An independent analysis of race differences in ratings of attractiveness in the Add Health Study

Jelte M. Wicherts

&

Scott Barry Kaufman

Note: The main findings of this report were reported in Scott Barry Kaufman's Blog on Psychology Today on May 21st, 2011.

This version: May 24, 2011

Correspondence should be addressed to: Dr. Jelte M. Wicherts, j.m.wicherts@uva.nl

In his controversial blog post entitled: *"Why Are Black Women Rated Less Physically Attractive Than Other Women, But Black Men Are Rated Better Looking Than Other Men?"* (Psychology Today website, May, 15, 2011) psychologist Satoshi Kanazawa from the London School of Economics (LSE) concluded that he had found that African American women were "objectively" less attractive than European American, Asian American, and Native American women. The results he presented were based on analyses of publically available data from the longitudinal Add Health study. The purpose of the current note is to present the results of an independent re-analysis of the Add Health data set in order to see whether Kanazawa's results are replicable.

Method

We downloaded the publically available datasets from the Add Health website, for Waves 1 through 4. Note that Kanazawa's analyses were limited to Waves 1 through 3. The variable biological sex (BIO_SEX) from Wave 1 was used to select female participants for the analyses. We used the following variables to make for each wave a categorization into four racial groups: S1Q4A RACE-WHITE-W3, S1Q4B RACE-BLACK/AFRICAN AM-W3, S1Q4C RACE-AMER INDIAN/NATIVE AM-W3, S1Q4D RACE-ASIAN-W3. The results of this categorization were almost identical for Waves 3 and 4 (all but 20 out of 2604 cases; 99.03%), but slightly different for Wave 1 (e.g., Wave 3 and Wave 1 were identical of 93.4% of the 2473 cases). We used the racial categorization from Wave 3 and 4 for analyses of these waves, but Wave 3 categorizations for the analyses of Waves 1 and 2 (so as to not have changing categorizations). We call these groups European American, African American, Native American, and Asian American.

Attractiveness ratings of participants were made by interviewers during home interviews during all four waves. Ratings were made on a 5-point scale with the scales representing 1= Very Unattractive, 2=Unattractive, 3=About Average, 4=Attractive, and 5=Very Attractive. In each wave, interviewers rated between 1 and 41 female participants. Sex, age, and race of the interviewers do not appear to be part of the publically available data and interviewers are simply indicated by an unique Interviewer ID in each wave.

Statistical Analyses

Missing data were dealt with simply by using pair-wise deletion and participants were not weighted by sampling weights. We note that both choices are not ideal for various reasons, but we wished to follow Kanazawa's analyses closely and he did not indicate to have

dealt with these issues in the Blog or in his published paper in which he analyzed the same data for different purposes (Kanazawa, 2011, Intelligence). Because interviewers typically made numerous ratings of different participants, the ratings of these participants are not statistically independent. Consequently, standard statistical tests will be too liberal (i.e., show p-values that are too small) because dependencies will lead to underestimates of standard errors and within-group Mean-Squares. Data are not strictly normal, but to keep the analyses straightforward, we use a mixed ANOVA with interviewer as a random effect and race (4 levels) as a fixed effect. Besides the ANOVA we also report a regular 5 by 4 cross-tab independence analysis but note that because of the violation of independence, the results will lead to inflated chi-squares. We used an of alpha = .05 throughout without corrections for multiple testing.

Results

First, we run for each of the four waves the Mixed-Effects ANOVA.

Tests of Between-Subjects Effects

Dependent Variable: Rating of attractiveness at Wave 1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	7470.896	1	7470.896	8153.163	.000
	Error	1446.847	1578.978	.916[a]		
Race3	Hypothesis	3.194	3	1.065	1.395	.243
	Error	429.894	563.427	.763[b]		
INTID1	Hypothesis	533.081	422	1.263	1.614	.000
	Error	134.846	172.277	.783[c]		
Race3 * INTID1	Hypothesis	237.691	308	.772	1.043	.308
	Error	1382.743	1868	.740[d]		

