Accepted for publication in *Learning and Individual Differences*

More than the eye of the beholder: The interplay of person, task, and situation factors in evaluative judgments of creativity

Damian P. Birney*

University of Sydney, Australia

Jens F. Beckmann

Durham University, UK

&

Yuan-Zhi Seah

University of Sydney, Australia

Running Head: Evaluating Creative Products

Keywords: Creativity; Evaluation; Judgment; Accuracy

*Corresponding Author

Damian P. Birney

School of Psychology

University of Sydney

Sydney, New South Wales, 2006

damian.birney@sydney.edu.au

Abstract

Judging creativity accurately is difficult. Individuals who are involved in product creation tend to overestimate the creativity of their work; Individuals not involved lack understanding of the creative process that led to the product under scrutiny. We studied creativity judgments in a tripartite person-task-situation framework. Under high, medium, or no structure conditions and different orders of evaluation, participants ($N = 90$) rated the creativity and purchase appeal of products created by themselves and others. Accuracy was defined as differences from consensus evaluations of participants not involved in production ($N = 30$). Moderator analyses suggest that externally set structure of the evaluation process (e.g., using a set of criteria) facilitates the quality of creativity judgment. In unstructured conditions, evaluating one's own product before evaluating a peer's leads to low accuracy, but higher levels of conscientiousness seem to mitigate potentially deleterious effects of lack of structure. Higher levels of openness facilitated accurate creativity judgments of peer-produced products, but not self-produced products. A person-task-situation approach is needed to fully unpack the complexity of processes underlying accurate evaluation of creativity.

Keywords: Creativity; Judgement; Person-Task-Situation; Evaluation Accuracy

1 Introduction

Creativity, when viewed from a product approach, incorporates the conceptual dimensions of novelty and originality and focuses on outcomes that are both useful and appropriate (Barron, 1988; Bleakley, 2004; Nickerson, 1999; Ruscio, Whitney & Amabile, 1998; Torrance, 1988). Innovation, broadly speaking, is the successful implementation of creativity (Hirst, Knippenberg & Zhou, 2009; Hulsheger, Anderson & Salgado, 2009; Klein & Knight, 2005; Mumford & Gustafson, 1988). There has been considerable work investigating the characteristics that make for a creative product and its successful implementation. However, considerably less attention has been given to the processes and structures that people employ in the evaluation of creativity. Evaluation is crucial to both creativity (Mumford, 1999) and innovation (Klein & Knight, 2005), in that it dictates which products to develop further and which to discard. History is littered with artistic and literary works, technological inventions, and scientific discoveries that were initially ignored or even disregarded because of poor or inaccurate evaluations (Elsbach & Kramer, 2003). The current paper contributes to research in the area by investigating factors that underlie accurate evaluation.

Evaluative judgment, as for the study of most human behaviour in psychological research, takes place in the tripartite context of the person, the task, and the situation. The *Person* dimension comprises all that can be subsumed under person-related psychological variables, such as attitudes, skills, abilities, and knowledge. A *Task* is defined as a specified requirement of behaviour (e.g., to solve a problem, to acquire knowledge, or make decisions). Behaviour in this regard is not limited to observable physical acts, it also includes cognition linked to processing information (Ferguson, 1956; Hackman, 1969; McGrath, 1984) – or in our case, evaluation. The *Situation* encompasses the circumstances or the situational context

in which the task is to be performed. A conceptual demarcation between task and situation seems challenging, mainly because experientially every task is linked to a particular situation (i.e., we cannot describe a task without any circumstantial reference). However, because the same task can be presented in different ways and contexts, tasks and situations need to be treated as conceptually independent when studying behaviour (Beckmann, 2010; see also the notion of "task environment" described by Newell & Simon, 1972, p. 55).

In the present study we focus on whether and how *person* and *contextual* factors interactively contribute to creativity judgements. Person factors considered include individual differences in *divergent thinking* skills and various dimensions of *personality*. Situational factors include the *level of involvement* in the actual creation of the product that is to be evaluated. The imposed *structure of the judgment* process and the *order* of judging (whether one evaluates one's own product first or not) are contextual variables also considered a part of the judging situation. In the following section we describe our investigative framework.

