or nonconscious) measures, as opposed to explicit (controlled or conscious) measures, might entertain more reliable associations with 2D:4D. This is because explicit mental representations, as gauged by self-report measures, exploit propositional structures that are accessible through introspection, malleable, and thus reflect more recent events/experiences, whereas implicit mental representations build on associative structures that are unaccessible through introspection, less malleable, and thus may reflect earlier events/experiences (Fazio & Olson, 2003).

Using the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), a widely applied experimental method for measuring implicit social cognition (Lane, Banaji, Nosek, & Greenwald, 2007), Schmukle et al. (2007) indeed obtained stronger associations (significant for men, but not significant for women) of 2D:4D with sex-role orientation, when measured implicitly rather than explicitly. However, this evidence again was based on a small (39 men and 42 women) single-sample design without replication attempt.

There are several generic approaches for clarifying conflicting research literatures, most of which have already been tried within 2D:4D research, including meta-analyses (e.g., Putz, McDaniel, Jordan, & Breedlove, 2008), individual-data reanalyses of studies (e.g., McFadden, Loehlin, Breedlove, Lippa, Manning, & Rahman, 2005), and multi-sample approaches (e.g., Troche et al., 2007). Replicability of findings, a straightforward scientific criterion, remains the via regis for establishing and furthering knowledge (Hendrick, 1990). For clarity, replications are differentiated into exact (or direct) vs. systematic (or conceptual) ones. Exact replications imply repeating studies using identical procedures, thus supporting (if successful) or contesting (if unsuccessful) a theory. Systematic replications use theoretical, procedural, or operational changes to the original study, thus extending (if successful) or limiting (if unsuccessful) a theory.

In the present research, we applied both types of replications. The rationale was as follows: if candidate individual difference variables indeed show stronger or more consistent associations with 2D:4D when measured implicitly rather than explicitly (Schmukle et al., 2007), this should generalize to other traits than sex-role orientation. Several studies have investigated associations of 2D:4D with types of sexually differentiated aggression (Austin, Manning, McIntroy, & Mathews, 2002; Bailey & Hurd, 2005; Benderlioglu & Nelson, 2004; Coyne, Manning, Ringer, & Bailey, 2007; Gallup, White, & Gallup, 2007; Hamson et al., 2008; Kuepper & Hennig, 2007; McIntyre, Barrett, McDermott, Johnson, Cowden, & Rosen, 2007; Millet & Dewitte, 2007). Of note, their findings appear less conflicting than the 2D:4D literature about sex-role orientation reviewed above and thus this line of inquiry possibly is more fruitful. A detailed review of these accounts is beyond the intended scope of the present work, but, suffice it to say, that the majority of the studies cited above yielded theory compliant within-sex associations of 2D:4D with various measures of aggression. Also, there is evidence for an early emergence of sex differences in aggression and related traits (e.g., childhood play) that are influenced by differences in early exposure to androgens (Hines, 2008). We therefore selected aggression (measured explicitly vs. implicitly) as the target trait for investigating associations with 2D:4D. This attempt of demonstrating generalizability constitutes systematic replication.

As a further procedural extension, we considered two additional measures of hand anatomy that similarly are thought to be putative pointers to sex-hormonal exposure. First, DR-L, the directional asymmetry (right-minus-left difference) in 2D:4D, typically is lower in men than in women (Manning, Churchill, & Peters, 2007; Voracek, Dressler, & Manning, 2007) and has therefore been proposed as another putative marker for prenatal androgen levels (Manning, 2002, pp. 21-22). For example, Coyne et al. (2007) found DR-L to be negatively correlated with indirect aggression in women. Second, absolute finger length (2D and 4D), for which within-sex correlations with adult height are only moderately positive (merely hovering around .50; Lippa, 2003), seems to reflect pubertal-adolescent androgen levels, because the sex difference therein is negligible prior to puberty, but very large thereafter (Lippa, 2006). Because of the only loose association with adult height and further because of evidence for significant cross-talk of the androgen receptor (the target of testosterone) with growth-factor signalling pathways (Nieschlag, Behre, & Nieschlag, 2004), measures like absolute finger length cannot be regarded as pointers to growth-hormone exposure exclusively, but may well be appropriate markers for pubertal-adolescent testosterone exposure. Up to now, the measure of absolute finger length has only rarely been considered in 2D:4D research (e.g., Jackson, 2008).

