RESEARCH PAPER

A Smile is Just a Smile: But Only for Men. Sex Differences in Meaning of Faces Scales

Achim Elfering · Simone Grebner

Abstract This study tests psychometrics of the faces scale that is often used in organizational psychology to assess emotions, attitudes, and well-being. In analyzing 10,584 two-categorical judgments ("sad" versus "happy") of 11 faces from 72 participants, women judged a face with a horizontal line as mouth ("neutral face") more often sad than happy and significantly more often sad than men did. Moreover, women adapted their judgments to the range of faces under study in showing a significant contrast effect. In a range of neutral to bright smiling faces a face showing a mild smile was judged to be sad, while in a range of neutral to heavy griming faces a mild grim was judged to be happy. Sex differences in meaning of faces within faces scales and their implications for the use of the faces scale in work and organizational psychology are discussed.

Keywords Field research methods · Survey research · Equivalence testing

1 Introduction

Since Kunin (1955) developed a faces scale to measure general job satisfaction, the psychometric validity of faces (i.e. 'smileys') as response options was proven in measuring general job satisfaction (Kunin 1955, 1998; Niklas and Dormann 2005; Wanous et al. 1991). The faces scale was also recommended for the assessment of well-being (Andrews

---

In this article, we use the term *sex differences* in the context of comparing the two sexes. The term *gender* is most often used in the context of societal norms and roles associated with each sex.

A. Elfering (✉)
Department of Psychology, University of Bern, Muesmattstr. 45, 3000 Bern 9, Switzerland
e-mail: achim.elfering@psy.unibe.ch

A. Elfering
National Centre of Competence in Research, Affective Sciences, CISA, University of Geneva, Geneva, Switzerland

S. Grebner
Central Michigan University, Mt. Pleasant, USA

and Withey 1976) and sensory judgements of various stimulus qualities (Chen et al. 1996). When Kunin (1955) validated his faces scale to measure general job satisfaction, he noted that individuals differ in evaluation of the 'neutral face' in the centre of the scale. These individual differences in evaluation were left unexplained. Kunin (1955) attributed this variability to his particular validation procedure involving a separate validation of smiling and frowning faces in two subsets. He did not test for person variables to explain variability in judgement and 'no effort was made to separate men and women or to limit the testing to a particular age or occupational group' (p. 71). This study tested whether systematic differences in evaluation of faces in the centre of a faces scale are related to gender and evaluation of restricted scale ranges.

1.1 The Mouth as an Indicator of Mood

When individuals respond to a faces scale they are asked to match their emotions or attitudes on a scale showing faces with only the curvature of the mouth line varying systematically from a large smile (concave arc) to a frown (convex arc). The mouth within faces was shown to be related to the recognition of the happy versus sad emotion dimension (Ekman 1972; Izard 1971) independently of other aspects of the faces (Keren and Lewis 1993). Meanwhile, the adherence to the mouth line as an indicator of emotional well-being seems to be restricted to Western culture, where people tend to express their emotions. In Eastern culture, positive mood is coded in the eye section of the face (Yuki et al. 2007).

Indeed, use of faces scales could lead to interpretation problems when individuals differ in their evaluation of faces. Kunin (1955) tested eleven different faces separately in two sets. Because of technical constraints Kunin's participants simply judged faces in a 'happy' and a 'sad' faces set, each including the neutral face, and Kunin expected that participants would place the neutral face in the end of the scale. He found 'in some cases, however, other [than the neutral] faces were placed in one of the ends of the scale [of the numerical rating scale]' (p. 73), indicating that the separate judgement of smiling and frowning faces caused this difference. Kunin (1955) did not trace the phenomenon in more depth, but averaged the two judgements of the neutral faces for further analysis. He did not report whether cultural differences and or sex were systematically related to variability in judgements. There is, however, reason to suggest that the use and understanding of faces as response options do depend on individual differences in emotion expression and emotion recognition that may differ systematically between cultures (Yuki et al. 2007) and between men and women (LaFrance et al. 2003). Emotion expression and emotion recognition may relate to understanding of faces in faces scales if one considers the process of judgement that Kunin himself described as when 'the respondent is required only to check the face, which "looks like he feels" rather than verbalizing his feelings' (Kunin 1998, p. 823). Consequently, one should expect sex-related differences in emotion expression and emotion recognition to be transferred to use of the faces scale. Sex differences in the cognitive processes involved in forming response, i.e. recognition, labelling and reporting, were often found in health-related quality-of-life assessment (Barofsky 2000). Thereby, information processing in women was more complex than in men; for instance, judgements on self-reported health in women compared with men included a wider range of health-related and non-health-related factors (Benyamini et al. 2000). It is a primary goal of this study to disentangle possible mechanisms that lead to variance in judgements concerning the centre of faces scales. Two influences have been mentioned so far: (1) the use of a restricted set of only smiling or only frowning faces; and (2) sex-related differences in recognition of faces. Hence, the research questions are:

