Can we believe judgements of human physical attractiveness?

Martin J. Tovée[a], Jennifer Taylor[b] & Piers L. Cornelissen[c]

[a] School of Psychology, Lincoln University, Lincoln, LN6 7TS, UK.

[b] Department of Psychology, University of York, York, YO10 5 DD, UK.

[c] Department of Psychology, Faculty of Health and Life Sciences, Northumbria University, Northumberland Building, Northumberland Road, Newcastle upon Tyne, NE1 8ST, UK.

Word count: 4,004 (excluding references and abstract).

*Requests for reprints should be addressed to Prof Tovée, School of Psychology, Lincoln University, Lincoln, LN6 7TS, UK. (e-mail: MTovee@lincoln.ac.uk).

Abstract

A key question in attractiveness studies is the validity of the reported results outside the narrow confines of the experimental paradigm used. Does the range of physical features in a set of pictures used to test attractiveness judgments predict the individual ratings of each body? Or does each stimulus have an attractiveness value independent of the range of attractiveness found in the image set of which it is a part? An additional problem is that because participants are often shown a relatively large array of images in a short space of time, this may produce perceptual biases which could cause a short-term shift in attractiveness preferences which are not usually found in real life mate choice decisions. To address this issue we asked 20 participants (10 male and 10 female) to judge the attractiveness of 20 digital photographs of female bodies. We then asked a different set of 400 people (who had not seen the body pictures) to judge the attractiveness of one of the bodies from the set (so each body was rated in isolation by 10 male and 10 female participants). We then compared the attractiveness judgement each body received when seen independently versus when it was seen within the context of a set of bodies. The results showed no significant difference between the two conditions which suggests that each body has an attractiveness value independent of the attractiveness of the other bodies with which it is viewed.

Keywords: Attractiveness ratings, Body Judgements, viewing context, contraction bias, adaptation**.**

1. Introduction

In laboratory studies of human attractiveness, whether of faces or bodies, a key potential confound is the methodology used to obtain attractiveness ratings. Most studies either ask participants to rank images from a small stimulus set, all of which are visible at any one time (e.g. Henss, 1995, 2002; Singh, 1993; Swami et al. 2006, 2012), or to rate a larger set of images, each presented separately in a random order, but with the participants having already been shown the whole set to acquaint them with the degree of variation in physical features (e.g. Fan et al., 2004; Furnham et al., 2005, 2006; Rilings et al., 2009; Smith et al., 2007a; Streeter & McBurney, 2004; Tovée et al., 1999, 2003). In both cases, it can therefore be argued that each face/body is rated in the context of the other faces/bodies within a given set, but that the rating of an individual face/body could be different if it were included in a different set of images. Critically, it is also possible that a participant might use a rating strategy and/or adapt their response scale to the specific task at hand; developing, on the fly, a heuristic which allows them to discriminate between the stimuli in given set, just for the sake of completing the task they have been given. For example, a participant might focus on a particular body feature which happens to show a great deal of variation in that particular set of images [such as body mass index (BMI) or waist-to-hip ratio (WHR)] and which can thus be used to efficiently differentiate between the images in their ratings. However, this may not be the strategy/feature they might otherwise use in actual attractiveness judgements during real world mate selection.

An additional problem is that because participants are often shown a relatively large array of images in a short space of time, this may cause sequential contraction bias. This is a transfer bias that occurs when observers judge the magnitude of a number of stimuli one directly after

another (Poulton, 1989). The previous stimulus becomes an additional reference magnitude against which the next stimulus is judged. Consequently, observers may underestimate the size of the difference between the previous stimulus and the next stimulus. Clearly, randomization of a very large stimulus set across a very large number of participants would very likely abolish such effects, as far as any central tendencies in the results are concerned. However, smaller stimulus sets with modest samples of participants – arguably the norm in such studies – might nevertheless be at risk, especially for some stimulus pairings where sequential contraction bias might be large (e.g. if an image representing the largest BMI in the set immediately follows an image representing the smallest BMI). Furthermore, when the range of responses available to a participant has an obviously recognisable centre, such as on a line or Likert scale (e.g. Swami and Tovée, 2012), observers may simply default to this centre value and ignore the ends of the scale. This value can then become the reference magnitude for responding so that observers increasingly select responses that are closer to the centre value than they would otherwise do (Poulton, 1989). Altogether, therefore, it is plausible that previously published laboratory based studies of human health and attractiveness may have been subject to one or more of these effects, possibly leading to poor face validity (Anastasi, 1988).