Tests of Between-Subjects Effects

Dependent Variable: Rating of attractiveness at Wave 2

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	5949.115	1	5949.115	7948.001	.000
	Error	771.286	1030.436	.749[a]		
Race3	Hypothesis	2.531	3	.844	1.613	.185
	Error	249.714	477.481	.523[b]		
INTID2	Hypothesis	342.465	300	1.142	2.447	.000
	Error	38.404	82.307	.467[c]		
Race3 * INTID2	Hypothesis	116.246	231	.503	.880	.891
	Error	868.962	1520	.572[d]		

Tests of Between-Subjects Effects

Dependent Variable: Rating of attractiveness at Wave 3

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	7000.266	1	7000.266	9133.801	.000
	Error	1357.758	1771.574	.766[a]		
Race3	Hypothesis	3.313	3	1.104	1.524	.207
	Error	443.425	611.932	.725[b]		
INTID3	Hypothesis	356.144	381	.935	1.211	.070
	Error	142.993	185.178	.772[c]		
Race3 * INTID3	Hypothesis	214.189	285	.752	1.113	.110
	Error	1303.514	1930	.675[d]		

Tests of Between-Subjects Effects

Dependent Variable: Rating of attractiveness at Wave 4

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	5334.793	1	5334.793	5616.276	.000
	Error	821.912	865.279	.950[a]		
Race4	Hypothesis	5.667	3	1.889	2.563	.054
	Error	384.050	521.094	.737[b]		
INTID4	Hypothesis	418.913	278	1.507	1.825	.000
	Error	156.762	189.870	.826[c]		
Race4 * INTID4	Hypothesis	207.911	263	.791	1.243	.008
	Error	1122.617	1765	.636[d]		

The mixed effects ANOVA for each of the four waves, race differences are not statistically significant after taking into account variation in ratings due to the interviewers. In all but the third wave, the variance component due to raters was significant different from zero. Note that this analysis is not ideal because it assumes normality and homogeneous error variances, both of which are not expected to be tenable in these data. In addition, although the total sample size is fairly large, the number of participants per interviewer is neither constant nor large. Hence, the power to detect race differences in these analyses should be studied further.

It is important to note that for Waves 3 and 4, the (incorrect) standard ANOVA with a fixed effect for race group, showed main effects that were not entirely impressive. Although in Wave 3 there was a significant effect: $F(3, 2596) = 2.67$, $p = .046$, in Wave 4 it did not reach significance: $F(3, 2310) = 0.87$, $p = .455$. Even a minor correction for dependencies will render the result in Wave 3 insignificant.

The effect of the different raters is also readily apparent when one computes the Pearson Moment Correlations or Spearman Rank-order correlations between ratings of the same female Add Health Participants across the four waves.

Correlations

		attract1	attract2	attract3	attrac4
attract1	Pearson Correlation	1	.300**	.188**	.122**
	Sig. (2-tailed)		.000	.000	.000
	N	3353	2515	2622	2756
attract2	Pearson Correlation	.300**	1	.239**	.106**
	Sig. (2-tailed)	.000		.000	.000
	N	2515	2517	2071	2134
attract3	Pearson Correlation	.188**	.239**	1	.136**
	Sig. (2-tailed)	.000	.000		.000
	N	2622	2071	2624	2327
attrac4	Pearson Correlation	.122**	.106**	.136**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	2756	2134	2327	2759

**. Correlation is significant at the 0.01 level (2-tailed).

Spearman's rho Correlations

		attract1	attract2	attract3	attrac4
attract1	Correlation Coefficient	1.000	.334**	.217**	.161**
	Sig. (2-tailed)	.	.000	.000	.000
	N	3353	2515	2622	2756
attract2	Correlation Coefficient	.334**	1.000	.265**	.136**
	Sig. (2-tailed)	.000	.	.000	.000
	N	2515	2517	2071	2134
attract3	Correlation Coefficient	.217**	.265**	1.000	.181**
	Sig. (2-tailed)	.000	.000	.	.000
	N	2622	2071	2624	2327
attrac4	Correlation Coefficient	.161**	.136**	.181**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	2756	2134	2327	2759

**. Correlation is significant at the 0.01 level (2-tailed).

Given that the ratings were taken on the same persons, these correlations can be interpreted as showing quite poor inter-rater reliability. However, it should be noted that there was some time lag between ratings. However, Wave 1 and Wave 2 ratings were taken within a relatively short time and are still quite low. The ratings at Waves 3 and 4 were both done when participants were adults, but also show correlations that are too low for typical standards of inter-rater reliability.