We refrained from basing conclusions on findings yielded from a single sample, even when this was considerably larger than the one of the only predecessor study (Schmukle et al., 2007). We therefore collected data from a second, independent sample, using identical procedures. That way, we conducted an exact replication of our own research.

Methods

Participants

Sample I were 114 men and 130 women, ages 16 to 59 years (M= 27.5, SD= 10.3). Sample II were 101 men and 132 women,
Peace
provocation, punch, revenge, violence, weapon; categories (tenderness) selected by the authors.

In the traditional IAT, genuinely intrapersonal associations nonconscious associations between paired concepts (commonly, indirectly measure the relative strengths of automatic, computer screens, whereby response latencies are analyzed to keyboard-based rapid sorting tasks of stimuli presented on.

1998). In the traditional IAT, genuinely intrapersonal associations are contaminated with extrapersonal (cultural) ones, because categories Pleasant vs. Unpleasant are likely perceived as normative. To reduce this contamination, personalized IAT variants (Olson & Fazio, 2004) substitute these with idiosyncratic categories (I like vs. I don’t like) and use evaluation-laden stimuli with strong individual preferences, but no social consensus. A considerable body of evidence attests to the predictive validity of IAT measures using a variety of attitudes and self-concepts (for reviews, see Fazio & Olson, 2003; Greenwald & Nosek, 2001).

A personalized aggression IAT was assembled as follows: stimuli for the idiosyncratic categories (I like vs. I don’t like; airplanes, beer, cleaning house, cigarettes, coffee, country music, darkness, disco, eating, football, garlic, jogging, Monday, motorcycles, opera, romance novels, spinach, storms, television, tequila) were adopted (Olson & Fazio, 2004) and those for the attributes (Aggression: anger, danger, fight, hothead, pain, provocation, punch, reverse, violence, weapon; Peace: calmness, freedom, health, heaven, joy, love, loyalty, pleasure, relaxation, tenderness) selected by the authors.

For a schematic overview, see Table 1. Stimuli from the categories (I like vs. I don’t like) and the attributes (Aggression vs. Peace) were presented in 5 blocks. Practice blocks (1, 2, and 4) introduced concepts and stimuli (with each word presented twice). Blocks 1 and 4 were similar, except that attribute position was switched. Critically, concepts were differentially paired in Blocks 3 and 5, in order to measure their degree of association. Order of these blocks was not counterbalanced to maximize individual differences measurement and thus to enable unconstrained target-trait correlations (Egloff & Schmukle, 2002).

The aggression IAT was run on laptops, with concept labels presented in the right and left upper screen corners and stimuli midway (intertrial interval: 150 ms). Participants were instructed to react as quickly as possible and used keys «E» vs. «I» for sorting to the left vs. right. Response latencies between stimulus presentation and keystroke were recorded throughout Blocks 1 to 5 in ms.

Finger-length measure. Following published 2D:4D ascertainment standards (Voracek, Manning, & Dressler, 2007), flatbed-scanned images of participants’ right and left palms were produced. From high-resolution laser printouts of scanned images, length of 2D and 4D was measured triplicate, using the same image, from the middle of the flexion crease proximal-most to the palm to the fingertip by experienced, mutually blinded investigators (not the authors), using digital vernier calipers accurate to 0.01 mm. Average measurements were used to calculate right-hand and left-hand digit ratio (R2D:4D and L2D:4D).

Procedure
After reporting basic demographic information, participants completed the IAT, the BPAQ, and the feeling thermometer in quiet facilities. Presenting an IAT prior to corresponding explicit measures minimizes carry-over or priming effects (if any) among measures, because implicit measures by their nature are less likely to elicit such effects (Egloff & Schmukle, 2002). Finally, palm scans were taken and participants thanked and debriefed.

Data analysis
Calculation of individual IAT results followed standard practice, namely the originally proposed IAT effect algorithm (for full procedural details, see Greenwald et al., 1998) which continues to be used by a clear majority of IAT studies. Briefly, the IAT effect (in units of logged milliseconds) is the difference between the mean natural log transformed valid response latencies in attribute trials were discarded from analysis (see Table 2 note; by definition, response errors in idiosyncratic category trials are undetectable). Also following standard practice (Greenwald et al., 1998), calculation of the reliability of the IAT was based on the 78 response-time differences from the trials of Block 5 and 3 (the first two out of the 80 trials of these blocks are omitted).

