BEHAVIOR RESEARCH METHODS

Revisiting Rossion & Pourtois with new ratings for automated complexity, familiarity, beauty and encounter.

Journal:	*Behavior Research Methods*
Manuscript ID	BR-Org-15-335.R2
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Forsythe, Alex; University of Liverpool, Psychology Street, Nichola; Staffordshire University Helmy, Mai; Menoufia University, Psychology

Revisiting Rossion & Pourtois with new ratings for automated complexity, familiarity, beauty and encounter.

Differences between norm ratings collected when participants are asked to consider more than one picture characteristic (Rossion and Pourtois, 2004) are contrasted with the traditional methodological approaches of collecting ratings separately for image constructs (Proctor & Vu, 1990). We present data that suggests that reporting normative data, based on methodological procedures that ask participants to consider multiple image constructs simultaneously, could potentially confounded norm data. We provide data for two new image constructs, beauty and the extent to which participants encountered the stimuli in their everyday lives. Analysis of this data suggests that familiarity and encounter are tapping different image constructs. The extent to which an observer encounters an object predicts human judgements of visual complexity. Encountering an image was also found to be an important predictor of beauty, but familiarity with that image was not. Taken together, these results suggest that continuing to collect complexity measures from human judgements is a pointless exercise. Automated measures are more reliable and valid measures, which are demonstrated here as predicting human preferences.

Subjective ratings have been an established method by which to produce normative data for language and picture research (see Proctor & Vu 1990, for a review). Paivio and his colleagues were one of the first to obtain normative ratings of concreteness, imagery and meaningfulness in what was to become one of the best-known sets of normative ratings for the imageability, concreteness and meaningfulness of words (Paivio, Yuille & Madigan, 1968). Their motivation for obtaining ratings was the lack of appropriate normative data for word characteristics that they wished to investigate in the course of their research. Prior research had sometimes relied on '*unspecified judgements by the experimenter alone*' (p2). Since Proctor & Vu, further norms have been reported for not only words but also icons and symbols (McDougall et al., 2000; 1999; Forsythe et al., 2003; 2008), most extensively for picture sets (e.g. Alario et al., 2004; Barry, Morrison, & Ellis, 1997; Bates et al., 2003; Bogka et al., 2003; Bonin, Barry, Méot, & Chalard, 2004; Bonin, Chalard, Méot, & Fayol, 2002; Catling & Johnston, 2006; Cuetos, Ellis, & Alvarez, 1999; Dell'Acqua, Lotto, & Job, 2000; Dimitropoulou et al., 2009; Ellis & Morrison, 1998; Lloyd-Jones & Nettlemill, 2007; Morrison & Gibbons, 2006; Morrison, Hirsh, & Duggan, 2003; Snodgrass & Yuditsky, 1996; Snodgrass & Vanderwart, 1980; Vitkovitch & Tyrrell, 1995; Weekes, Hao, Shu, Liu, & Tan, 2007; Zevin & Seidenberg, 2002; Rossion & Pourtois, 2004) and recently for art (Forsythe et al., 2011).

Following on from the initial classic work by Snodgrass & Vanderwart (1980), Rossion & Pourtois were interested in examining visual complexity. Snodgrass & Vanderwart suggested how, in episodic memory tasks, complexity is likely to influence stimulus recognition. The extra detail depicted in an object may give an image added novelty, and this novelty may slow the recognition process. The authors felt it likely that increased complexity would influence the speed at which pictures are categorised, man-made objects being simpler would be categorized most quickly, and naturalistic complex images, such as insects or trees would be categorised more slowly. Some categorical reaction time advantage has been reported - natural categories tend to be responded to more quickly than other natural categories—although this seemed to be mainly a function of diagnostic colour for example such as fruits/vegetables versus animals, rather than a function of complexity (Rossion and Pourtois, 2004).

Other researchers have reported this variability in complexity effects. Some have suggested that increased complexity can enhance performance (Biederman, 1987; Lloyd-Jones & Luckhurst, 2002), other have argued that visual complexity increases processing

time (and hence naming time) at, or before, the stage of object recognition (Alario et al., 2004; Ellis & Morrison, 1998; Humphreys, Riddoch, & Quinlan, 1988). One reason for variations in complexity effects is possibly explained by the way in which researchers have attempted to quantify what is complex. The metrics used to determine complexity within images differed between researchers and in some cases complexity was confounded with other variables such as concreteness (McDougall et al., 1999; and see Forsythe et al., 2008). As such researchers have sought to find ways to standardise the measurement of complexity (Forsythe et al., 2008) or, as previously mentioned, to develop sets of standardised images for use in testing.

Measuring Complexity

The study of visual complexity emerged from the empiricist tradition. The tradition is based on the premise that people make poor intuitive judges and understanding could only be advanced through quantification in controlled laboratory settings. When unusual, unexplainable results emerged, Gestalt psychology developed to explain them. The Gestaltists set out to understand the processes of perception, not through the meticulous analysis of patches of light, shape and colour, but through an analysis of the whole, configuration or form (Hochberg, 1986). Their philosophy was that sensations are not elementary experiences; we "see" shape and form regardless of where the image falls on the retina or what neurons process the various image components. What was important was constancy. One such law generated through the Gestalt movement was Prägnanz. The Prägnanz principle contends that the forms that are actually experienced take on the most parsimonious or 'best' arrangement possible in given circumstances. In other words, of all the possible perceptual experiences to which a particular stimulus could give rise, the one most closely fitting to the concept of 'good' will be experienced.