Cohen's Kappa for all comparisons are given below and can be seen to be lower than .20 in all comparisons. According to Landis and Koch's (1977) guidelines, Kappa values below .20 are considered to show "slight agreement". Combined, the results show that there is little agreement in ratings over the four waves.

		attract1	attract2	attract3	attrac4
attract1	Cohen's Kappa	1.000	.196	.100	.076
	Asymp.SE	.	.015	.014	.013
	N	3353	2515	2622	2756
attract2	Cohen's Kappa	.196	1.000	.141	.075
	Asymp.SE	.015	.	.016	.015
	N	2515	2517	2071	2134
attract3	Cohen's Kappa	.100	.141	1.000	.099
	Asymp.SE	.014	.016	.	.015
	N	2622	2071	2624	2327
attrac4	Cohen's Kappa	.076	.075	.099	1.000
	Asymp.SE	.013	.015	.015	.
	N	2756	2134	2327	2759

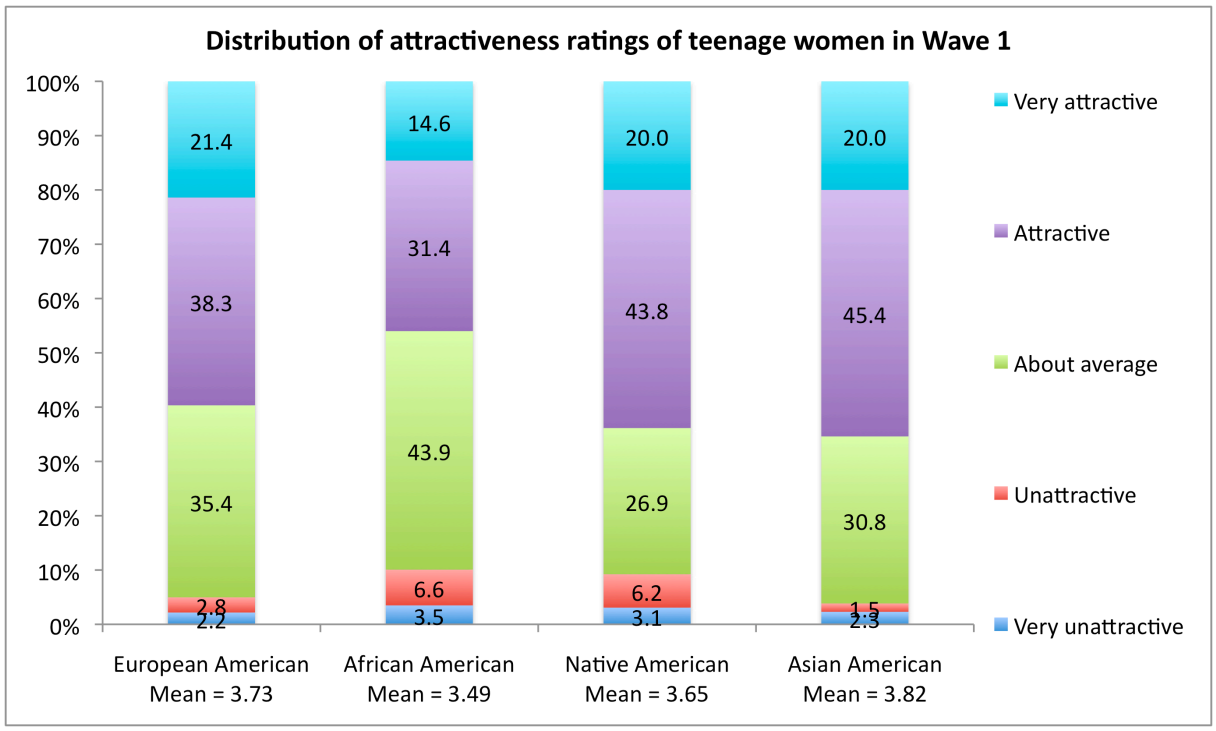
In Wave 1 participants were still teenagers (age in years: M = 15.9 years, SD = 1.8) and there appears to be some differences in terms of the distribution of the ratings of attractiveness.