(1) Do individuals differ considerably in judgements of the 'neutral face' as the centre within faces scale?

(2) Are these individual differences in judgement systematically related to sex?

1.2 Theoretical Explanation of the Test Procedure Effect which Kunin Observed

Kunin observed differences in judgement when subjects rated range-restricted subsets of faces, i.e. scales with only smiling or only frowning faces. Range effects in judgements are common (Sarris 2004). By the beginning of the last century, Hollingworth (1910) had already demonstrated that judgements were shifted toward a mean magnitude in the respective mental representation of subjective stimulus scales by underestimation of squares above the mean magnitude, and vice versa. He called the mean magnitude the indifference point (IP). Ideally, the 'neutral midpoint' would represent the IP in the psychophysical sense, the stimulus value that is equally likely to be judged as happy or sad (Sarris 2004). Beside the degree of individual differences in IP (research question 1) and sex-related differences in correspondence of the IP to the midpoint of the scale (research question 2), another important question is whether the IP differs in faces scales that include either smiling or frowning faces as response options. For measurement purposes, the restriction to smiling or frowning faces may sometimes be useful. For instance, faces scales that have only smiling faces might be considered useful to avoid ceiling effects in measuring positive emotion. On the other hand, faces scales that show only frowning faces are frequently used to measure pain in children because lack of pain does not imply happiness when they are suffering from disease (Chambers et al. 1999). Whether the response options of a scale are composed symmetrically around the 'neutral face' or asymmetrically, i.e. with predominantly positive or negative options, presumably affects the evaluation of the 'neutral face'. Within the theoretical framework of frame of reference theory (Sarris 2004), the evaluation of the 'neutral face' depends on the remaining response options that function as a contextual background or frame of reference (Sarris 2004). Contextual background should induce a *contrast effect*. In assuming a contrast effect, we expect that a small smile would appear comparably less happy when most response options show a more intense smile, compared with a series of faces, where the face with a small smile is presented amongst faces that show different degrees of frowning faces. The third research question therefore addresses whether the range-restricted subsets of faces, i.e. scales with only smiling or only frowning faces, lead to systematic differences in evaluation of faces around the centre of the scale. The hypotheses are:

(1) Individuals differ considerably in IP.

(2) Men and women differ in judgements of 'neutral faces' in faces scales. We expect sex differences in IP.

(3) The IP differs systematically between stimulus contexts ('positive', 'neutral', 'negative'), showing a psychophysical contrast effect.

2 Method

2.1 Participants

Seventy-two undergraduate psychology majors (48 women and 24 men, age $M = 24.4$ years, SD $= 2.8$ years) volunteered to participate. Except for correctional glasses ($n = 24$),

participants had no vision problems and were naive with respect to the hypotheses. Participants were German-speaking students of a university in Switzerland.

2.2 Materials

The stimuli were printed on paper. They included eleven faces ranging from a broad smile to a deep scowl. The faces differed only in the curvature of the mouths. Five faces showed raised corners of the mouth in steps of ten angular degrees ($10°$ to $50°$) and a further five faces showed hanging corners of the mouth ($-10°$ to $-50°$). One face showed a horizontal line as mouth. This face is called the 'neutral face' and is mostly used as midpoint of the scale (Niklas and Dormann 2005).
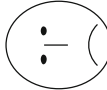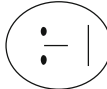
There were three contextual conditions (the 'neutral', 'negative', and 'positive' conditions). The 'neutral' condition included three faces with hanging corners of the mouth (i.e. 'sad' faces), the 'neutral' face and three faces with raised corners of the mouth (i.e. 'happy' faces). In the 'negative' condition, there were five 'sad' faces, the neutral and one 'happy' face. In the 'positive' condition, the range of faces of the 'negative condition' was mirrored, including only a 'sad' face, a 'neutral' face, and five 'happy' faces (Table 1).