Kofka (1935) proposed that the term 'good' means symmetrical, simple, organised and regular. In his study of psychological organisation Kofka explained the tendency to create psychologically, simple order patterns from a wide range of perceptual stimuli. This early study of 'simplicity' evolved into the study of 'complexity', with theorists attempting to re-write the Gestalt Law of simplicity within a more formal framework (Attneave, 1954; Attneave & Arnoult, 1956; Hochberg & Brooks, 1960). Both Hochberg and Attneave acknowledged that shape was a multidimensional variable that would vary with the complexity of an image, with Hochberg and Brooks going on to developed what was the first

semi-automated measure of image complexity, arguing that relying solely on human judgments of complexity would mean that they had no way of predicting just how complex or simple an image would appear.

Later Approaches to Visual Complexity

Following the work of Attneave and Arnold, complexity has received less attention, in part because no universally acceptable metric existed. Those measures that had been developed historically were not particularly well supported within a theoretical framework (Johnson, et al., 1996). For example, Geiselman et al., (1982) developed an index of discriminability between graphic symbols and identified nine 'primitive' attributes; e.g. numbers of straight lines, arcs, quasi angles and blackened-in elements. Symbols selected for high discriminability using this metric were responded to faster than those with lower discriminability. Garcia et al (1984) also sought to count the number of primitive attributes in icons and signs in order to determine how concrete, or pictorial, the icon was. Unfortunately for the authors, this proved to be a much better measure of visual complexity than concreteness (see McDougall et al, 1999). Garcia et al. reported that icons that are pictorially similar to their real world counterparts are more likely to be judged as complex. This has been found not to be the case, complexity is more closely related to search efficacy (McDougall et al., 2000). A more valid and reliable measure of complexity would enable researchers to determine more accurately the effects of extra detail and intricacy on performance.

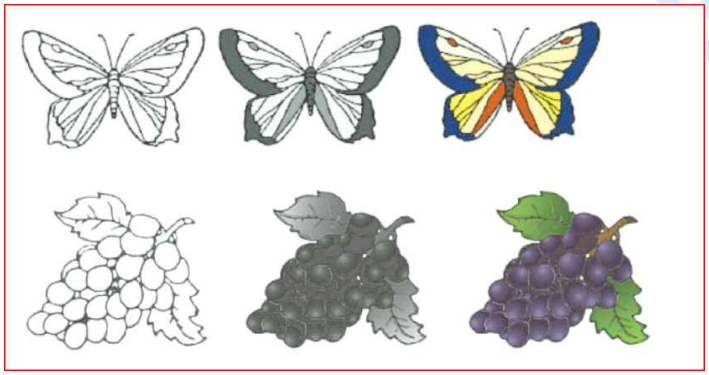
Forsythe et al., (2003; 2008) tested several automated measures of complexity based on measurements of the changes in image intensity (Beck et al., 1991; Harwert et al., 1978; Sutter et al., 1989; Vassilev & Mitov, 1976). More recent work (Forsythe et al., 2011), examined the relationship between information processing models suggested by Shannon & Weaver (1949) and image compression as a measure of visual complexity. When images contain few elements or are homogenous in design, there are few message alternatives and as such the file string contains mostly numbers to be repeated. A more complex picture will have a less predictable number string. These measures seem to have good reliability when compared with human judgements of visual complexity (Forsythe et al., 2008; 2011) and have contributed towards a general disposition towards the development complexity metrics in the field (Marin & Leder, 2013; Machado, Romero, Nadal, Santos, Correia, & Carballal, 2015).

Familiarity and Complexity

The idea that observers make poor judges of visual complexity is important when considering the conventions of data collection. Rossion and Pourtois (2004) collected ratings on a number of stimulus variables (familiarity, concreteness, complexity etc.,) for new versions of pictures in the style of Snodgrass and Vanderwart (1980). Previously only line drawing versions of these images had existed and Rossion & Pourtois created versions in outline, colour and Greyscale (Figure 1).

Historically, groups of raters had been employed in the collection of picture norms, with each group being asked only to consider one image construct (see Protor and Vu, 1990 for a review). Rossion & Pourtois did not follow this tradition; rather groups of twenty subjects performed both the complexity *and* familiarity tasks. This could be problematic as Forsythe et al., provide experimental evidence that when observers are made more familiar with objects with no semantic content (i.e. nonsense shapes) they begin to rate those objects as less complex than they actually are, suggesting that the complexity data collected from Rossion and Pourtois could be confounded with familiarity judgements. This confound would somewhat explain the large correlations between complexity and familiarity judgements reported in the Rossion & Pourtois data set; correlations which are atypical in most other picture data sets

Figure 1: Examples of the Rossion and Pourtois images



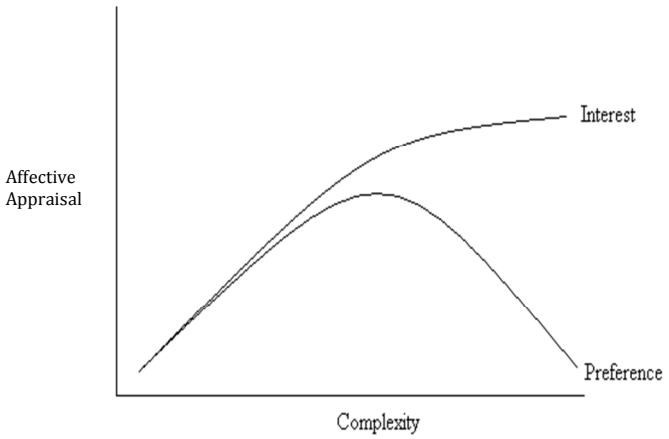
Of course, if familiarity is a part of the construct of complexity, then this is what researchers and designers may need to take into consideration rather than simply arriving at the best, context free measure of visual complexity. The collection of ratings that contain both a complexity and familiarity component is not necessarily inherently bad and

removing familiarity effects is not an advantage in its own right, it depends on what one wants. With this caveat in mind, we present a contrast analysis with the Rossion and Pourtois data set for the variables familiarity and complexity, with new ratings collected by separate groups of observers. We also provide for researchers automated data for the variable visual complexity, which is based on the Gif metric reported by Forsythe et al., (2011).