To address this (potential) problem, in the current study of human female physical attractiveness, we asked individual participants to judge just one image from a set of stimuli, but recruited a large number of participants so that each image could be rated many times. Rating images in isolation means that one participant's judgement can only be made by comparing the given image to an internalized reference, and we can safely assume that this has been learnt over the course of their life time. Critically, this is of course the source of information we want to interrogate in the first place (Cornelissen et al., 2013, 2015; Rhodes,

et al., 2013; Robinson & Kirkham, 2013; Winkler and Rhodes, 2005). What participants will not be able to do is discover a temporary heuristic, which is based purely on arbitrary features distributed in a particular set of test images. As a second step in the current study, we compared these 'single-shot' results to a conventional paradigm in which a different set of participants saw the entire stimulus set first and then made sequential judgements about each of them, presented in random order, one after the other. If the 'single-shot' and conventional 'multi-body' data converge on the same pattern of results, we can have greater confidence in the validity and applicability of previous laboratory based studies.

2.0 Method

2.1 Sample size and stimuli

As stimuli, we used images of females as previous research has suggested a clear relationship between varying the BMI of the women in such images and the attractiveness ratings given to them (e.g. Tovée et al., 1998, Swami & Tovée, 2005; Smith et al., 2007b). Thus by varying the images along this dimension we expected to see a significant change in the ratings of different images. In our analysis, we also included WHR as an explanatory variable. WHR has been extensively used as a potential predictor of attractiveness judgements of female bodies (Fan et al., 2004; Furnham et al., 2005, 2006; Rilings et al., 2009; Streeter & McBurney, 2004) and although it is a comparatively weak predictor in studies using images of real women, it can have a significant effect in small image sets when the range of BMI is constrained (e.g. Tovée et al., 2002).

For the single-shot study we had to find a compromise between having as many images rated as possible, where each image would be rated individually by both male and female

observers, and making the study logistically feasible in terms of the total number of participants to be recruited and tested – ours was not an online study. Our approach to solving this problem was two-fold. First, we sought to minimize the number of parameters to be estimated in any multiple regression model of attractiveness ratings. In general terms, sample sizes tend to increase as the number of regression coefficients to be estimated increases. Therefore, minimizing the number of parameters to be estimated should help to reduce the required sample size for a given statistical poser, thereby making the study more logistically feasible. However, for studies of physical attractiveness, there is a particular problem. If the BMI of the women in the stimuli is allowed to vary as much as is typical in the general population, e.g. ~ 12 to ~ 40, then attractiveness ratings are best described as an asymmetric inverted U-shape function of BMI: attractiveness judgements increase rapidly from BMI ~12 to a peak at ~19, and then decrease more slowly from this peak towards a minimum at BMI ~40. From a statistical modelling point of view, this requires a third order polynomial for BMI, comprising three parameter estimates. However, we can restrict the number of parameters to be estimated for BMI to one, by selecting stimuli from the range ~19 to ~40 which corresponds to a linear decrease in attractiveness. This first criterion of a stimulus set with a restricted BMI range was met by using the colour image database reported by Smith et al. (2007a), which comprises 42 images whose BMI varies from 18.42 to 26.68.

The next question was how many images should be selected from this database as a stimulus subset and still achieve adequate statistical power? Using the data from Smith et al., (2007a), we found that a model with linear fits for BMI and WHR had an r-square of ~ 0.55, equating to an effect size f^2 of 1.2. Based on this, to achieve a power (1-β) of 0.9 and an α of 0.05 in the current single-shot study, we would need ~14 images to be rated. Therefore, for the current study, we selected a subset of 20 images from Smith et al. (2007a) by evenly

sampling from the space described by a scatterplot of image WHR plotted as a function of BMI. The mean, minimum, maximum and standard deviation for BMI and WHR of the 20 images we selected were: BMI: M = 22.46, Min = 18.42, Max = 26.68, SD = 2.41; WHR: M = 0.75, Min = 0.68, Max = 0.84, SD = 0.043. Clearly, by deliberately restricting the BMI range of the stimuli in this way, there is the possibility that the relative ranges of BMI and WHR in the stimulus set may not reflect those in the general population. A bias like this should not affect the research question at hand because it applies equally to the data from the two methodologies. Nevertheless, to quantify it, we compared the ranges of BMI and WHR in the stimulus set to those observed in the Health Survey for England (HSE 2012), defined in terms of z-scores. For BMI the range in z-scores was -1.32 to 0.21 (difference = 1.52) and for WHR it was -1.54 to 0.75 (difference = 2.29). Therefore, relative to the population at large, there was a wider range of WHR in our stimulus set than there was for BMI. For the single-shot study, each image was rated once by 10 male and 10 female observers, with different sets of 20 participants (10 male, 10 female) assigned to each image, giving a total of 400 participants in all. The women in the images wore identical unsupportive flesh coloured vests and briefs to ensure standardised clothing. Additionally, their faces were blurred both for anonymity and to ensure participants only rate body attractiveness (and not facial attractiveness). Examples of our stimuli are shown in Figure 1 A.