Interobserver repeatabilities of finger-length measurements were assessed with single-score intraclass correlation coefficients (ICC; two-way mixed-effects model with absolute-agreement definition; Voracek et al., 2007). Significance was set to p < .05 (two-tailed), and ICCrep, the probability of replicating an effect (Killeen, 2005), was calculated for selected findings for demonstration purposes.

### Table 1

<table>
<thead>
<tr>
<th>Block</th>
<th>Trials</th>
<th>Block type</th>
<th>Concept label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>Practice</td>
<td>Peace</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>Practice</td>
<td>I like</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>Test</td>
<td>Peace + I like</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>Practice</td>
<td>Aggression</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>Test</td>
<td>Aggression + I like</td>
</tr>
</tbody>
</table>
Results

Sex differences in variables

Descriptive statistics and tests for sex differences are shown in Table 2. Replicating previous evidence (Manning, 2002) in both samples, men had lower 2D:4D than women. This sex effect was small-to-medium, somewhat weaker for L2D:4D than for R2D:4D, and nearly absent for DR-L. Sex differences in absolute finger length were large, but invariably larger for 4D than for 2D, which pattern is the basis for the sex effect in 2D:4D.

Implicitly measured aggression was somewhat (but not significantly) stronger in men than in women in both samples.Explicitly measured aggression showed no sex effect (feeling thermometer) or generally small ones that were inconsistent across samples (BPAQ). Specifically, significantly higher Physical aggression scores of men relative to women in Sample I failed to replicate in Sample II, and significant sex differences in Anger scores were directionally inconsistent across samples.

Reliabilities of variables

These are displayed in Table 3 (diagonal entries). Consistent with prior evidence, in both samples finger-length and digit ratio measurements were highly repeatable (Voracek et al., 2007), internal consistency figures of the IAT high (Lane et al., 2007), and those of BPAQ subscales tallied to reference values (Herzberg, 2003).

Sex-specific associations among variables

All interrelations of variables are set out in Table 3 (off-diagonal entries). As expected, correlations among absolute finger lengths and digit ratios were strongly positive (upper-left corner entries) in both samples. Specifically, correlations among BPAQ measures were consistently positive, with the highest correlations observed for Anger and Hostility subcales.

### Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men M (SD)</th>
<th>Women M (SD)</th>
<th>t</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Digit ratio</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2D:4D</td>
<td>0.957 (0.032)</td>
<td>0.969 (0.032)</td>
<td>-3.02</td>
<td>.003</td>
<td>-0.37</td>
</tr>
<tr>
<td>L2D:4D</td>
<td>0.956 (0.032)</td>
<td>0.967 (0.029)</td>
<td>-2.67</td>
<td>.008</td>
<td>-0.36</td>
</tr>
<tr>
<td>DR-L</td>
<td>0.001 (0.019)</td>
<td>0.003 (0.026)</td>
<td>-0.65</td>
<td>.52</td>
<td>-0.09</td>
</tr>
<tr>
<td>Absolute finger length (mm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2D</td>
<td>74.20 (3.99)</td>
<td>68.39 (4.43)</td>
<td>10.69</td>
<td>&lt;.001</td>
<td>1.37</td>
</tr>
<tr>
<td>R4D</td>
<td>72.43 (4.04)</td>
<td>67.55 (4.44)</td>
<td>8.64</td>
<td>&lt;.001</td>
<td>1.14</td>
</tr>
<tr>
<td>L2D</td>
<td>77.59 (4.43)</td>
<td>70.61 (4.39)</td>
<td>11.66</td>
<td>&lt;.001</td>
<td>1.49</td>
</tr>
<tr>
<td>L4D</td>
<td>75.64 (3.96)</td>
<td>69.81 (4.71)</td>
<td>10.02</td>
<td>&lt;.001</td>
<td>1.32</td>
</tr>
<tr>
<td>Average finger length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAT (log ms)</td>
<td>0.26 (0.22)</td>
<td>0.31 (0.24)</td>
<td>1.57</td>
<td>.12</td>
<td>-0.22</td>
</tr>
<tr>
<td>BPAQ-Physical aggression</td>
<td>-0.98 (0.73)</td>
<td>-1.34 (0.57)</td>
<td>4.27</td>
<td>&lt;.001</td>
<td>0.55</td>
</tr>
<tr>
<td>BPAQ-Verbal aggression</td>
<td>0.31 (0.51)</td>
<td>-0.29 (0.55)</td>
<td>-0.19</td>
<td>.85</td>
<td>-0.04</td>
</tr>
<tr>
<td>BPAQ-Anger</td>
<td>0.03 (0.64)</td>
<td>0.05 (0.60)</td>
<td>-0.22</td>
<td>.82</td>
<td>-0.03</td>
</tr>
<tr>
<td>BPAQ-Hostility</td>
<td>-0.37 (0.66)</td>
<td>-0.52 (0.63)</td>
<td>1.66</td>
<td>.10</td>
<td>0.23</td>
</tr>
<tr>
<td>BPAQ-Total</td>
<td>0.61 (0.71)</td>
<td>-0.40 (0.71)</td>
<td>2.37</td>
<td>.02</td>
<td>0.30</td>
</tr>
<tr>
<td>Feeling thermometer (degrees)</td>
<td>25.54 (20.71)</td>
<td>26.33 (18.99)</td>
<td>-0.31</td>
<td>.76</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