Beauty and Complexity

In the study of beauty Berlyne's (1970) curvilinear relationship with visual complexity has received the most attention. Berlyne argued that complexity increases linearly with preference until an optimum level of visual arousal is reached (Figure 2). At this point further increases in complexity would elicit a down turn in arousal and preference would decrease. In other words, when visual stimuli are of low complexity (i.e. simple), preference and judgements of beauty will also be low. People will seek to maintain a level of arousal that supports their preferred level of stimulation. Individuals who are highly aroused will seek out certainty, whereas those low on arousal will seek more stimulating environments. Berlyne's theory has received mixed support because it has poor predictive validity; it is not possible to determine the point of the cusp. There is also some suggestion that when familiarity for an image is controlled for, the relationship between beauty and visual complexity is much more linear in nature (Forsythe et al., 2011).

Figure 2. Berlyne (1971), the effect of complexity on preference and interest.



Exposure and beauty

The competing tensions between beauty and visual complexity perhaps generate some degree of arousal between the existing and the unexpected. We also know that repeated exposure is sufficient to enhance positive attitudes (Zajonic, 1968) perhaps because people are uncertain about how to deal with objects that are novel, repeated exposure acts to make the stimulus more accessible to the individual. Increases in preference are possible with even the slightest repeated exposure, hence the term mere exposure effect (meaning that the object is just accessible to perception). As with the beauty and visual complexity, mere exposure is subject to an inverted U shaped relationship. Whilst preference increases initially with exposure to a stimulus, later repeated exposure elicits a decrease in preference (Moreland & Topolinski, 2010). Bornstein (1989) attributes this effect to "attributional discounting". Liking increases with repeated exposure, but observers attribute some of this liking with the exposure processes: *I have seen it often therefore I like it because it is familiar to me.* If however, observers are not aware of repeated exposure to a stimulus, the discounting effect does not occur. When exposure frequencies and familiarity ratings were used to predict preference, each variable contributed differently to preference judgements and suggesting that the mere exposure effect could take place without learning occurring (Moreland & Zajonc, 1977; Moreland & Zajonic, 1979), suggesting that exposure contributes to preference regardless to how familiar an object is.

Although psychologists have known for some time that preference and exposure are related (Fechner 1876) the mere exposure phenomena is still of significant interest to the field. Contemporary research is linking exposure to perceptual fluency or the speed to which a stimulus is processed and greater perceptual fluency generates a positive affect (see Moreland & Topolinski, 2010 for a review). With this in mind we offer new ratings for the Rossion & Pourtois picture sets (line drawings, grey scale and colourised images) for beauty and exposure to these pictures. Exposure is operationalized by self-report measures of how often participants 'encounter' the images. Such measures have not been collected to date and the work of Moreland, Zajonc suggest that that encountering something on a regular or irregular basis may not necessarily be the same in concept as being familiar with an image and that familiarity and exposure may contribute to judgements of beauty in different ways. We may be familiar with what a butterfly is, but we do not necessarily encounter the insect every day. Encounter is a variable, which perhaps captures exposure effects and could be potentially useful to researchers when attempting to measure preferences for pictures.

Method

Participants

11 different groups of 30 participants from three United Kingdom University student populations (n=330) took part in this experiment. For 10 of the groups participants rated only one image construct for one image type (colourised, grayscale or line drawing). Rossion and Pourtois (2004) collected their data set from French speaking students. It could be argued that any differences in data sets could be due to cultural factors. To determine if this were the case, as a control, ratings were collected from one group (n=30) for both complexity and familiarity simultaneously, with the aim of determining if any cultural differences existed between the French and UK population.

Stimuli

The Rossion and Pourtois (2004) image sets for colourised, grey scale and line drawings were presented in a PowerPoint presentation on a screen resolution of 1024 x 768. In the Rossion and Pourtois (R&P) methodology, each stimulus was preceded by an attention signal (!) for 500 ms and, after a brief blank screen (150 ms), was presented for 3000 ms, however because participants rated for two constructs they were exposed for

6000 ms in total. For complete comparability of data, our participant's viewed each image for 6000 ms.

Procedure

Participants were asked to rate pictures on a Likert scale from 1 – 5. Ratings were collected for the variables of complexity, familiarity and encounter for colourised, grey scale and line drawings. A score of 1 represented a picture that was not at all familiar; a score of 5 was an image that was very familiar. For encounter, participants were asked to consider how often they encountered the items in the pictures. A score of 1 was not very often and a score of 5 very often. Complexity was described as the 'amount of detail or intricacy' (Snodgrass & Vanderwart, 1980) in the image with a score of 1 being very simple and a score of 5 being very complex.

An additional set of ratings was obtained for beauty for the line drawing picture set. Participants were asked to consider on a 5-point scale the extent to which something was considered to be beautiful [not at all or very much]. Norms were not obtained for the colourised or grey scale sets as it was considered that colour and shading could act as mediating factors in judgements of beauty.

Automated measures

The R&P picture sets (n=260) were analysed using the two most reliable compression measures as possible automated measures of visual complexity (Forsythe et al., 2011). Jpeg (lossy compression) is a technique that reduces the size of the image file by removing redundant information, but generally assumes that some loss of information is acceptable, this means that Jpeg compression does not always reconstruct an image to its original format and is susceptible to the inclusion of compression artefacts also known as pixilation. Gif (lossless compression) works on a similar principle except that when the image is to be recovered no image loss occurs, for this reason it works well in compressing images that have sharp transitions such as diagrams, text or line drawings. However, because Gif retains more of the integrity of the image Gif can only compress to 50% of the image size. Here images were compressed both in GIF and JPEG to a 50% compression size.