2.2 Participants

For the single-shot study, 400 hundred students from the University of York (200 males) agreed to take part in the study, and were recruited by opportunity sampling face to face (this was not an online study). To ensure that cross cultural variation was controlled for, only participants who rated themselves as: White British, White Irish or White Other were asked to participate. This is due to research that shows that preferences for female attractiveness, in

particular with regards to weight and WHR, may be culturally dependent (e.g. Boothroyd et al., 2016; Furnham et al., 2002; Swami et al., 2012, 2013). The female participants' ages were M= 21.24 years, SD = 4.15 years and the males' M = 21.81 years, SD = 5.34 years. For the multi-body study, 20 additional students were recruited (10 male). For this separate sample of observers, the participants' ages were M= 20.80 years, SD = 4.67 years and M = 21.33 years, SD = 4.91 years for females and males respectively.

2.3 Procedure

For the single-shot study the investigator showed each participant one image, printed in colour on an A4 sheet, and asked the participant to make two judgements about it. One was an attractiveness rating on a visual analogue scale (VAS) and the other was a two alternative forced choice (2AFC) decision (yes/no judgement). The VAS was 90mm in length, the left side was labelled ' very unattractive' and the right side 'very attractive'. Participants were asked to mark with a pencil line on this scale how attractive they perceived the image to be. The 2AFC judgement required the participant to decide categorically whether they found the woman in the image attractive or not. The 2AFC judgement was included in order to provide an estimate of reliability of attractiveness judgements in the single-shot data. The order in which participants made the VAS versus 2AFC ratings was counterbalanced across participants.

The multi-body experiment was run on a Windows PC using E-prime. The images were presented on a 17 inch LCD monitor (1280x1024-pixel resolution, 32-bit colour depth) viewed from a distance of about 60cm. First, all 20 images were presented in random order so that participants could become aware of the range of shapes, sizes, and features of the female volunteers in the stimuli. Then participants saw each image a second time, again presented in

a random order, but this time for ten seconds after which they made a VAS judgement on a response sheet.

3.0 Results

3.1 Univariate data

**********Table 1 about here **********

It is clear from Table 1 that there is good agreement between the VAS and 2AFC methods of image rating as well as good agreement between male and female observers for the single-shot task. Similarly, for the multi-body sample, we calculated Cronbach's alpha separately for male and female participants (i.e. 0.89 and 0.95 respectively), as well as the Pearson correlation between the mean males' and females' ratings per image ($r = 0.95$, $p < .0001$). In general, the pattern of results comparing male and female performance across the single-shot and multi-body studies suggest that males and females showed good agreement about the relative ranking of the attractiveness of the women in the stimuli. Therefore, we collapsed the single-shot and multi-body data across gender in order to simplify the further analyses.

3.2 Multivariate data

**********Figure 1 about here**********

Figure 1B and 1C show plots of mean VAS attractiveness as a function of BMI and WHR respectively, separately for the multi-body (blue) and single-shot (red) datasets. Consistent with Smith et al. (2007a) and Tovée et al (2003), in both of whose studies a restricted BMI range was used, we found linear reductions in attractiveness as both BMI and WHR increased. Moreover, the simple regression lines for BMI and WHR for the single-shot data lay entirely within the 95% confidence intervals for the multi-body data, leading us to expect no statistically significant differences between the modes of study.

Based on our stimulus selection, we assumed that only linear terms should be required for BMI and WHR in the dummy regression analysis reported below. To check this, we first analysed the single-shot and multi-body data separately, and tested whether there was any evidence for non-linearity for either BMI and/or WHR, by adding second order polynomial terms. We found no statistically significant evidence for non-linearity, and therefore only fitted linear terms in the analyses to follow.

We wanted to quantify the relationship between attractiveness ratings, BMI and WHR while simultaneously testing whether the method of study influences these relationships. If multi-body studies of human physical attractiveness have little ecological validity, then we should see a statistically significant effect of study method – i.e. a difference in the slope and/or intercept of the regression of attractiveness on BMI and WHR when comparing the fit for the single-shot data with the multi-body data. However, if the results from standard multi-body investigations of human physical attractiveness can be extrapolated to the real world, then we should see no significant effect of study method.