Each contextual condition was printed on a separate sheet of paper with 49 faces printed in a seven-by-seven matrix. Every face was shown seven times in a latin-squared balanced order of positions and neighbouring stimuli. Each of the three sheets of 49 faces was preceded by a 'calibration sheet' showing sixteen faces, including two different faces, repeated in balanced order: a sad (-20 angular degrees) and a happy face (20 angular degrees). The 'calibrating sheets' were used to prevent carry-over effects between contextual sheets. Every sheet's headline asked: 'How do these people feel?' For every face, two response alternatives were printed below the face, including 'sad' on the left, and 'happy' on the right. In addition to the six sheets, an introductory sheet asked for age, sex, visual problems, use of correctional glasses and emo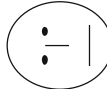tional state of well-being ("how do you feel at the moment?", with six verbally described response options: 'very sad' (1), 'sad' (2), 'a bit sad' (3), 'a bit happy' (4), 'happy' (5) and 'very happy' (6). The last paper sheet asked again for the time and comments.

2.3 Procedure

The experiment was conducted with small groups of five to ten students. The instruction was given by word-of-mouth. Participants learned that the experiment was on stimulus perception and included recognition of emotional expression in faces. The researcher also stressed that judgements were not related to any aspect of (emotional) intelligence or performance. Participants were instructed to follow their immediate impulse. There was no speed instruction or time limit. Participants were asked to judge all faces from the upper to the lower rows from left to right. They were also told not to inspect or change prior judgements on sheets they finished earlier but to make only immediate corrections of their judgements. Participants judged the faces by marking the response alternatives with a pencil. Following the first 'calibration sheet', the experiment started with the 'neutral' contextual condition. The order of the 'negative' and 'positive' context condition was counterbalanced across participants. On average, the experiment lasted 8 min (SD = 2.4 min).

Table 1 Experimental stimuli in two-alternative two forced-choice training and generalization tests

Stimuli (angular degrees)	−50	−40	−30	−20	−10	0	10	20	30	40	50
Training Stimuli											
** Control series**											
** Neutral**			☺	☺	☺	☺	☺	☺	☺		
Context series											
** Negative**		☺	☺	☺	☺	☺	☺				
** Positive**					☺	☺	☺	☺	☺	☺	☺
(Training)	☺			☺				☺			

2.4 Design

The experimental design included a within factor. Each participant judged faces that were presented in three contextual sets ('neutral', 'negative' and 'positive') with seven stimuli in each set. The experiment also included a comparison of judgements by sex.

2.5 Calculation of the Indifference Point

The IP marks the face that is equally frequently judged as 'sad' or 'happy'. Hence, the IP is the (asymptotic) face that receives 50% 'happy' judgements. The unit of IP is the angular degree of the mouth line, and the IP is calculated in this experiment from 49 individual judgements on seven different faces for each contextual set. Figure 1 shows an individual response distribution and the IP for the 'neutral' set of faces.

2.6 Equivalence Testing

Equivalence in judgements of men and women on the neutral face was tested with the non-central t-test by Anderson and Hauck (1983). This t-test does not test the differences of two means for significant difference but for difference less than a predefined allowable maximum (for an example see Bi 2005). The maximum allowable difference in the neutral face was predefined as the standard deviation of judgements in the total sample.

2.7 Data Analysis

Data contain information at the person level and the judgement level, with judgements nested within persons. This leads to statistical problems. On the one hand, a simple aggregation of judgement-related information at the person level would lead to loss of information and power. On the other hand, analyses on the judgement level (disaggregated data) would lead to a 'blown up' data set regarding the higher-level units (persons), and
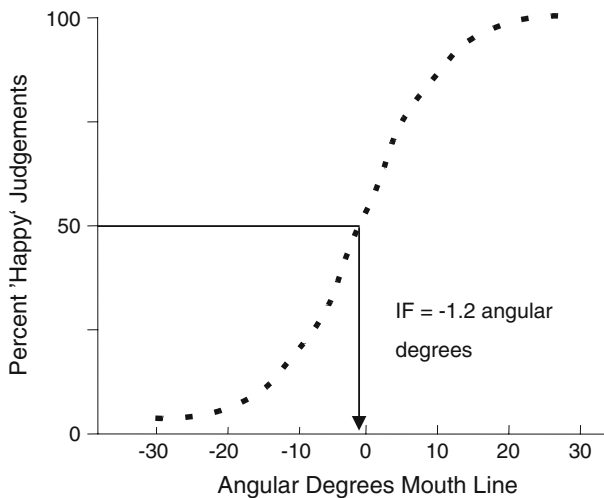