Results

Analysis 1: Rossion and Pourtsis contrasted with image constructs collected in isolation

Table 1: Summary statistics for the three image sets.

(n=260)	Complexity		Familiarity		Encounter	Gif	JPEG	Beauty
	Forsythe	R&P	Forsythe	R&P				
Line								
Mean	2.66	2.77	4.17	3.59	3.33	3140.84	2169.18	3.15
StDev	.85	1.03	.55	1.01	1.13	943.24	420.98	.63
Skew	.15	.11	-.48	-.32	-.28	.76	.52	.86
Skew Error	.15	.15	.15	.15	.15	.15	.15	.15
Kurtosis	-.74	-1.02	-.70	-1.04	-1.24	.51	.21	-.67
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	.30
Minimum	1.04	1.00	2.72	1.06	1.07	999	1698	1.74
Maximum	4.60	4.82	5.00	5.00	4.97	7285	33056	4.66
Grey Scale								
Mean	2.65	2.89	2.76	3.52	3.17	4738.21	1992.17	
StDev	.67	1.03	1.50	.94	.57	1408.04	352.89	
Skew	-.09	-.02	.22	-.31	.69	.22	.74	
Skew Error	.15	.15	.15	.15	.15	.15	.15	
Kurtosis	-.68	-1.12	-1.39	-1.04	.35	-.43	.87	
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	
Minimum	1.09	1.06	1.83	1.41	2.07	1698	1326	
Maximum	4.17	4.88	4.96	5.00	4.89	33056	7193	
Colour								
Mean	2.61	2.01	3.59	3.43	2.79	4781.22	2083.27	
StDev	.77	.94	.83	1.01	1.18	1412.77	359.51	
Skew	.14	.21	-.23	-.15	.28	.23	.80	
Skew Error	.15	.15	.15	.15	.15	.15	.15	
Kurtosis	-.73	-1.11	-1.10	-1.31	-1.26	-.46	1.16	
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	
Minimum	1.04	1.00	1.73	1.53	1.11	1811	1433	
Maximum	4.46	4.65	5.42	5.00	5.00	29859	8389	

Table 1 shows the means, standard deviations, kurtosis, skew and errors for each of the automated and human counts of complexity, familiarity and encounter. Subjective judgements of familiarity show evidence of skew and on histogram inspection it is apparent that for the Rossion and Pourtois picture set participants perceive a large number of very familiar images. For the line drawing set, no images received mean ratings below the mid-point of 3, for Grey scale and colour ratings started at point 2.

For visual complexity, no statistically cross-cultural differences were identified between the UK group and the French group of participants for ratings of complexity and familiarity collected simultaneously. For familiarity, there were no statistically significant differences between the R&P data set and UK data collected simultaneously, suggesting again that any differences in collected norms will not be due to cross cultural differences.

The differences between the R&P data set and the new norm data reported here were examined using analysis of variance (GLM). There are significant differences between the R&P ratings and new ratings reported here, with large effect sizes across the familiarity and complexity image categories for Line drawings, $F(1,518)=486.18$, $p<.01$, η_p^2 .47, colourised drawings $F(1,518)=157.26$, $p<.01$, η^2.23 and grey scale, $F(1,518)$ 160.96, $p<.01$, η^2.24 In addition to the main effect, significant interactions were found between the picture sets (Forsythe et al., & R&P) and the norm scores (Figures 3&4) for Line Drawings, $F(1,518)=74.89$, $p<.01$, η_p^2 .13 and for Grey scale drawings $F(1,518)$ 9.20, $p<.01$. η_p^2 .02. When compared to the R&P data sets, mean familiarity ratings across data sets for line drawing and grey scale are higher, where as the mean complexity ratings are lower. This difference seems to be less pronounced for colourised drawings.

Figure 3: Line drawings mean responses across groups

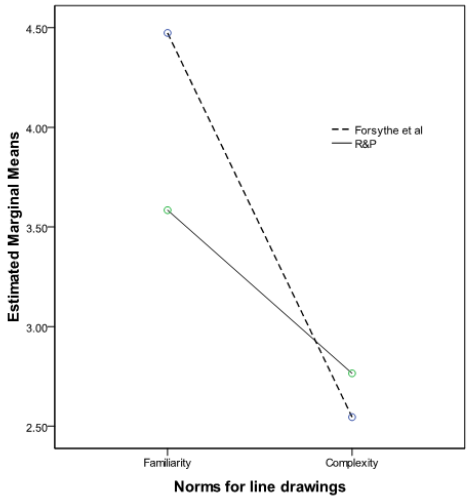
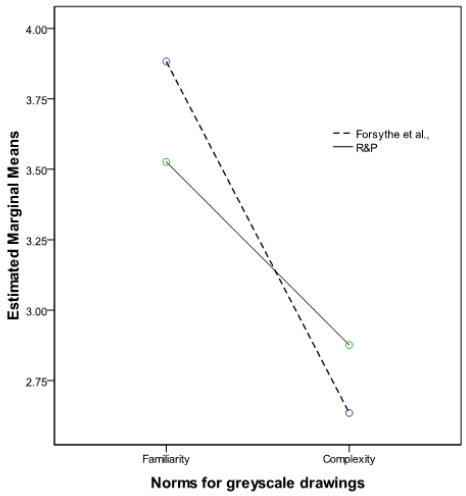


Figure 4: Grey scale mean responses across groups



Analysis 2: Correlations between image constructs and automated complexity.