First-line entries= Sample I (114 men, 130 women; for IAT: 95 and 107), second-line entries= Sample II (101 men, 132 women; for IAT: 79 and 120). d= Cohen’s d (average group difference [male mean minus female mean] divided through square root of weighted mean of group variances). BPAQ= Buss-Perry Aggression Questionnaire (mean item responses). P values are two-tailed.
subscale and between these and the feeling thermometer were consistently positive (lower-right corner entries). Implicitly measured aggression was largely dissociated from the explicit aggression measures, one noteworthy exception being a highly consistent, replicable, or theory compliant association of digit ratios with explicit or implicit aggression measures (details omitted), thus suggesting that the nil findings of the main analysis were robust and not influenced by outliers.

As for sex-specific associations of finger-length measures with implicit or explicit aggression measures (upper-right and lower-left corner entries), crucially, out of 8 (finger-length measures) × 7 (aggression measures) = 224 computed correlations only 4 were significant. Not a single of these associations replicated across samples or would have survived multiple-tests correction, and 3 of them were directionally opposite to theory, with the exception of a negative correlation between R2D:4D and Physical aggression for men in Sample II (r = -.23, p < .05). Overall, the proportion of significant correlations (4 / 224 = 1.8%) was even below the expected error rate of statistical tests (5%, if p < .05), when the null hypothesis (no effect) actually is true.

Re-running the correlational analysis using Spearman’s rank-order correlation coefficient also failed to yield any discernible pattern of consistent, replicable, or theory compliant associations of digit ratios with explicit or implicit aggression measures (details omitted), thus suggesting that the nil findings of the main analysis were robust and not influenced by outliers.

Discussion

This study was only the second one to examine associations of 2D:4D with a candidate trait measured both explicitly and implicitly. Samples I and II combined were almost six times larger than in the predecessor study (Schmukle et al., 2007). Based on a subsample of 39 men, the predecessor reported a significant correlation of r = .33 between R2D:4D and an implicit (IAT-based) measure of gender self-concept (i.e., masculinity-femininity). Assuming that this correlation magnitudinally was typical (i.e., that associations between 2D:4D and implicitly measured target