Fig. 1 Determination of the indifference point (IP) from 'sad' versus 'happy' judgements on faces

spurious significances could result (see Hox 2002). In consideration of the nested data, a multilevel linear model approach was employed that allows for testing of the influence of judgement-related variables and person-related measures, and cross-level interactions of judgement and person-related variables. The dependent variable in multilevel regression analysis was the IP, one IP for every contextual set of faces. For 72 participants judging three contextual sets of faces, the total number of IP was 216. The IP was regressed on contextual set (coded as −1 for 'negative', 0 for 'neutral', and 1 for 'positive' sets of faces) as predictor variable on IP level (Level 1), and sex (coded 0 for women, and 1 for men), the duration the experiment took (grand mean centred), correctional glasses (coded 0 for no correction, and 1 for correction), and baseline mood (grand mean centred) as predictor variables on participant level (Level 2). The two-way interaction between contextual set and sex was also included into the model to test for potential sex differences in context effects.

3 Results

Generally, there were fewer 'happy' judgements than 'sad' judgements (mean percentage of happy judgements = 45%, SD = 8%), and this difference was more pronounced in judgements of women (mean = 43%, SD = 8%) than of men (mean = 48%, SD = 8%). This sex difference in use of response alternatives was statistically significant ($t(70) = -2.94, P = .004$). The only other sex-related difference was in age, with men being significantly older than women (mean age in men was 25.9 years versus 23.7 years in women, $t(70) = -3.58, P = .001$). There were no significant differences in mean levels of baseline mood at the beginning of the experiment. The average baseline mood level was 4.08 in men, (SD = .72), and 4.40 in women (SD = .68; $t(70) = 1.81, P = .074$). The non-significant correlation between sex and baseline mood was −.21. When correlations on IP-level ($n = 216$) were examined, sex showed some statistically meaningful associations with other study variables. Sex was negatively associated with IP ($r(216) = -.29, P < .001$), i.e. the IP was more positive in women. The expected positive association between IP and context was small ($r(216) = .13, P = .066$).

3.1 Individual Differences in IP

Evidently individual variation in levels of IP was large, and it was large compared with the effects of contextual sets of faces. Indeed, in multilevel analysis, the variance components model—the model only includes the constant, which is free to vary on both levels, but no predictor variables—showed that 72% of the variation in IP can be attributed to the person level (ICC = .72, Table 2).

3.2 Sex Differences in IP

Figure 2 shows the mean percentages of 'happy' judgements for the midpoint face of the scale, and the smallest smile, respective lightest frown. Best correspondence would be indicated by percentages around 50% for the midpoint face. It is apparent from Fig. 2 that the neutral face was more often judged to be sad rather than happy. Out of 504 judgements on the neutral face in the 'neutral' contextual condition (seven judgements each from 72 participants), there were 365 "sad" (72%) and 139 "happy" (28%) judgements. Judgements significantly deviated from chance expectation in binomial test with a probability of

Table 2 Linear multilevel model predicting the IP in judgements of faces

Intraclass-correlation	.72			
Predictor variables				
Fixed effects	*Unstandardised Param.*	*SE*	*Standardised Param.*	*P*
Intercept	5.153	4.527		
Level 1 (Person)				
Sex	−3.603	1.317	−.286	.006
Duration	.060	.247	.024	.814
Correctional glasses	.211	1.289	.017	.872
Mood	−.169	.876	−.020	.849
Level 2 (Judgements)				
Context	1.283	.302	.177	.000
Level 1 × level 2 interactions				
Context × sex	−1.120	.524	−.090	.032
Random effects				
VAR intercept person level	22.590	4.267		.000
VAR intercept series level	8.773	1.034		.000
2*log likelihood (IGLS)	1238.040			