Before the following correlations were calculated 6 outliers were removed on visual inspection of stem and leaf plots and the outlier labelling rule from across the three image sets. These outliers related to images that had unusually large compression scores relative to the remainder of the data set. Table 2 details the correlations between the different variables. For the Line drawing set, the correlation between Gif compression and human judgements of complexity is r_s.78, p<.01, and for Gif compression r_s.67, p<.00. These small differences are explained by the lossless technique favoured by Gif compression, a method that works best with images that have sharp transitions. Jpeg is known to perform better on images that have high colourisation and this finding is supported by the larger correlate

(r_s=.61, p<.01). For the Grey scale set there seems to be little difference between the two correlations. Table 2 also demonstrates that measuring visual complexity with compression techniques produces scores that do not correlate strongly with judgements of familiarity.

Table 2: Significant Spearman correlations.

Line	**Complexity**	**Familiarity**	**Encounter**
Familiarity	-.46	1.00	
Encounter	-.48	-.87	1.00
Gif	.78	-.29	-.29
JPeg	.67	-.25	-.25
Grey Scale			
Familiarity	-.42	1.00	
Encounter	.32	-.79	1.00
Gif	.55	-.15	$_{ns}$ -.04
JPeg	.58	-.24	$_{ns}$ .00
Colour			
Familiarity	-.44	1.00	
Encounter	.47	-.89	1.00
Gif	.54	$_{ns}$-.17	$_{ns}$.14
JPeg	.61	-.24	.21

Analysis 3: Beauty, exposure and visual complexity

When automated measures for visual complexity are applied there seems to be limited evidence for an inverted U-shaped relationship between complexity and beauty (Figs 5&6). The computerised measures suggest that there is a much sharper climb in preference for images that are above the mean point of visual complexity.

Fig 5 Human judgments of beauty, contrasted with computerised measures of complexity

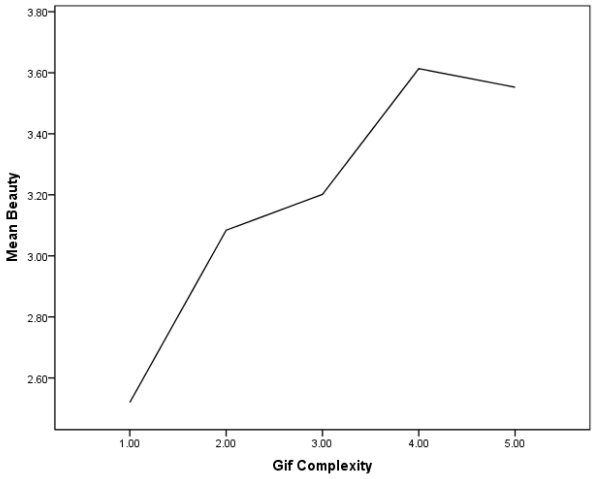
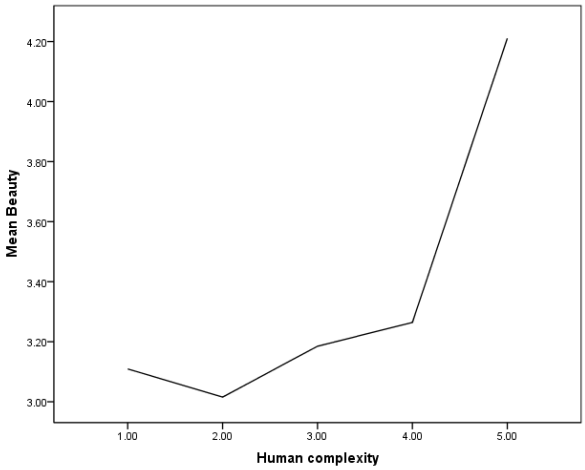


Fig 6 Human judgments of beauty, contrasted with human judgements of complexity



Predictors of Visual Complexity and Beauty in Human judgements

The variables Gif, beauty, familiarity and encounter were regressed onto the dependent variable human judgements of visual complexity. High correlations between the variables familiarity and encounter contributed to high co-linearity within the model. The model was revised with encounter and familiarity entered separately. The overall model accounted for a moderate percentage of the variance in human judgements of visual complexity (r_2.62) F(3,256)138.15, p<.01 with Gif emerging as the largest individual predictor variable (β.66, t=16.25, p<.01), followed by encounter (β-.31, t=-7.52, p<.01). For the revised model, which included familiarity rather than encounter, the overall model variance remained the same. Similar to the encounter variable, familiarity makes a negative individual contribution (β-.29, t=-7.11, p<.01). Scatterplot examination confirmed that images that are less familiar or encountered less frequently are also rated as more complex, findings that complement results reported elsewhere (Forsythe et al., 2008). Neither beauty nor familiarity predicted visual complexity.

The individual contributions towards the dependent variable 'beauty' were smaller (r_2.27 (F(3,256) 6.46, p<.01). The significant predictors were Gif complexity β.23, t=2.60, p<.01 and the extent to which viewers encountered the images (β-.25, t=3.64, p<.01). Again, scatterplot examination determined that images that are encountered less often are perceived as less beautiful. Previous research has suggested that complexity is an important factor in beauty. The data reported here suggest Gif complexity contributed in a small way to perceptions of beauty, but that beauty has no significant relationship with human judgements of visual complexity or familiarity with an image.

Discussion

Rossion & Pourtois (2004) were interested in measuring complexity because of its impact on processing speed. By developing three new sets of images in greyscale, colour and as line drawings they were able to collect normative data (naming agreement, familiarity, complexity and imagery judgements) for images created similar to those of Snodgrass and Vanderwart (1980). Through the provision of two new object data sets, the authors made an important contribution towards studies of object recognition in normal and clinical populations. These ratings were consequentially used to examine reaction and naming times for these pictures. The authors reported some categorical reaction time advantage—that is, some categories tend to be responded to more quickly than others—

although this seemed to be mainly a function of diagnostic colour in categories, such as fruits/vegetables versus animals. However, because the authors overlooked convention and collected ratings for familiarity and complexity from the same group of participants some of the reported norms may be confounded.