To do this we ran a multiple regression model in which we used dummy explanatory variables to code for the single-shot versus the multi-body datasets (Zar, 1984), and included

both BMI and WHR as continuous explanatory variables. We used PROC REG in SAS v9.4 (SAS Institute, North Carolina, US) to fit the following dummy regression model:

$$y_i = (\beta 0 + \delta.d) + (\beta 1 + \beta 2).BMI_i + (\beta 3 + \beta 4).WHR_i + \varepsilon_i$$

where: y_i is the outcome, i.e. estimated attractiveness for image i, ε_i is the model error, $(\beta 0 + \delta.d)$ is the model intercept including the dummy term $\delta.d$, $(\beta 1 + \beta 2)$ is the regression coefficient for BMI, with $\beta 2$ as its dummy component, and $(\beta 3 + \beta 4)$ is the regression coefficient for WHR, with $\beta 4$ as its dummy component. We chose the multi-body study as the control level, and set the dummy variables for it to zero accordingly. When the dummy variables are zero, the regression equation behaves as if the relationship between attractiveness, BMI and WHR is estimated for multi-body observers only. However, when the dummy variables are non-zero, both the coefficient δ for the additional intercept component for single-shot observers and coefficients $\beta 2$ and $\beta 4$ for the additional slope component for single-shot observers are estimated. Note however, that in practice, all parameters are estimated simultaneously by PROC REG from the dummy regression model, using both datasets; it is not necessary to run separate models. If either δ and/or $\beta 2$ and/or $\beta 4$ are statistically significantly different from zero, then this would mean that two separate regression lines (or, strictly speaking, regression planes) are required for an adequate fit for BMI/WHR; one for the single-shot data and another for the multi-body data. In this situation, we would have clear evidence that the two experimental methodologies produce different effects on observers' judgements. However, if δ and $\beta 2$ and $\beta 4$ are not statistically significantly different from zero, then single regression lines (i.e. a single regression plane) would provide an adequate fit to BMI and WHR for both data sets. In this alternative situation, we would have no evidence that the two experimental methodologies produce different effects on observers' judgements.

**********Table 2 about here **********

The dummy regression model explained 58.3% of the variance in VAS attractiveness ratings. Consistent with Smith et al (2007a) and Tovée et al. (2003), and as can be seen in Table 2, we found negative, linear relationships between attractiveness judgements and both BMI and WHR. Critically, as can be seen in Table 2, we found that both the slopes and intercept for the single-shot data were not statistically different from those for the multi-body data because the dummy regression parameters were not significantly different from zero (i.e. $\beta_2 = 0.12$, $p = .28$; $\beta_4 = 0.82$, $p = .42$ and $\delta = -6.31$, $p = .19$). Therefore, this analysis supports the idea that no significant differences can be found between multi-body and single-shot studies of human physical attractiveness.

4.0 Discussion

There were no significant differences in the intercepts or slopes for the regressions of attractiveness on BMI and WHR when comparing the single-shot and multi-body experiments, suggesting equivalent levels of agreement between the two paradigms. Therefore it is unlikely that previous laboratory based studies, which usually use some version of the multi-body paradigm, will have been seriously confounded either by the psychophysical biases discussed in the Introduction, or by participants using task strategies and/or reference ranges which are noticeably different from each other. Therefore, these results suggest that both ways of measuring attractiveness judgements are either likely to be

broadly representative of real-life mate choice decisions, or, of course equally flawed – the design of this experiment cannot distinguish between these two possibilities.

Although we could not detect statistically significant differences between the two paradigms, there was nevertheless a trend to suggest a potential difference in the pattern of the responses as illustrated in figure 1. There is a sharper gradient in the relationship between the attractiveness ratings and the physical characteristics for the multi-body conditions, so with more data points (i.e. more bodies to be rated) and a wider range of BMI, the relationship might have become statistically significant. Based on a regression model with only BMI, and its accompanying dummy variables, for an α of 0.05 and power (1-β) ranging between 0.8 – 0.9, power calculations indicate that we would have required 120-160 bodies (over the same BMI range) to obtain such a result. Therefore, even if we were to carry out such a study, the implication is that any difference in the expected effect size between paradigms is rather weak, and would have little impact on day-to-day judgements of attractiveness.

One methodological difference between the single-shot group and the multi-body group is that for the former, data were collected using printed photographs of women's bodies rather than presenting the images on a computer monitor. Therefore, it is conceivable that this might lead to a difference in how bodies were judged. However, a comparison of data collected with the same set of images using print-outs of body photographs as well as computer displays found no differences in the pattern of ratings when the two modes of presentation were compared (Mo et al. 2014). The current results also suggest that there is no significant difference between using a 2-AFC versus a VAS method of recording attractiveness judgements. Again, a similar result was reported by Tovée et al. (2012), who found no difference in judging bodies for body mass or for health on a scale of 0 to 9 versus using a 2-AFC paradigm.