1.1 Creativity criteria and evaluation accuracy

Contemporary research has for the most part adopted a product approach to creativity in lieu of focusing directly on the person, the process, or the environment. This has been because, first, creative people are typically judged to be creative by what they produce (Kaufman, Christopher & Kaufman, 2008). Second, product characteristics explain the most variance in evaluations of creativity, far more than person or process dimensions (Demirkan & Hasirci, 2009); and most importantly, third, because a product approach has been seen for some time to provide access to what are considered to be the main contributors of creativity: Environmental factors, processes used, and attributes of the individual producing the creative product (Amabile, 1988).

Derived from this view, the most common method to ascertain product creativity is through some form of expert judgement (e.g., Dailey & Mumford, 2006). This approach has

been applied to the evaluation of real-world creative products ranging from TV shows of the "got talent" variety to Nobel Prizes. Various guidelines have been developed on how to use consensus judgement systematically and rigorously (e.g., Amabile, 1982; Baer, Kaufman & Gentile, 2004). The consensus-based expert judgement regarding the creativity of a given product serves as the reference to determine the 'accuracy' of a specific evaluation. That is, accuracy is defined by the degree of correspondence or alignment between the judgement provided by an individual rater and what a consensus group has agreed upon. In the current research, the consensus group is defined as typical, potential consumers of the created product. We include willingness to purchase the product (*purchase appeal*) as an additional indicator of the *utility* criterion of creativity.

Our study of how individuals deal with the *task* of evaluating a creative product focuses on three *situation* factors, structure, involvement, and order, and two categories of *person* factors, ability and personality.

1.1.1 Structure of evaluation

Structure, in the form of a prescribed set of evaluation criteria, is a situational factor that is expected to impact upon how the evaluation task is performed and consequentially, the quality (e.g., accuracy) of the judgement (Gary, Birney, & Wood, submitted). The most unstructured approach is simply to ask people to provide a summary rating of how creative they believe a product is. This is often the basis of consensus scores where the unstructured, 'naturalistic', or intuitive evaluations of experts are obtained (without using a scoring rubric, Kaufman, Baer, Cole & Sexton, 2008) and then aggregated (Amabile, 1982).

Criterion-bound methods are often developed as alternatives to naturalistic ratings and commonly used to structure evaluation (e.g., O'Quin & Besemer, 2006). Besemer (2000) argues that while natural, intuitive judgments are useful, they can result in snap judgments and less considered processing. Structured evaluation methods enforce a more conscious and

deliberative evaluation process (Wood, Beckmann & Birney, 2009, Beckmann, Beckmann, Birney & Wood, 2015). Gary, Birney, and Wood (submitted) have argued for the efficacy of using similar approaches for structuring analogical reasoning. Positioned between a naturalistic, intuitive evaluation process and a structured approach is the *implicit criteria evaluation method*, which relies on individual evaluators explicating their own implicit judgment standards and using these as criteria for summary judgments (Weinstein, 1980).

The general expectation is that structured evaluation methods lead to more effective and generally more accurate judgments (Beckmann & Schumacher, 2004; Meehl, 1954). We therefore hypothesize that all else equal, structured evaluation methods will result in more accurate evaluations for creativity and purchase appeal of a product.

1.1.2 Involvement

Another situational variable that is in the focus of this study is the level of *involvement* in the creation of the product to be judged. *Product-involved evaluators* are likely to be more knowledgeable about the product and make evaluations cognizant of the idiosyncratic features of the creation process. Involvement in product development may also create higher levels of vested interests in favourable evaluations than would be expected of uninvolved, more impartial judges. Product-involved evaluators can be *self-evaluators* who are directly involved or chiefly responsible for the creative output under scrutiny, or *peer-evaluators* who are either only marginally involved in the creation of the product or have been involved in the creation of a similar product but not the one under scrutiny.

Domain experts seem to use different implicit criteria than laypeople for evaluating creativity (Runco & Bahleda, 1986; Sternberg, 1985). Evaluation criteria not only differ as a function of expertise, they also tend to vary intra-individually (Charles & Runco, 2001; Runco & Chand, 1994). In our study we specifically compare the evaluations made by product creators (self) with those provided by others who were engaged in the same task but

not having produced the presented evaluation target (peers). Within social comparison research, the consistently documented above-average effect, where people rate themselves to be above average on an assortment of traits and attributes, is hinged on a difference in perspective between people evaluating themselves versus evaluating others (e.g., Klar & Giladi, 1999; Chambers & Windschitl, 2004). One explanation proposed for differences in creativity judgments is that unequal involvement and familiarity with the creation process and the product leads to biases and/or differences in cognitive processing (Chambers & Windschitl. Runco and Smith (1992), on the other hand, found that people were more accurate at evaluating the originality (in terms of statistical rarity) of their own responses in a divergent thinking test, than of responses provided by others.