The aim of the study reported here was to examine the extent to which collecting data from different groups of participants would alter the Rossion & Pourtois data for familiarity and complexity. Our results suggest that when ratings are collected from different groups of observers, scores differ from those reported by Rossion & Pourtois. New norms reported here are systematically rated as more familiar and less complex than previously recorded, with large correlations between complexity norms collected in isolation and compression measures of complexity (Table 2).

If a theory of the perception of complexity as mediated by top-down processing is correct, longer exposure, combined with the request for judgements of familiarity should lead to pictures being judged as *less* complex and *more* familiar than ratings gathered for complexity alone. Here that does not seem to be the case. Complexity ratings collected separate from familiarity presented mean scores that are lower than norms reported by Rossion and Pourtois. Familiarity ratings collected separately from complexity present a higher norm average than Rossion and Pourtois.

For judgements of familiarity, the data reported here has a larger minimum score, for example 2.72 for line drawings, compared with Rossion & Pourtois (1.06). Such a large minimum score suggests that for the line drawing set, in particular, observers did not feel that there were many unfamiliar pictures in the set and such observations were not found to reflect a cross-cultural effect. Asking observers to consider scoring an image on two variables increases exposure time and the observer would be able to make a thought-out response in regard to how familiar they are with the object in question and how much detail they could see in the object. Ratings for pictures that have been gathered for visual complexity and familiarity could then lead to judgements that are more complex than ratings gathered in isolation. In the original Rossion & Pourtois study each image had an exposure time of 3000ms. If participants were to view the image twice then exposure time would increase to 6000ms. Time then could have facilitated greater consideration of detail and complexity; however, in the study reported here all images were presented for 6000ms, suggesting that his explanation could not hold. A more likely explanation is that perhaps observers became confused. Having rated 260 images for one image construct very quickly

(3000ms), Rossion & Pourtois required participants to repeat the activity again for a different image construct. Fatigue and interference could have influenced the results with observers inadvertently rating the images for the wrong image construct or simply becoming bored with the activity. This would explain the unusual anomalies in the distributions of scores between the new data reported her and the R&P image sets, particularly for ratings of visual complexity.

Beauty, exposure and visual complexity

The second aim of this study was to further examine the relationship between beauty, exposure to an image and visual complexity. Data reported here suggests a more linear relationship between judgements of beauty and complexity, a relationship that is somewhat different from the predictions of Berlyne who argued for an inverted-U shaped relationship. A much sharper climb in preference for images that are above the mean point of visual complexity is evident (Figs 5 & 6).

Whilst the high correlations between the variables encounter and familiarity would suggest they are to some degree tapping the same image constructs, regression analysis presented the encounter variable as explaining more of the variance both in judgements of visual complexity and judgments of beauty. Overall our model explains 62% of the variance in human judgements of visual complexity, with GIF compression emerging as the largest predictor variable and 'encounter' emerging as a marginally stronger predictor variable than familiarity. It would seem then that the number of times in which we encounter an object is a good predictor of human judgements of visual complexity, with images encountered less often being considered more complex. The difference between the two constructs builds on previous research (Forsythe et al., 2008) which suggests that repeated exposure to shapes with meaningless content (i.e. no semantic information), and therefore completely unfamiliar, can reduce perceived visual complexity.

Previous research has suggested that complexity is an important factor in beauty but our model only predicted a moderate amount of the variance; with compression complexity and the extent to which participants encountered the image emerging as significant individual contributors. Familiarity with the image did not predict beauty, nor did human judgements of complexity. Taken together, these analyses suggest that continuing to collect complexity judgements based on human ratings is a pointless exercise and that researchers should consider further analysis of the extent to which participant's

are exposed to, meet with or encounter an image, rather than simply how familiar the subject mater is.

Conclusions

The relationship between complexity and familiarity resurrects an old argument that complexity is meaningless; it is the way in which a stimulus is perceived that is important, not the number of elements (Rump, 1968). If complexity correlates negatively with familiarity is it intrinsically bad? If familiarity is a part of the construct of complexity, then this is what researchers and interface designers may need to take into consideration rather than simply arriving at the best, context free measure of visual complexity. Compression techniques offer researchers the most reliable and user-friendly option for the quantification of visual complexity, they are also unbiased - they are not affected by familiarity with an image set. These metrics have a strong theoretical basis (information theory), produce good approximations of human judgments, and have been demonstrated her as being able to predict human behaviour. However, it is a reality that visual complexity is related to familiarity and researchers should consider what it is that they want from a measure of visual complexity and if removing familiarity from the equation is warranted. With this in mind, the reported statistics by Rossion & Pourtois are not intrinsically incorrect; they are simply an alternative way by which to measure picture constructs.

Our findings also determine that perhaps considering the extent to which a person actually encounters an object on a day-to-day basis may be a useful image construct. Encounter seems to explain some of the variance in how people reach judgments of image complexity and in how aesthetically pleasing one finds an object.

References

Alario, F.-X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers,* 31, 531-552.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review,* 61, 183–193

Attneave, F., Arnoult, M.D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin*, 53, 452–471

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology,* 50A, 560-585.

Bogka, N., Masterson, J., Druks, J., Fragkioudaki, M., Chatziprokopiou, E.-S., Economou, K. (2003). Object and action picture naming in English and Greek. *European Journal of Cognitive Psychology*, 15, 371-403.

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, and Computers, 31,* 650-658

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonava, E., De- vescovi, A., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review,* 10, 344-380.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguisitic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers,* 35, 158-167

Beck, H., Graham, N., Sutter, A. (1991) Lightness differences and the perceived segregation ofregions and population. *Perception and Psychophysics.* 49(3), 257–269.