The results suggest that attractiveness ratings are indeed based on comparison to an internal reference template, like a lookup table, and that the ratings of individual bodies are not substantially influenced in the short term by the range of the physical features within a given image set. An alternative hypothesis is that to be adaptive, attractiveness judgements should be malleable. They should be provisional and relative to the other bodies available at a specific time and subject to change as the range of other bodies in the environment changes (i.e. changes in visual diet). There is some evidence for this kind of adaptive change in the gradual adaptation of people's internal template to a heavier level with the general increase in heavier bodies in Western populations (Robinson and Kirkham, 2013; Oldham and Robinson, 2015) and a shift in preferences towards a lower body size through exposure to thinner bodies on television in rural populations in Nicaragua (Boothroyd et al., 2016). Another example is the differences found in the attractiveness ratings made by people resident in Kwa Zula Natal and people who had recently left that South African province and moved to the UK (Tovée et al., 2006). In this case, there is a significant shift in the ratings of individual bodies made by the migrant group towards the same pattern of preferences shown by long-term residents of the UK. However, these changes appear to be happening over the course of months and do not show the kind of very rapid change needed for the range of physical features in a body set to significantly affect the individual rating of a body. Instead the results reported here suggest that a body may have a relatively stable attractiveness value with a degree of independence relative to the attractiveness of other bodies in an environment. More rapid adaptation to environmental changes in body size or shape have been reported, but these studies have used exposure to extreme examples of a body feature to create short-term perceptual adaptation (Winkler and Rhodes, 2005; Boothroyd, Tovée & Pollett, 2012; Rhodes et al., 2013) and this kind of extreme adapting stimulus is uncommon in the real world. So although in evolutionary terms, it would be adaptive for preferences to change over time to compensate

for changes in the environment or movement between environments, these changes should not occur too rapidly or there would be constant fluctuations in preferences. This would disrupt an observer's ability to judge the attractiveness of the bodies around them by adding significant noise to their judgements. Thus, it would adaptive to have a slower response time in changing body preferences to minimise the effect of short-term fluctuations in the sizes of bodies in the visual diet.

This is not to say that a body's relative attractiveness in mate selection may not change with changes in the distribution of body types in their environment (so for example, a body may always be a 6 and so be relatively attractive if all the other bodies in an environment are 5s, but it will become relatively less attractive with the addition of bodies who are rated as a 7). In this scenario, changes in judgements of the attractiveness of a particular body will occur with longer term changes to the size and shape of their internal template. This provides a degree of continuity and stability in attractiveness judgements while at the same time allowing for a gradual modification of the judgements to compensate for changes in environment and circumstance.

5.0 Data Availability

All the data used in this study are available at the following url:

https://sites.google.com/site/unnpsych/datasets/Tovee_et_al_EHB_2016_full.csv?attredirects=0&d=1

Acknowledgments: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Boothroyd, L.G., Jucker, J.L., Thornborrow, T., Jamieson, M.A., Burt, D.M., Barton, R., Evans, E.H. & Tovée, M.J. (2016). Television exposure predicts body size ideals in rural Nicaragua. *British Journal of Psychology*, 2016, (in press). doi: http://dx.doi.org/10.1111/bjop.12184.

Boothroyd, L.G., Tovée, M.J. & Pollett, T.V. (2012). Visual diet versus associative learning as mechanisms of change in body size preferences. *PLoS ONE*, 7, e48691. doi: http://dx.doi.org/10.1371/journal.pone.0048691

Cornelissen, P.L., Johns, A. & Tovée, M.J. Body size over-estimation in women with anorexia nervosa is not qualitatively different from female controls. *Body Image*, 10, 103-111. doi: http://dx.doi.org/ 10.1016/j.bodyim.2012.09.003

Cornelissen, K.K., Bester, A., Cairns, P., Tovée, M.J. & Cornelissen, P.L. (2015). The influence of personal BMI on body size estimations and sensitivity to body size change in anorexia spectrum disorders. *Body Image*, 13, 75-85. doi: http://dx.doi.org/10.1016/j.bodyim.2015.01.001

Fan JT, Liu F, Wu J, Dai W (2004) Visual perception of female physical attractiveness. *Proceedings of the Royal Society B-Biological Sciences* 271: 347–352. doi: http://dx.doi.org/10.1098/rspb.2003.2613

Furnham, A., Moutafi, J. & Baguma P. A (2002). cross-cultural study on the role of weight and waist-to-hip ratio on female attractiveness. *Personality and Individual Differences*, 32, 729–745. doi: http://dx.doi.org/10.1016/S0191-8869(01)00073-3.