---

Table 3

<table>
<thead>
<tr>
<th>R2D:4D</th>
<th>L2D:4D</th>
<th>D_R-L</th>
<th>R2D</th>
<th>R4D</th>
<th>L2D</th>
<th>L4D</th>
<th>Average</th>
<th>IAT</th>
<th>Physical</th>
<th>Verbal</th>
<th>Anger</th>
<th>Hostility</th>
<th>Total</th>
<th>Feeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2D:4D</td>
<td>.92***</td>
<td>.83***</td>
<td>.31**</td>
<td>.20*</td>
<td>-.39***</td>
<td>.15</td>
<td>-.32**</td>
<td>-.11</td>
<td>.04</td>
<td>-.04</td>
<td>.01</td>
<td>.10</td>
<td>.06</td>
<td>.04</td>
</tr>
<tr>
<td>L2D:4D</td>
<td>.66***</td>
<td>.84***</td>
<td>.30**</td>
<td>.20*</td>
<td>-.26**</td>
<td>.26*</td>
<td>-.31**</td>
<td>-.05</td>
<td>.14</td>
<td>.02</td>
<td>.07</td>
<td>.09</td>
<td>.14</td>
<td>.08</td>
</tr>
<tr>
<td>D_R-L</td>
<td>.49***</td>
<td>.85***</td>
<td>.42***</td>
<td>.20*</td>
<td>-.07</td>
<td>.22*</td>
<td>-.21*</td>
<td>.04</td>
<td>.09</td>
<td>-.11</td>
<td>.05</td>
<td>-.15</td>
<td>.04</td>
<td>-.08</td>
</tr>
<tr>
<td>R2D</td>
<td>.13</td>
<td>.13</td>
<td>.02</td>
<td>.99***</td>
<td>.82***</td>
<td>.86***</td>
<td>.73***</td>
<td>.92***</td>
<td>.03</td>
<td>-.08</td>
<td>.09</td>
<td>-.01</td>
<td>.01</td>
<td>-.02</td>
</tr>
<tr>
<td>R4D</td>
<td>.17*</td>
<td>.17*</td>
<td>.98***</td>
<td>.85***</td>
<td>.80***</td>
<td>.72***</td>
<td>.91***</td>
<td>.01</td>
<td>-.19</td>
<td>-.08</td>
<td>-.12</td>
<td>-.02</td>
<td>-.13</td>
<td>-.05</td>
</tr>
<tr>
<td>L2D</td>
<td>.01</td>
<td>.165</td>
<td>.18*</td>
<td>.95***</td>
<td>.89***</td>
<td>.99***</td>
<td>.84***</td>
<td>.92***</td>
<td>.08</td>
<td>.06</td>
<td>.01</td>
<td>.03</td>
<td>.11</td>
<td>.03</td>
</tr>
<tr>
<td>L4D</td>
<td>.03</td>
<td>.06</td>
<td>.03</td>
<td>.74***</td>
<td>.71***</td>
<td>.99***</td>
<td>.91***</td>
<td>.94***</td>
<td>-.05</td>
<td>-.11</td>
<td>-.03</td>
<td>-.10</td>
<td>-.03</td>
<td>-.09</td>
</tr>
<tr>
<td>Average</td>
<td>-.15‡</td>
<td>-.06</td>
<td>-.11</td>
<td>.95***</td>
<td>.97***</td>
<td>.97***</td>
<td>.97***</td>
<td>.99***</td>
<td>.03</td>
<td>-.07</td>
<td>.06</td>
<td>-.01</td>
<td>.03</td>
<td>-.01</td>
</tr>
<tr>
<td>IAT</td>
<td>-.20*</td>
<td>-.10</td>
<td>-.13</td>
<td>-.08</td>
<td>.02</td>
<td>-.01</td>
<td>-.04</td>
<td>-.01</td>
<td>.95</td>
<td>.03</td>
<td>-.10</td>
<td>-.05</td>
<td>.01</td>
<td>-.03</td>
</tr>
<tr>
<td>Physical</td>
<td>.05</td>
<td>.09</td>
<td>.05</td>
<td>.05</td>
<td>-.06</td>
<td>-.05</td>
<td>-.09</td>
<td>-.06</td>
<td>.05</td>
<td>.77</td>
<td>.20*</td>
<td>.46***</td>
<td>.27**</td>
<td>.74***</td>
</tr>
<tr>
<td>Verbal</td>
<td>-.01</td>
<td>-.01</td>
<td>-.17‡</td>
<td>-.05</td>
<td>.01</td>
<td>-.01</td>
<td>-.01</td>
<td>-.07</td>
<td>.28**</td>
<td>.59</td>
<td>.34***</td>
<td>.20*</td>
<td>.51***</td>
<td>.40***</td>
</tr>
<tr>
<td>Anger</td>
<td>-.02</td>
<td>.09</td>
<td>-.13</td>
<td>-.08</td>
<td>-.07</td>
<td>-.05</td>
<td>-.09</td>
<td>-.07</td>
<td>.08</td>
<td>.42***</td>
<td>.41***</td>
<td>.72</td>
<td>.46***</td>
<td>.79***</td>
</tr>
<tr>
<td>Hostility</td>
<td>-.06</td>
<td>-.11</td>
<td>.17*</td>
<td>-.14</td>
<td>-.12</td>
<td>-.05</td>
<td>-.13</td>
<td>.16</td>
<td>.57***</td>
<td>.24**</td>
<td>.58</td>
<td>.56***</td>
<td>.80***</td>
<td>.27**</td>
</tr>
<tr>
<td>Total</td>
<td>-.02</td>
<td>.11</td>
<td>-.10</td>
<td>-.09</td>
<td>-.09</td>
<td>-.06</td>
<td>-.11</td>
<td>-.09</td>
<td>.07</td>
<td>.72***</td>
<td>.56***</td>
<td>.78***</td>
<td>.74***</td>
<td>.82</td>
</tr>
<tr>
<td>Feeling</td>
<td>-.06</td>
<td>.19*</td>
<td>.14</td>
<td>.04</td>
<td>.01</td>
<td>-.06</td>
<td>.02</td>
<td>.02</td>
<td>.08</td>
<td>.54***</td>
<td>.22*</td>
<td>.32***</td>
<td>.19*</td>
<td>.45***</td>
</tr>
</tbody>
</table>