Dell'Acqua, R., Lotto, L., & Job, R. (2000). Naming times and stan- dardized norms for the Italian PD/DPSS set of 266 pictures: Direct comparisons with American, English, French, and Spanish published databases. *Behavior Research Methods, Instruments, & Computers,* 32, 588-615.

Dimitropoulou, M., Panagiotis Blitsas, J.A., and Carreiras, M. (2009). A standardized set of 260 pictures for Modern Greek: Norms for name agreement, age of acquisition, and visual complexity, *Behavior Research Methods,* 41, 2, 584-589.

Garcia, M., Badre, A.N., Stasko, J.T. (1994). Development and validation of icons varying in their abstractness. *Interacting with Computers,* 6, 2, 191–211

Hochberg, J.E., Brooks, V. (1960). The psychophysics of form: Reversible perspective drawings of spatial objects. *American Journal of Psychology,* 73, 337–354

Hochberg, J.E. (1986). Perception, 2nd edn. Prentice-Hall, Englewood Cliffs.

Fechner, G.T. (1876). *Vorschule der Asthetik*, Leipzig, Breitkopf und Hartel

Forsythe, A., Nada, M., Sheehy, N., & Cela-Conde, J. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology, 102,* 49–70.

Forsythe, A., Mulhern, G., Sawey, M. (2008). Confounds in pictorial sets: the role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods,* 40, 1, 116–129.

Forsythe, A., Sheehy, N., Sawey, M. (2003). Measuring icon complexity: an automated analysis. *Behavior Research Methods, Instruments, and Computers,* 35, 334–342

Geiselman, R.E., Landee, B.M., Christen, F.G. (1982) Perceptual discriminability as a basis for selecting graphic symbols. *Human Factors* 24, 329–337

Johnson, C.J., Paivio, A., Clark, J.A. (1996) Cognitive components of picture naming. *Psychological Bulletin,* 120(1), 113–139

Lloyd-Jones, T.J. & Nettlemill, M. (2007). Sources of error in picture naming under time pressure. *Memory & Cognition*. 35, 816-836.

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., & Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, 16043-57.

Marin, M. M., & Leder, H. (2016). Effects of presentation duration on measures of complexity in affective environmental scenes and representational paintings. *Acta Psychologica*, *163*, 38-58.

McDougall, S.J.P., Bruijn de, O., Curry, M.B (2000) Exploring the affects of picture characteristics on user performance: The role of picture concreteness, complexity and distinctiveness. Journal of Experimental Psychology: Applied 6, 291–306

McDougall, S.J.P., Bruijn, D.O., Curry, M.B. (1999). Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity and semantic distance for 239 symbols. *Behavior Research Methods,* 31, 3, 487–519.

Morrison, C.M. Gibbons Z. (2006) Review of lexical determinants of semantic processing speed, *Visual Cognition,* 13, pp. 949–967

Morrison, C. M., Hirsh, K. W., & Duggan, G. B. (2003). Age of acquisition, ageing and verb production. Normative and experimental data. *Quarterly Journal of Experimental Psychology*, 56A, 705-73

Proctor, R. W., & Vu, K.-P. L. (1999). Index of norms and ratings, *Behavior Research Methods, Instruments, & Computers*, 31, 659-667.

Rossion, B., Pourtois, G. (2004) Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception*, 33, 217–236

Rump, E.E. (1968). Is there a general factor of preference for complexity? *Perception& Psychophysics,* 3, 346–348

Shannon, C.E., Weaver, W. (1949) The mathematical theory of communication. University of Illinois Press, Urbana.

Snodgrass, J.G., Vanderwart, M. (1980) A standardized set of 260 pictures. Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental psychology*. *Human Learning & Memory*, 6, 174–215.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snod- grass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers,* 28, 516-536.

Sutter, A., Beck, J., Graham, N. (1989) Contrast and spatial variables in texture segregation: Testing a simple spatial-frequency channels model. *Perception and Psychophysics*, 46(4), 312–332.

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology,* 48A, 822-848.

Vassilev, A., Mitov, D. (1976) Perceptual time and spatial frequency. *Vision Research* 16, 89–92

Weekes, B. S., Shu, H., Hao, M., Liu, Y., & Tan, L. H. (2007). Predictors of timed picture naming in Chinese. *Behavior Research Methods*, 39, 335-342.

Zevin, J.D. & Seidenberg, M. S. (2002) Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47, 1-29

(n=260)	Complexity		Familiarity		Encounter	Gif	JPEG	Beauty
	Forsythe	R&P	Forsythe	R&P				
Line								
Mean	2.66	2.77	4.17	3.59	3.33	3140.84	2169.18	3.15
StDev	.85	1.03	.55	1.01	1.13	943.24	420.98	.63
Skew	.15	.11	-.48	-.32	-.28	.76	.52	.86
Skew Error	.15	.15	.15	.15	.15	.15	.15	.15
Kurtosis	-.74	-1.02	-.70	-1.04	-1.24	.51	.21	-.67
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	.30
Minimum	1.04	1.00	2.72	1.06	1.07	999	1698	1.74
Maximum	4.60	4.82	5.00	5.00	4.97	7285	33056	4.66
Grey Scale								
Mean	2.65	2.89	2.76	3.52	3.17	4738.21	1992.17	
StDev	.67	1.03	1.50	.94	.57	1408.04	352.89	
Skew	-.09	-.02	.22	-.31	.69	.22	.74	
Skew Error	.15	.15	.15	.15	.15	.15	.15	
Kurtosis	-.68	-1.12	-1.39	-1.04	.35	-.43	.87	
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	
Minimum	1.09	1.06	1.83	1.41	2.07	1698	1326	
Maximum	4.17	4.88	4.96	5.00	4.89	33056	7193	
Colour								
Mean	2.61	2.01	3.59	3.43	2.79	4781.22	2083.27	
StDev	.77	.94	.83	1.01	1.18	1412.77	359.51	
Skew	.14	.21	-.23	-.15	.28	.23	.80	
Skew Error	.15	.15	.15	.15	.15	.15	.15	
Kurtosis	-.73	-1.11	-1.10	-1.31	-1.26	-.46	1.16	
Kurtosis Error	.30	.30	.30	.30	.30	.30	.30	
Minimum	1.04	1.00	1.73	1.53	1.11	1811	1433	
Maximum	4.46	4.65	5.42	5.00	5.00	29859	8389	

Table 2: Significant Spearman correlations.