Furnham, A. Petrides K.V. & Constantinides, A. (2005). The effects of body mass index and waist-to-hip ratio on ratings of female attractiveness, fecundity, and health. *Personality and Individual Differences*, 38, 1823-1834. doi: http://dx.doi.org/10.1016/j.paid.2004.11.011

Furnham, A., Swami, V. & Shah, K. (2006). Body weight, waist-to-hip ratio and breast size correlates of ratings of attractiveness and health. *Personality and Individual Differences*, 41, 443-454. doi: http://dx.doi.org/10.1016/j.paid.2006.02.007

Henss, R. (1995) Waist-to-hip ratio and attractiveness. A replication and extension. *Personality and Individual Differences*, 19, 479–488. doi: http://dx.doi.org/10.1016/0191-8869(95)00093-l.

Henss, R. (2000) Waist-to-hip ratio and attractiveness of the female figure. Evidence from photographic stimuli and methodological considerations. *Personality and Individual Differences*, 28, 501–513. doi: http://dx.doi.org/10.1016/s0191-8869(99)00115-4

Mo, J.J.Y., Cheung, K., Gledhill, L.J., Pollet, T.V., Boothroyd, L.G. & Tovée, M.J. (2014). Perceptions of female body size and shape in China, Hong Kong, and the United Kingdom. *Cross-Cultural Research*, 48, 78-103. doi: http://dx.doi.org/10.1177/1069397113510272

Oldham, M. & Robinson, E. (2015). Visual weight status misperceptions of men: Why overweight can look like a healthy weight. *Journal of Health Psychology*, 20, 1–10. doi: http://dx.doi.org/10.1177/1359105314566257

Poulton, E.C. (1989). Bias in Quantifying Judgements. Hove, UK: Erlbaum.

Rhodes, G., Jeffery, L., Boeing, A., & Calder, A. (2013). Visual Coding of Human Bodies: Perceptual Aftereffects Reveal Norm-Based, Opponent Coding of Body Identity. *Journal of Experimental Psychology: Human Perception and Performance*, 39, 313-317. doi: http://dx.doi.org/10.1037/a0031568.

Rilling, J.K., Kaufman, T.L., Smith, E.O., Patel, R., and Worthman, C.M. (2009). Abdominal depth and waist circumference as influential determinants of human female attractiveness. *Evolution and Human Behavior*, 30, 21–3. doi: http://dx.doi.org/10.1016/j.evolhumbehav.2008.08.007

Robinson, E., & Kirkham, T.C. (2013). Is he a healthy weight? Exposure to obesity changes perception of the weight status of others. *International Journal of Obesity*, 38, 663-667. doi: http://dx.doi.org/10.1038/ijo.2013.154.

Singh, D. (1993) Body shape and women's attractiveness - the critical role of waist-to-hip ratio. *Human Nature* 4: 297–321. doi: http://dx.doi.org/10.1007/bf02692203

Smith, K.L., Cornelissen, P.L. & Tovée, M.J. (2007a). Color 3D bodies and judgements of human female attractiveness. *Evolution and Human Behavior*, 28, 48-54. doi: http://dx.doi.org/ 10.1016/j.evolhumbehav.2006.05.007.

Smith, K.L., Tovée, M.J., Hancock, P.J.B., Bateson, M., Cox, M.A.A. & Cornelissen P.L. (2007b). An analysis of body shape attractiveness based on image statistics: Evidence for a dissociation between expressions of preference and shape discrimination. *Visual Cognition*, 15, 927-953. doi: http://dx.doi.org/10.1080/13506280601029515

Streeter, S.A. & McBurney, D.H. (2003) Waist-hip ratio and attractiveness. New evidence and a critique of 'a critical test.'. Evolution and Human Behavior, 24, 88–98. doi: http://dx.doi.org/ 10.1016/s1090-5138(02)00121-6

Swami, V. & Tovée, MJ. (2005). Male physical attractiveness in Britain and Malaysia: A cross cultural study. Body Image, 2, 383-393. doi: http://dx.doi.org/10.1016/j.bodyim.2005.08.001.