First-line entries= Sample I, second-line entries= Sample II. Diagonal entries= ICCs (finger-length measurements) or Cronbach’s α (aggression measures). Off-diagonal entries= Pearson correlations, above-diagonal entries= men, below-diagonal entries= women. ‡ p < .10, * p < .05, ** p < .01, *** p < .001 (two-tailed)
traits typically would yield \( r \) figures of about .30, the a priori estimate of statistical power (Elashoff, 2000) to detect this theoretical \( r \) value of .30 would have merely been 45% with twosided testing and a sample size of 39. This suggests that the finding of Schmukle et al. (2007) may well be a chance finding. However, given sample sizes of 79, 95, 107, or 120 (as were available for the tests for within-sex associations of digit ratios with the IAT in Samples I and II of the present study; see Table 1), the a priori power estimates to detect an expected \( r \) value of .30 with twosided testing amount to 76%, 84%, 88%, or 91%, respectively. These values of statistical power are sufficiently high and thus suggest that the present two studies were not statistically underpowered and that their nil findings are unlikely to be due to statistical type II errors.

The current findings are clear-cut. Overall, there was no indication of theory compliant, reliable, or replicable within-sex associations of putative markers for prenatal (2D:4D, \( D_{45} \)) or pubertal-adolescent (absolute finger length) androgen action with either implicit or explicit measures of aggression. The lack of reliable 2D:4D associations with BPAQ scores is in line with nil findings of 2D:4D studies that also used the BPAQ (Austin et al., 2002; Coyne et al., 2007). Some recent studies (Millet & Dewitte, 2007, 2008, 2009; van den Bergh & Dewitte, 2006) point to the possibility that contextual cues may moderate 2D:4D effects, particularly cues eliciting or priming aggressive versus nonaggressive (altruistic and prosocial) behavior. By its design, this eventuality could not be addressed in the present research. However, it is worth mentioning that (similar to other strands of evidence from 2D:4D research) at present such context-dependent interaction effects for associations of 2D:4D with target traits have only been reported by one research group (and thus await independent replication) and furthermore have only been tested within a limited range of behavioral responses, namely largely for economic aggression measures (which calls for the need to demonstrate generalizability of such assumed interaction effects to a wider range of behavior). In our study, 2D:4D was also not reliably associated with aggression measured implicitly with a personalized IAT variant. In this regard, our systematic replication (or generalizability test) of the findings of Schmukle et al. (2007) failed: it is therefore uncertain (and remains to be tested anew) whether 2D:4D really is stronger or more reliably or anyway associated with sexually differentiated (and thus conceivably hormonally influenced) candidate individual difference variables, when these are assessed implicitly rather than explicitly. Further, explicit and implicit aggression measures themselves were largely dissociated, a finding not uncommon in IAT research (Lane et al., 2007).