Line	Complexity	Familiarity	Encounter
Familiarity	-.46	1.00	
Encounter	-.48	-.87	1.00
Gif	.78	-.29	-.29
JPeg	.67	-.25	-.25
Grey Scale			
Familiarity	-.42	1.00	
Encounter	.32	-.79	1.00
Gif	.55	-.15	$_{ns}$ -.04
JPeg	.58	-.24	$_{ns}$ .00
Colour			
Familiarity	-.44	1.00	
Encounter	.47	-.89	1.00
Gif	.54	$_{ns}$-.17	$_{ns}$.14
JPeg	.61	-.24	.21

Figure 1: Examples of the Rossion and Pourtois images

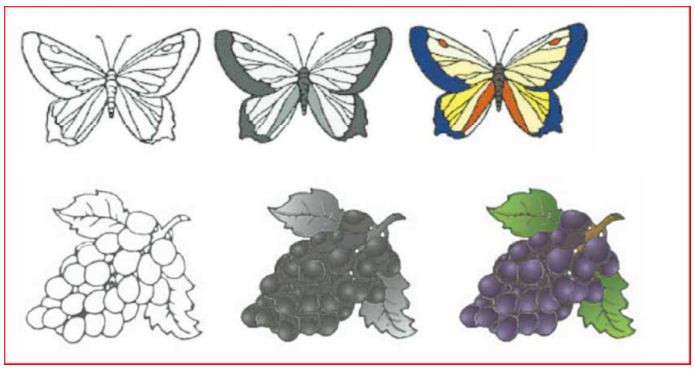


Figure 2. Berlyne (1971), the effect of complexity on preference and interest.

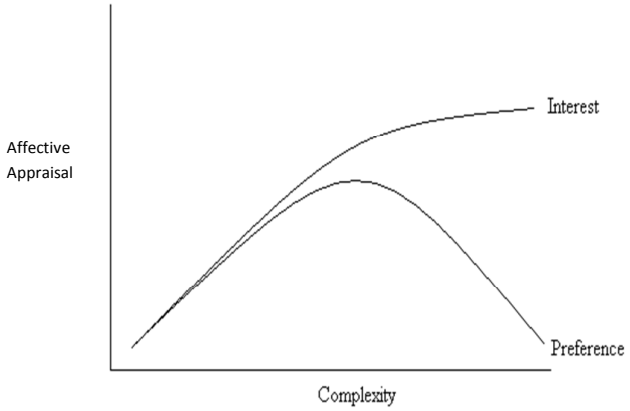


Figure 3: Line drawings mean responses across groups

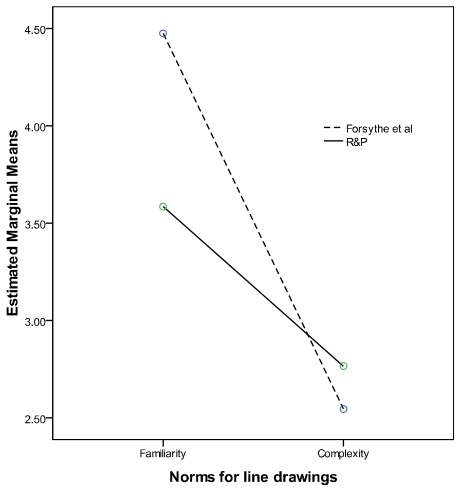


Figure 4: Grey scale mean responses across groups

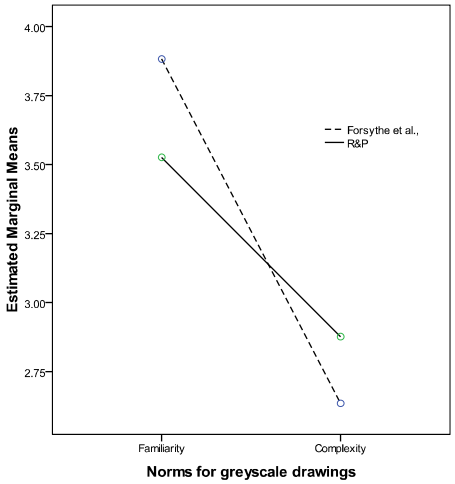


Fig 5 Human judgments of beauty, contrasted with computerised measures of complexity

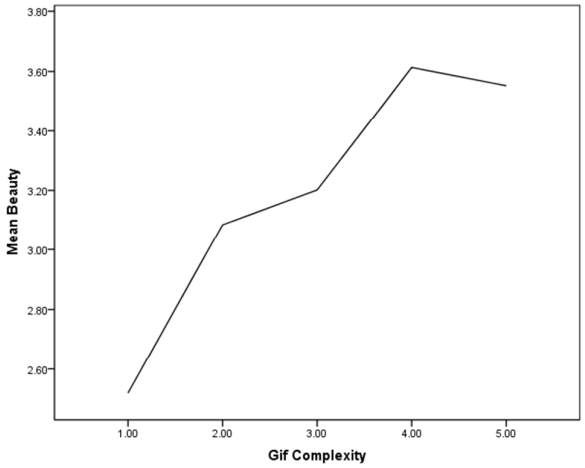


Fig 6 Human judgments of beauty, contrasted with human judgements of complexity