Note: Sample size $N = 216$ IF from (IF aggregated from 10,584 stimuli judgements made by 72 participants). Unstandardised Param. = unstandardised fixed regression parameter estimates; SE = standard error; Standardised Param. = standardised fixed regression parameter estimates; IGLS = Iterative Generalised Least Squares. After the standard errors the two-sided significance level of the Wald-Test (parameter estimates/standard error) is indicated. Random Effects = Variance and covariance estimates of parameters that are allowed to vary on level 2. Codings for sex: 0 = female, 1 = male; correctional glasses: 0 = no, 1 = yes; context: −1 = sad faces (negative) (range −50° to 10°), 0 = symmetrical faces condition (−30° to 30°), 1 = happy faces (positive) (−10° to 50°)

.5 ($P < .001$). The fit to a binomial chance model differed between men and women. In men there were 57% 'sad' and 43% 'happy' judgements, while there were 80% 'sad' and 20% 'happy' judgements in women. In separate binomial tests, judgements did not differ from chance level in men ($P = .076$), but differed significantly from chance level in women ($P < .001$).

The non-central *t*-test (Anderson and Hauck 1983) was used to test whether recognition of the neutral face was equivalent in men and women. The alternative hypothesis in equivalence testing stated that sex-related differences in response were less than the similarity limit that was set to the standard deviation of the total sample. Because the non-central *t*-test revealed a *P* value of $P = .15$ that was larger than the risk of concluding similarity when the two response patterns are in fact different (α was set to 5%), we must reject equivalence in recognition of the neutral face.

3.3 IP in 'Positive', 'Neutral', and 'Negative' Subsets of Faces

The third hypothesis addressed systematically differences in level of IP between stimulus contexts ('positive', 'neutral' and 'negative'). A psychophysical contrast effect would suggest that a mild smile appears happier in a 'negative' than a 'positive' situation. Accordingly, a slightly sad expression will be judged to be sadder in a 'positive' than a 'negative' context. In terms of the IP, the contrast effect turns out in an IP deviation in the direction of the predominant faces within a context: in a negative situation, the IP should
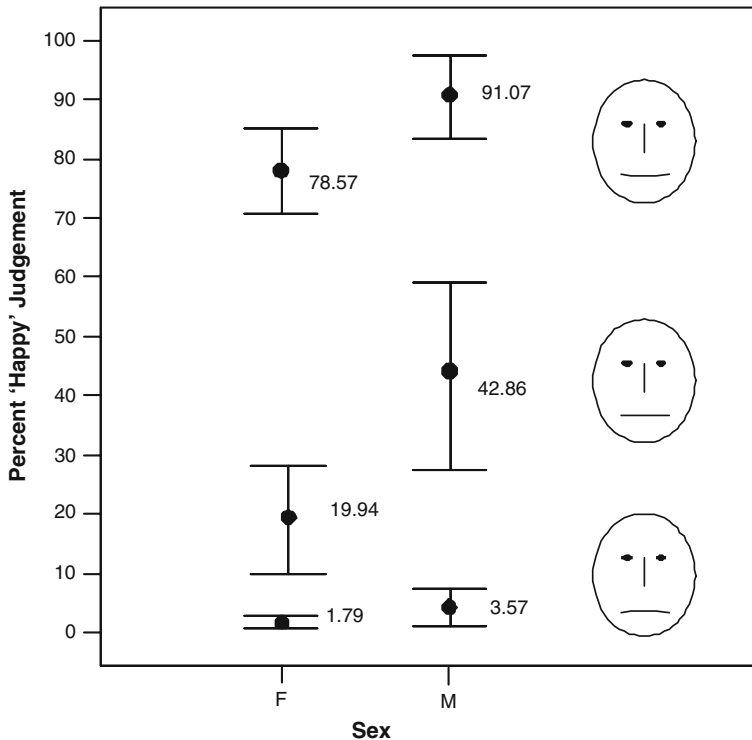
Fig. 2 Men's and women's mean percentage and 95% CI for 'happy' judgements on three faces in the neutral context condition

be lower than in the positive situation. What is individually perceived as 'neutral' is a sadder expression at a negative and is a comparably happier expression at a positive. Hence, the mean IP levels of different contextual situations differ in the expected directions ('negative situation' mean level = 2.98 angular degrees, SD = 5.40; 'neutral situation' mean level = 3.79 angular degrees, SD = 5.66; 'positive situation' mean level = 4.80 angular degrees, SD = 6.62). Although significant as a linear trend factor in multilevel analysis (B = 1.28, SE = .30, $P < .001$, Table 2), the magnitude of the effect is small compared with the large individual differences in IP. Moreover, a significant interaction between sex and context (B = − 1.12, SE = .52, $P < .05$) indicates that the contrast effect of scale range is restricted to women (Fig. 3).