Product-uninvolved evaluators can be sub-divided into *judges* and *consumers*. Judges are 'Appropriate Observers' (Amabile, 1982) or 'Domain Gatekeepers' (Csikszentmihalyi, 1990) with established expertise to (arguably rightly) determine whether the output is deemed to be creative or not. Consumers on the other hand are those individuals directly affected by, or are the intended target group for the product in question. They are not necessarily experts and are likely to lack the breadth of experiences that judges possess. We hypothesize that the level of involvement in creation has an impact on the evaluation of a creative product. Products will be rated as being more *creative* and as having higher *purchase appeal* by product-involved raters than by product-uninvolved raters. Also, product-involved individuals will rate their products consistently more favourably than they will rate the products of others who worked on a similar task.

1.1.3 Order of evaluation

The third situational variable included in this study refers to the effects of the chronological context (i.e., *order*) in which judgements are made. Specifically, we are interested in the potential effects of evaluating one's own product after or before evaluating a

peer's product. Two person-related constructs relevant to how individuals deal with the situational factor of evaluation order are (1) *egocentrism* and (2) *trait underestimation of others*. Egocentrism refers to instances when thoughts about the self loom larger than thoughts about others, which in turn results in a disproportionate weighing of self-referent information (Chambers & Windschitl, 2004). In a similar vein, Klar and Giladi (1999) argue that *trait underestimation of others* reflects a lack of awareness about the level of the trait or ability in others. Both lines of social comparison research converge to suggests that someone who self-evaluates first would tend to focus heavily on their own work and rate themselves with limited awareness of the abilities or creativity of the works of others. Conversely, evaluation of a peer's product first would lead to a clearer awareness of the creativity of products produced by others, self-judgments more proportionately weighted, and thus more accurate ratings of creativity and purchase appeal of their own creations.

1.1.4 Individual differences

Person-related facets underlying evaluation are the psychological characteristics of the evaluator. In the present study we aim at exploring how creative abilities of the evaluator (*flexibility* and *fluency* of divergent thinking) as well as personality dimensions contribute to the formulation of a judgement of creativity.

Divergent thinking (DT) ability has been seen as a proxy for creativity with links to evaluative skills (Runco, 2003). Developmentally, high DT individuals are assumed to have had more opportunity to practice and hone creativity-specific evaluation skills than people who derive fewer ideas (Silvia, 2008). However, this is in contrast to the work of Groborz and Necka (2003) who reported no such links between creative abilities and evaluation skills. In our study we further test whether or not divergent thinking facilitates accurate self- and peer-evaluations of creativity, and if so, under what conditions.

The links between creativity and *personality traits* have been extensively studied.

Openness to experience has consistently been found to be positively related to creativity (e.g., King, Walker & Broyles, 1996; McCrae, 1987). Conscientiousness has been found to be negatively related to creative behaviours in the workplace (George & Zhou, 2001). Agreeableness has also been found to be deleterious to creativity (King, Walker & Broyles), whereas positive links with extraversion and neuroticism have been suggested (McCrae). In contrast to the wealth of research conducted on the links between creative ability and personality, only a few studies have looked at the connection between personality and *evaluations* of creativity. Silvia's (2008) finding that openness to experience and agreeableness positively predicts evaluation accuracy while conscientiousness negatively predicts accuracy of self-rated creativity is a rare example.

1.1.5 Moderating effects

Evaluation is a complex task of processing information in the context of the interplay between personal and situational factors. Therefore, considering situational factors such as structure, product involvement, and evaluation context (i.e., order), and person factors such as creative abilities and personality in isolation has low promise of gaining valuable insights. Hence, in our analyses we aim to comprehensively investigate moderating mechanisms through which those factors might contribute to the quality of evaluative judgments of creativity. From the research previously reported, we hypothesise that a structured approach will have a pronounced beneficial affect in biased or unfavourable conditions. That is, when psychological characteristics are at deleterious levels (e.g., when divergent thinking and/or conscientiousness is low), when evaluators are more product-involved, and when self-evaluation is conducted before having the opportunity to appraise the work of peers.