Race3 * attract1 Crosstabulation

Count

		attract1					
		Very unattractive	Unattractive	About average	Attractive	Very attractive	Total
Race3	European American	40	52	598	666	360	1716
	African American	23	47	277	208	102	657
	Native American	3	4	46	39	23	115
	Asian American	4	4	29	49	28	114
Total		70	107	950	962	513	2602

When tested without taking into account the dependencies, Pearson's Chi-square is 50.4, *DF* = 12, $p < .001$.



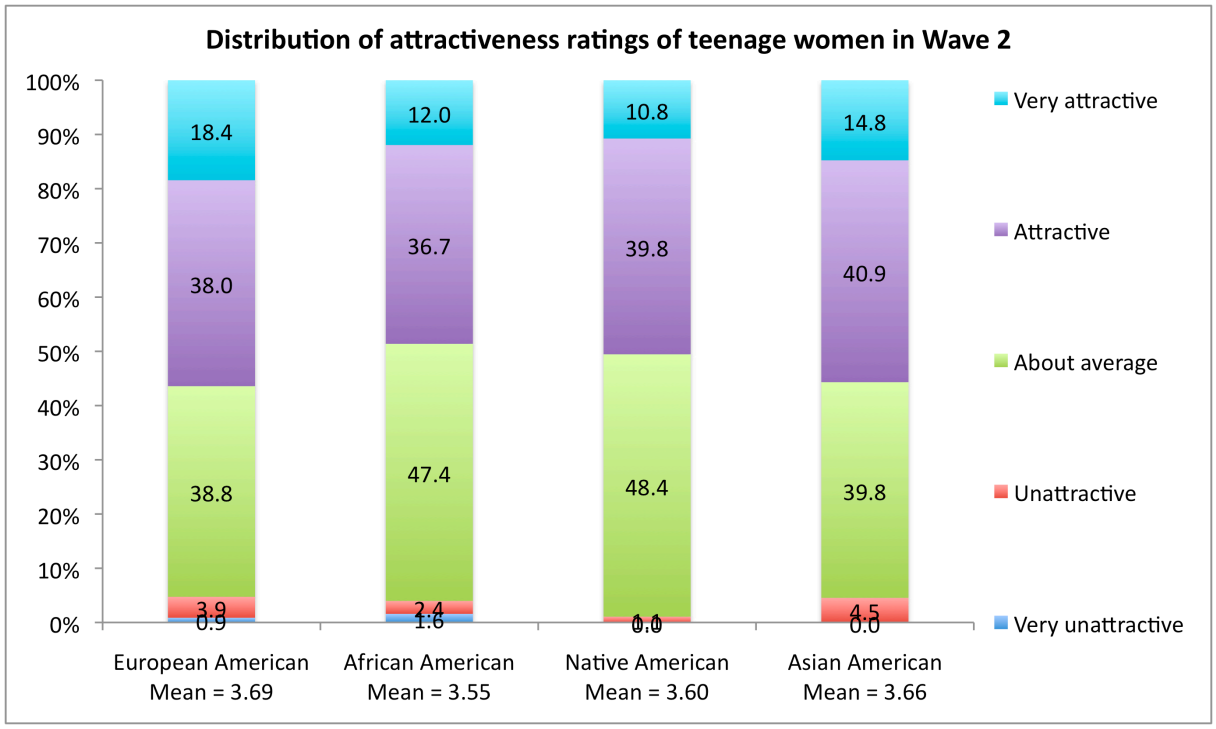
In Wave 2, average age of the female participants was 16.5 (SD = 1.6) and the race differences become much less pronounced.

Race3 * attract2 Crosstabulation

Count

		attract2					
		Very unattractive	Unattractive	About average	Attractive	Very attractive	Total
Race3	European American	12	53	533	521	253	1372
	African American	8	12	238	184	60	502
	Native American	0	1	45	37	10	93
	Asian American	0	4	35	36	13	88
Total		20	70	851	778	336	2055

When tested without taking into account the dependencies, Pearson's Chi-square is 27.8, *DF* = 12, *p* =.006.