4 Discussion

This study tested the measurement equivalence of the faces scale response options. Results showed considerable variation in judgements between individuals (individual differences account for 72% of variation in judgements). Women judged faces in the centre of the scale to be 'sadder' than men did, and only women judged faces in the centre of the scale to be 'sadder' within a 'positive' set of faces than in a 'negative' set, and vice versa.

In the process of judging, subjective values must be translated to the response system of the questionnaire item (Wedell 1990). For instance, when we are asked for our level of life

Fig. 3 Mean IP and 95% CI of facial expression by men and women in three context conditions
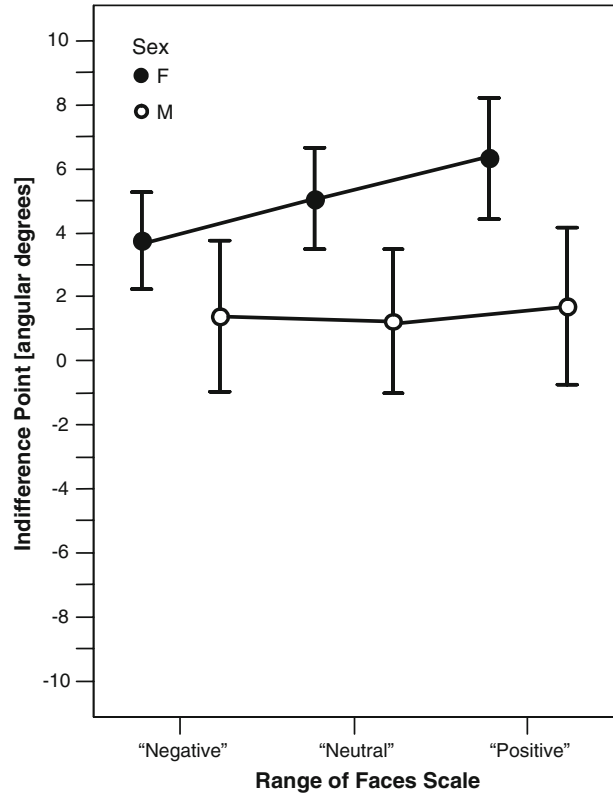
satisfaction, our internal impression of satisfaction must be translated to a matching numerical value that is verbally anchored. Participants using a faces scale match their own emotional state, level of agreement or job satisfaction to the expression of faces. Therefore, as regards faces scales it is central to psychometrics that the response alternatives have the same meaning for respondents. Results from this study showed that there is a lack of agreement in evaluations of the faces in the centre of the faces scale, and individual variability is strong. Kunin (1955) also found large individual variability. Interestingly, overall variability in nine of eleven response options in his male faces scale was less than in his abstract circular faces scale, and he therefore recommended the male faces scale for use (and later work showed equivalence of the male faces scale and a female faces scale irrespective of the sex of raters, Dunham and Herman 1975). Because of Kunin's practice of averaging the neutral faces judgements from two scale sets, however, interpretation of the variability of the neutral faces is difficult.

4.1 Consequences of Sex Differences in Evaluation of Faces

In this study, there were systematic sex differences in judgements of the neutral face. Women, but not men, judged neutral faces merely as sad. Hence, the same internal impression of, for example, job satisfaction is matched differentially: a relatively neutral internal impression is more likely to be translated to the midpoint face of the faces scale by men, while women are more likely to chose a slight smile. Consequently, in measuring job