Swami, V. & Tovée, M.J. (2005b). Female physical attractiveness in Britain and Malaysia: A cross-cultural study. *Body Image*, 2, 115-128.http://dx.doi.org/10.1016/j.bodyim.2005.02.002

Swami, V., Caprario, C., Tovée, M.J. & Furnham, A. (2006). Female physical attractiveness in Britain and Japan: A cross-cultural study. *European Journal of Personality*, 20, 69-81. doi: http://dx.doi.org/10.1002/per.568

Swami, V., Mada, R. & Tovée, M.J. (2012). Weight discrepancy and body appreciation of Zimbabwean women in Zimbabwe and Britain. *Body Image* 9, 559-662. doi: http://dx.doi.org/10.1016/j.bodyim.2012.05.006

Swami, V., Tovée, M.J. & Harris, A.S. (2013). An examination of ethnic differences in actual-ideal weight discrepancy and its correlates in a sample of Malaysian women. *International Journal of Culture and Mental Health*, 6, 96-107. DOI: http://dx.doi.org/10.1080/17542863.2011.643315

Swami, V. & Tovée, M.J. (2012). The Impact of Psychological Stress on Men's Judgements of Female Body Size. *PLoS One* 2012, 7, e42593. doi: http://dx.doi.org/10.1371/journal.pone.0042593

Tovée, M.J., Benson, P.J., Emery, J.L., Mason, S.M. & Cohen-Tovée, E.M. (2003). Measurement of body size and shape perception in eating-disordered and control observers using body-shape software. *British Journal of Psychology*, 94, 501-516. doi: http://dx.doi.org/10.1348/000712603322503060

Tovée, M.J., Edmonds, L. & Vuong, Q.C. (2012). Categorical perception of human female physical attractiveness and health. *Evolution and Human Behaviour*, 33, 85-93. doi: http://dx.doi.org/10.1016/j.evolhumbehav.2011.05.008.

Tovée, M.J., Maisey, D.S., Emery, J.L. & Cornelissen, P.L. (1999).Visual cues to female physical attractiveness. *Proceedings of the Royal Society B: Biological Sciences*, 266, 211-218. doi: http://dx.doi.org/ 10.1098/rspb.1999.0624

Tovée, M.J., Reinhard, S., Emery, J.L. & Cornelissen, P.L. (1998). Optimum body-mass index and maximum sexual attractiveness. *Lancet*, 352, 548-548. doi: http://dx.doi.org/ 10.1016/S0140-6736(05)79257-6

Tovée, M.J., Swami, V., Furnham, A., & Mangalparsad, R. (2006). Changing perceptions of attractiveness as observers are exposed to a different culture. *Evolution and Human Behaviour*, 27, 443-456. doi: http://dx.doi.org/ 10.1016/j.evolhumbehav.2006.05.004

Winkler, C. & Rhodes, G. (2005). Perceptual adaptation affects attractiveness of female bodies. *British Journal of Psychology*, 96,141-154. doi: http://dx.doi.org/ 10.1348/000712605X36343

Zar, J.H. (1984). Biostatistical analysis (4^{th} ed.). NJ: Prentice Hall.

Table Legends

Table 1: The Pearson correlations between the mean VAS scores for each image and their probability of being rated attractive, separately for male and female raters, for the single-shot data. All correlations were statistically significant at $p < .005$.

Table 2: The outcome of the dummy regression modelling in which mean VAS scores (the outcome) were predicted by WHR, BMI and dummy variables coding for the single-shot versus the multi-body methodologies.

Figure Legends

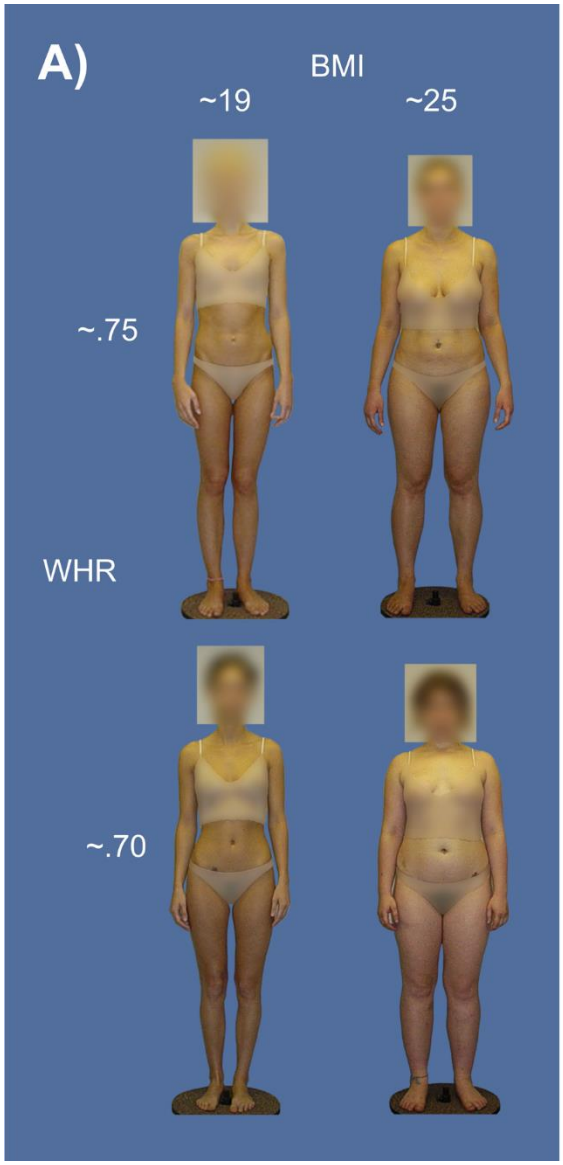
Figure 1: A) Examples of stimulus images varying in BMI (~19 and ~25) and WHR (~0.7 and ~0.75).Plots of mean VAS attractiveness per image, averaged across male and female observers, as a function of B) BMI and C) WHR. Red dots with a simple regression line through the data, in red, represent the single-shot data. Blue dots with a simple regression line, in blue, represent the multi-body data. Pink and cyan shaded regions represent the 95% CI for the single-shot and multi-body simple regressions of attractiveness on BMI and WHR respectively. Please note that these confidence intervals are derived individually from these simple regressions, and not the dummy regression model.