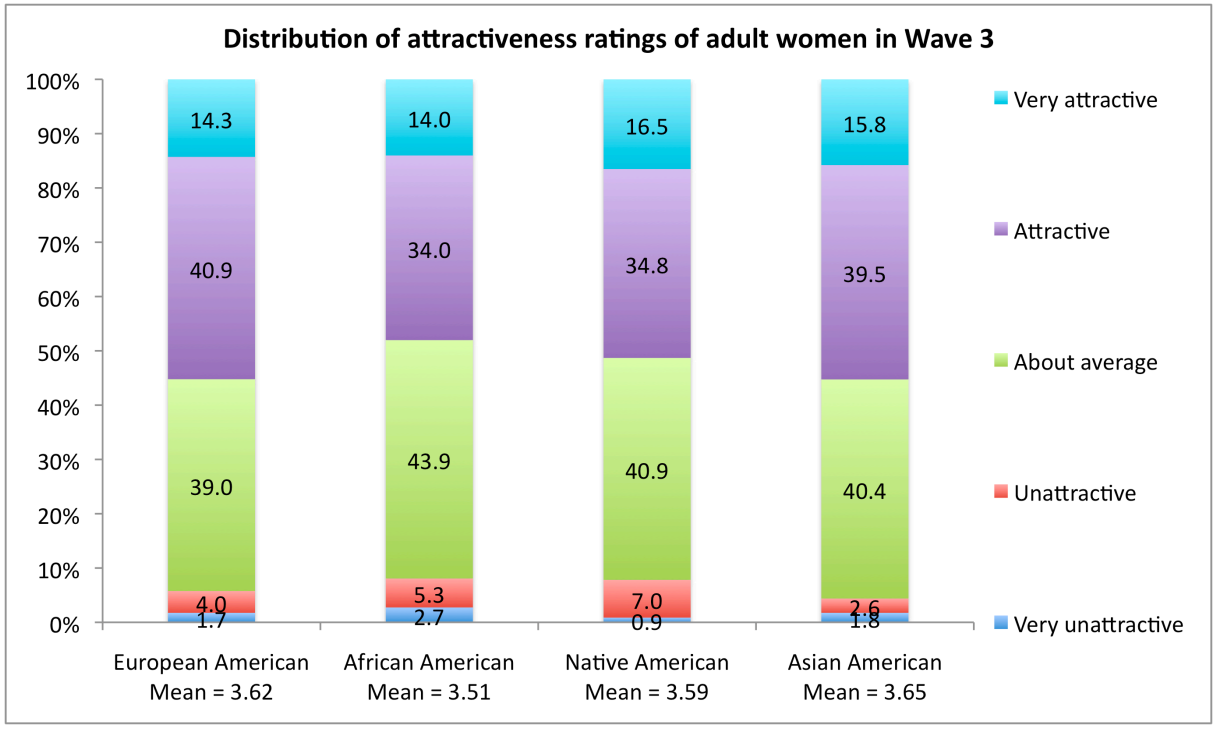
In Wave 3, all participants reached legal adulthood (age in years: M = 22.1, SD = 1.8) and Kanazawa's results no longer hold.

Race3 * attract3 Crosstabulation

Count

		attract3					
		Very unattractive	Unattractive	About average	Attractive	Very attractive	Total
Race3	European American	30	69	669	702	245	1715
	African American	18	35	288	223	92	656
	Native American	1	8	47	40	19	115
	Asian American	2	3	46	45	18	114
Total		51	115	1050	1010	374	2600

When tested without taking into account the dependencies, Pearson's Chi-square is 17.3, *DF* = 12, *p* = .138.