satisfaction women may show higher levels of job satisfaction than men even when there is no real difference in true levels. Indeed, higher job satisfaction in women than in men is reported in the literature and is known as the 'satisfaction paradox' (Kim 2005). Higher levels of job satisfaction in women did contrast with self-reports of working conditions that were less favourable than in men (Kim 2005). Most of these differences mostly disappeared when working conditions and rewards were controlled for (Lefkowitz 1994), and were small when they remained. We do not know, however, whether studies that showed the 'satisfaction paradox' used the faces scale more frequently. It is, however, concluded from this study that the faces scale may induce spurious differences in mean levels that may result from sex-related differences in perception and evaluation of response options. If so, correlations between satisfaction/well-being measures and other measurements may be overestimated when some of the variation in satisfaction is sex-related by a difference in information integration, and this difference in information integration or underlying factors also affects the other measure, for instance, of positive and negative affectivity. Indeed, a recent meta-analysis showed that associations between job satisfaction and positive affectivity depended on the job satisfaction questionnaire instrument, with faces scales showing stronger associations with negative affectivity than other scales (Kaplan et al. 2008). Furthermore, the bias of low measurement equivalence of the neutral face should be comparably lower in highly skewed measures of job satisfaction and/or life satisfaction, simply because the neutral face is rarely chosen, because of a high mean satisfaction. Instead, whenever a normal distribution of judgements around the neutral face is more common, e.g. in measurement of agreement to complex issues, the bias from low sex-related equivalence is potentially higher. As a consequence, when normal distribution around the centre of the scale is expected, and especially when sex differences are of interest, the researcher should rely on multiple types of scales to measure the construct of interest. Further research should address the consequences arising from the finding that faces scales tend to be more symmetric for men than for women.

4.2 Sex-related Differences in Information Integration during Evaluation of Faces

Potential influence of scale range as contextual information reflects one source of complexity within information integration during faces evaluations. Judgements on many stimulus dimensions were shown to depend on stimulus context (Poulton 1989) and there is evidence that recognition of emotion in natural faces also depends on stimulus context (Webster et al. 2004). In this study, effects of scale range were small compared with sizeable individual differences in judgements. The contrast effect was restricted to women: they tend to perceive the same face differently, depending on other stimuli. A little smile is rather sad within bright smiles while the same smile is judged happy when most other faces are sad. Women therefore referred to contextual information when they evaluated the faces and this relativism of women's judgement is an indicator of comparably more complex information integration than that of men. Given that women seem to be more overtly expressive than men and smile more often than men (LaFrance et al. 2003), we assume that these differences in behaviour and the sex differences in judgements of faces both rely on differences in encoding and decoding of emotions. Future studies should test for this mechanism in more depth. The finding therefore supports previous studies that report more complex information processing in women than men, e.g. self-assessed health judgements are based on a wider range of health-related and non-health-related factors than are men's (Benyamini et al. 2000).

4.3 Limitations

The results need some replication in other participants than undergraduate psychology students. The finding in our sample that there were no sex-related differences in mood at the beginning of the experiment is uncommon and may limit the generalisability of the results to other populations. Furthermore, we cannot rule out that the gender-related differences in information integration which we found are typical for highly educated individuals.

In sum, study results showed that information integration during evaluation of response options was more complex in women than in men. More complex information integration in women during the whole process of judgement should often result in lower associations between judgements and objective indicators. Indeed, the stronger association of self-assessed health with mortality in men seems to depend on simpler information integration in men who rely predominantly on life-threatening diseases (Spiers et al. 2003). Consideration of predominantly life-threatening disease information in men may also contribute to the higher self-assessed health values in men which are consistently found (Pinquart and Sörensen 2001). Accordingly, the gender-related job satisfaction paradox might also be (partly) based on more complex job satisfaction judgements in women. As a practical issue for research in health-related quality of life and job satisfaction, instruments should be tested for gender equivalence. The use of the faces scale should include equivalent smiling and frowning faces, i.e. scales should be symmetrically constructed. Further research is needed to test whether sex-related differences in negative affectivity are associated with gender-related differences in information integration; this study showed that sex-related differences were not the result of differences in current mood.

References

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinic trials. *Communications in Statistics Theory and Methods, 12*, 2663–2692. doi:10.1080/03610928308828634.

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: Americans' perceptions of life quality*. New York: Plenum.

Barofsky, I. (2000). The role of cognitive equivalence in studies of health-related quality-of-life assessments. *Medical Care, 38*(9), 125–129.

Benyamini, Y., Leventhal, E. A., & Leventhal, H. (2000). Gender differences in processing information for making self-assessments of health. *Psychosomatic Medicine, 62*, 354–364.