Diverse knowns of previous research replicated in both samples, thus attesting to effect robustness, procedural validity, and data typicality. For instance, measurements of finger length and digit ratios were highly repeatable, internal consistency figures of the IAT were high (see Mierke & Klauser, 2003) and those of BPAQ subscales dovetailed to published normative data for its German form. Sex differences in digit ratios were significant, small-to-medium sized, somewhat more pronounced for R2D:4D than for L2D:4D, noticeably smaller and nonsignificant (but directionally theory compliant) for \( D_{45} \). Those in absolute finger length were large, but invariably larger for 4D than for 2D, affirming the position that it is men’s relatively longer 4D that accounts for the sex effect seen in 2D:4D. BPAQ subscales were throughout positively interrelated and likewise throughout positively correlated with a simple, single-item explicit aggression measure of high face validity (feeling thermometer). The novel personalized IAT variant for assessing aggression showed potential to yield sex effects in the predicted direction, although magnitudinally these were small.

In contrast, the failure of sex effects in the BPAQ to replicate reliably and directionally consistent across samples suggests psychometric shortcomings of this instrument’s internal structure (see Herzberg, 2003), which also have become apparent in several adaptations of the BPAQ in other languages (Dutch version: Meesters, Muris, Bosma, Schouten, & Beuving, 1996; Japanese version: Nakano, 2001) as well as in an alternative German version of the BPAQ (von Collani & Werner, 2005). For instance, in one of their samples, von Collani and Werner (2005) found significant, but small, sex differences in Physical aggression scores (men higher; \( d_{p} = 0.31 \) and Anger scores (women higher; \( d = -0.25 \)), whereas no significant sex differences in Verbal aggression or Hostility scores. The sex difference in Physical aggression scores replicated in another sample and was medium-sized (men higher; \( d = 0.64 \)), but the sex difference in Anger scores did not replicate, and, on the other hand, there was a significant sex difference in Verbal aggression scores (contrary to expectation, men higher; \( d = 0.22 \)) in this second sample, which had been absent in the first sample. Also, the sex difference in total BPAQ scores in the second sample was, although statistically significant, magnitudinally modest (men higher; \( d = 0.21 \)).

More generally, the current nil findings add to a growing number of failures in 2D:4D research to replicate initially reported findings from small-N, single-sample designs in subsequent investigations providing large-scale or multi-sample evidence or both (e.g., Medland, Loehlin, & Martin, 2008). Presently, the majority of 2D:4D research findings is unreplicated. Time will show which of these will turn out unreproducible. In this regard, the combined approach of systematic replication attempts of prior evidence, along with exact replication attempts of one’s own research, as pursued here, appears beneficial and of general interest for the 2D:4D research program. In tandem with large-scale, multi-sample studies and meta-analyses, this should help sorting out veridical findings from those due to chance, selective publishing, or related phenomena. To achieve this goal, making original nil findings and replication failures visible will be of utmost importance. By the same token, one conceivable implication of the current design is to adjust publication acceptance criteria in this field, such that a successful replication of 2D:4D associations with target traits must be presented.

A final proposal concerns the reporting of replication statistics. This should assist in gauging the veridicality of findings. For example, calculated \( p_{rep} \) values for the sex effect in digit ratios observed in our samples (Table 2) ranged from 0.94 to 0.98, indicating a high degree of trustworthiness. Indeed, the sex effect in 2D:4D undoubtedly constitutes the most robust and replicable 2D:4D research finding so far. However, \( p_{rep} \) figures for observed correlations between finger-length and aggression measures were much lower (many of them < .50), representing a high degree of uncertainty to see the same effect replicated in another research report. Or in one’s own next sample, as we did.
Kuepper, Y., & Hennig, J. (2007). Behavioral aggression is associated with
Herzberg, P.Y. (2003). [Psychometric evaluation and validity of the Ger-
Hampson, E., Ellis, C.L., & Tenk, C.M. (2008). On the relation between
Coyne, S.M., Manning, J.T., Ringer, L., & Bailey, L. (2007). Directional asymmetry (right-left differences) in digit ratio (2D:4D) predict indi-
Millet, K., & Dewitte, S. (2009). The presence of aggression cues inverts the relation between digit ratio (2D:4D) and pro-social behaviour in a dictator game. British Journal of Psychology, 100, 151-162.