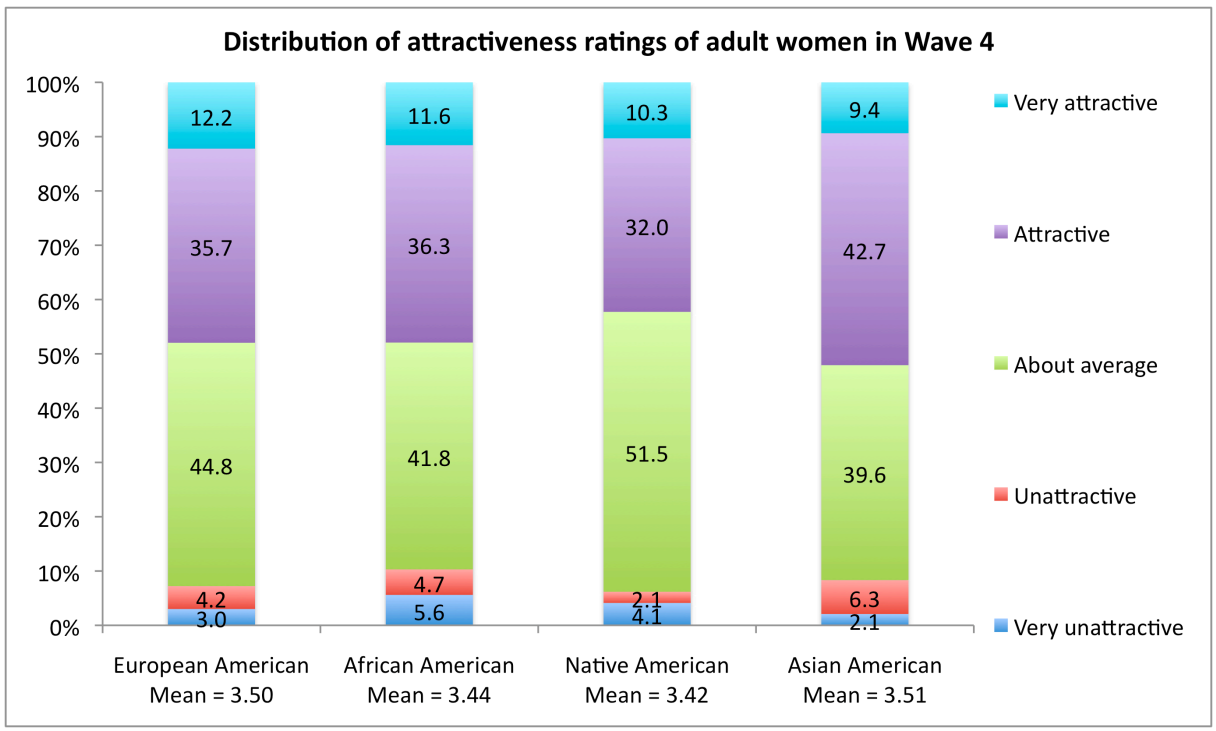
In Wave 4, all participants are adults (age in years: M = 28.9, SD = 1.8) and there are no systematic differences between the four groups in terms of ratings of attractiveness.

Race4 * attrac4 Crosstabulation

Count

		attrac4					
		Very unattractive	Unattractive	About average	Attractive	Very attractive	Total
Race4	European American	47	66	701	559	191	1564
	African American	31	26	231	201	64	553
	Native American	4	2	50	31	10	97
	Asian American	2	6	38	41	9	96
Total		84	100	1020	832	274	2310

When tested without taking into account the dependencies, Pearson's Chi-square is 15.6, *DF* = 12, $p = .210$.



Conclusions

Kanazawa claimed to have found that "black women are […] far less attractive than white, Asian, and Native American women" and that "In each wave, black women are significantly less physically attractive than women of other races". He also claimed that the ratings were "objective". Neither of these claims follow from the current re-analyses of the ratings of female participants in the Add health study, because (1) there was substantial variation in ratings made by different interviewers, (2) the ratings across the waves show very little convergence, (3) the ratings in which Kanazawa did find clear differences (on the basis of an analysis that did not take into account dependencies due to raters) were made when Add Health participants were not adults but teenagers, (4) the ratings that were done of attractiveness of adult women did not show overall systematic differences between the four racial groups.

The analyses presented here are by no means perfect and it is clear that additional statistical work is needed to fully come to grips with the statistical intricacies of the Add Health data that have a bearing on race differences in attractiveness ratings. However, the mixed effects ANOVAs that dealt with the statistical dependencies (that Kanazawa ignored) are expected to be quite robust to model violations and showed clear results that (for the most relevant adult data in Waves 3 and 4) converged with more traditional analyses. Given the pattern of the results in Waves 3 and 4, there appears to be little empirical ground for the claim that as adults African American women are rated to be less attractive on average than women from other racial groups.

Kanazawa interpreted his results incorrectly as having a bearing on attractiveness of women because the ratings were taken mostly when Add Health participants were teenagers. Kanazawa did not include the most relevant data to test his hypotheses (Wave 4). Statistically speaking his analyses are incorrect because he did not take into account the clear statistical dependencies that exist because of variation due to raters. Despite these problems, the means he presented in his figures were quite close to the ones we computed.