Bi, J. (2005). Similarity testing in sensory and consumer research. *Food Quality and Preference, 16*, 139–149. doi:10.1016/j.foodqual.2004.03.003.

Chambers, C. T., Giesbrecht, K., Craig, K. D., Bennett, S. M., & Huntsman, E. (1999). A comparison of faces scales for the measurement of pediatric pain: Children's and parents' ratings. *Pain, 83*, 25–35. doi:10.1016/S0304-3959(99)00086-X.

Chen, A. W., Resurreccion, A. V. A., & Paguio, L. P. (1996). Age appropriate hedonic scales to measure food preferences of young children. *Journal of Sensory Studies, 11*, 141–163. doi:10.1111/j.1745-459X.1996.tb00038.x.

Dunham, R. B., & Herman, J. B. (1975). Development of a female faces scale for measuring job satisfaction. *The Journal of Applied Psychology, 60*, 629–631. doi:10.1037/0021-9010.60.5.629.

Ekman, P. (1972). Universal and cultural differences in facial expressions of emotion. In J. K. Cole (Ed.), *Nebraska symposium on motivation* (Vol. 19, pp. 207–283). Lincoln, NE: University of Nebraska Press.

Hollingworth, H. L. (1910). The central tendency of judgment. *Journal of Philosophy, Psychology, and Scientific Methods, 7*, 461–469. doi:10.2307/2012819.

Hox, J. J. (2002). *Multilevel analysis*. Mahwah, NJ: Erlbaum.

Izard, C. E. (1971). *The face of emotion*. New York: Appleton.

Kaplan, S. A., Warren, C. R., Barsky, A. P., Thoresen, C. J.(2008). A note on the relationship between affect(ivity) and differing conceptualizations of job satisfaction: Some unexpected meta-analytic findings. *European Journal of Work and Organizational Psychology*. Retrieved November 04, 2008, from http://www.informaworld.com/10.1080/13594320701873264.

Keren, G., & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

Kim, S. (2005). Sex differences in the job satisfaction of public employees: A study of Seoul Metropolitan Government, Korea. *Sex Roles, 52*, 667–681. doi:10.1007/s11199-005-3734-6.

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65–78. doi:10.1111/j.1744-6570.1955.tb01189.x.

Kunin, T. (1998). The construction of a new type of attitude measure. *Personnel Psychology, 51*, 823–824. doi:10.1111/j.1744-6570.1998.tb00739.x.

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: A meta-Analysis of sex differences in smiling. *Psychological Bulletin, 129*, 305–334. doi:10.1037/0033-2909.129.2.305.

Lefkowitz, J. (1994). Sex-related differences in job attitudes and dispositional variables: Now you see them. *Academy of Management Journal, 37*, 323–349. doi:10.2307/256832.

Niklas, C. D., & Dormann, C. (2005). The impact of state affect on job satisfaction. *European Journal of Work and Organizational Psychology, 14*, 367–388. doi:10.1080/13594320500348880.

Pinquart, M., & Sörensen, S. (2001). Gender differences in self-concept and psychological well-being in old age: A meta-analysis. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 56*, 195–213.

Poulton, E. C. (1989). *Bias in quantifying judgements*. Hillsdale, NJ: Erlbaum.

Sarris, V. (2004). Frame of reference models in psychophysics. In C. Kaernbach, E. Schröger, & H. Müller (Eds.), *Psychophysics beyond sensation: Laws and invariants of human cognition* (pp. 69–88). Mahwah, NJ: Erlbaum.

Spiers, N., Jagger, C., Clarke, M., & Arthur, A. (2003). Are gender differences in the relationship between self-rated health and mortality enduring? Results from three birth cohorts in Melton Mowbray, United Kingdom. *The Gerontologist, 43*, 406–411.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1991). Overall job satisfaction: How good are single-item measures? *The Journal of Applied Psychology, 82*, 247–252. doi:10.1037/0021-9010.82.2.247.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557–561. doi:10.1038/nature02420.

Wedell, D. H. (1990). Methods for determining the locus of context effects in judgment. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases: Advances in psychology* (pp. 285–302). Amsterdam: North-Holland.

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and in the United States. *Journal of Experimental Social Psychology, 43*, 303–311. doi:10.1016/j.jesp.2006.02.004.