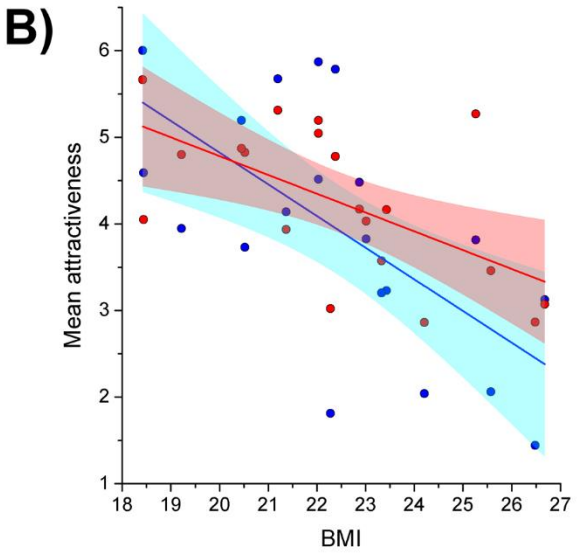
Table 1

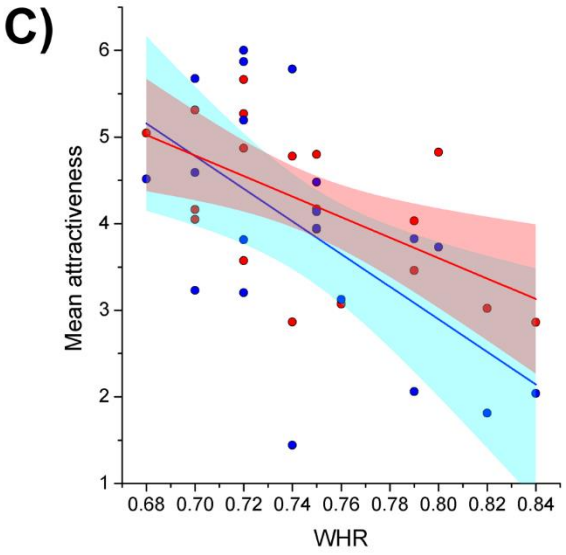
	Male VAS	Male 2AFC	Female VAS
Male 2AFC	0.97		
Female VAS	0.66	0.65	
Female 2AFC	0.63	0.61	0.90

Table 2

Explanatory variable	Term in equation	Parameter estimate	t-value	p-value	95% CI
Intercept	β_0	21.36	6.44	< .001	14.62 – 28.10
Dummy intercept	Δ	-6.31	-1.35	.19	-15.84 – 3.22
BMI	β_1	-0.30	-3.74	< .001	-0.46 – -1.34
Dummy BMI	β_2	0.12	1.11	.28	-0.10 – 0.35
WHR	β_3	-14.43	-3.26	.003	-23.44 – -5.42
Dummy WHR	β_4	5.16	0.82	.42	-7.59 – 17.90


A)
BMI
~19
~25
~.75
WHR
~.70


B)
Mean attractiveness
BMI
18 19 20 21 22 23 24 25 26 27
1 2 3 4 5 6


C)
Mean attractiveness
WHR
0.68 0.70 0.72 0.74 0.76 0.78 0.80 0.82 0.84
1 2 3 4 5 6