1.2 Aims and Objectives

The aim of the current research is to shed light on the evaluation process and in turn, build on research focused on improving evaluative accuracy. To this end, we conducted a

study that examines: (1) differences in the degree of evaluation structure; (2) differences in task-involvement, (3) differences in order of the evaluation, (4) individual differences with regard to creative abilities and personality traits, and (5) the interactive effects of evaluation structure, task involvement, order of the evaluation, and personal characteristics on accuracy.

We asked participants to create a greeting card and then evaluate the creativity and purchase appeal of both their own card and a card of their peers. Before providing a general summary rating, participants either evaluated the given card along a structured set of criteria, came up with their own implicit criteria against which to judge the cards, or simply provided an intuitive summary rating without any prior evaluation. Half of the participants rated their own card before rating a peer's card (the other half rated their own card last). A separate group of participants served as consensus raters and evaluated a large selection of the created cards, but did not create any themselves. Details of the design are as follows.

2 Method

2.1 Participants

Participants were 120 business students (65 female; mean age 22.58 years, *SD* = 3.47; 40 postgraduates) from a major Australian university. Participants rated their proficiency in English using a visual analogue scale ranging from 0 ("not proficient") to 100 ("very proficient") yielding a mean score of 84.49 (*SD* = 15.54; *range* = 49 – 100).

2.2 The Stimuli: Create a greeting card

Participants were randomly assigned to design either a *Happy Belated Birthday* card or a *Thank You for Your Support* card using *Tux Paint* (www.tuxpaint.org). Tux Paint was chosen over traditional programs such as Microsoft© Paint in order to minimize effects of prior experience. Tux Paint was designed for children, is user-friendly, required minimal training, and offered the experimenters extensive control over program functionality.

2.3 Experimental Conditions

Participants were randomly allocated to either one of the three Evaluation Structure conditions (Groups 1 to 3), or to the consensus condition (Group 4). Order of evaluation was also manipulated. Half of the sample in Groups 1 to 3 evaluated their own card first before rating a random peer's, while the other half evaluated their peer's first.

Group 1 – Structured: Participants completed the *Creative Product Semantic Scale* (CPSS) to structure their judgement process. The CPSS is composed of 9 subscales each containing five semantic differential items (45 items in total) with contrasting adjectives at opposite ends of a seven-point response scale, e.g., common 1-2-3-4-5-6-7 astounding (Besemer & O'Quin, 1999; O'Quin & Besemer, 1989). The CPSS factor analyses into three dimensions – *Novelty* (originality and surprise), *Resolution* (understandable and useful), and *Style* (elegant and well-crafted), as listed in Table A1 of the Appendix. Participants marked the number along the semantic differential they thought best described the card being evaluated. After completing the CPSS, participants' summary ratings of creativity and purchase appeal were obtained. Comprehensive analyses of the CPSS subscale reported in the Appendix indicated the scale to be reliable and appropriate for use.

Group 2 – Unstructured: In contrast to the CPSS's structured approach, participants in the *naturalistic* group evaluated the greeting cards by simply providing a summary rating for creativity and purchase appeal. No additional items or instructions were given.

Group 3 – Semi-Structured: Participants first listed their idiosyncratic evaluation criteria for creative products in general. They then evaluated the greeting cards by rating them on the criteria they had listed before going on to provide summary ratings of creativity and purchase appeal. These evaluations were thus based on participants' own implicit theories of creativity. A total of 190 unique criteria were listed.

Group 4 – Consensus: Each participant in the consensus group provided summary creativity and purchase appeal ratings for 45 greeting cards systematically sampled from the

90 cards designed by participants in the three evaluation groups. These ratings functioned as the basis of the criterion of evaluation accuracy (described in Section 2.5.3). This approach to rating was based on the *Consensus Assessment Technique* (CAT; Amabile, 1982). The CAT is a well-validated method for determining creativity (Baer, Kaufman & Gentile, 2004; Kaufman, Lee, Baer & Lee, 2007) under the condition that participants are: (1) familiar or have experience with the domain in question, (2) make independent judgments, (3) assess more than just creativity, (4) rate targets in relation to each other and not some absolute standard, and (5) rate targets in different random orders.

Greeting cards are arguably familiar to our participants who are likely to have been card creators, card purchasers and card recipients at some point in the past. Fulfilling conditions 2 to 5, each participant rated 45 greeting cards in differing random orders for both creativity and purchase appeal independently (which resulted in each card being rated by 15 independent raters). To assess CPSS subscale adequacy for greeting cards, consensus participants then sorted a selection of the "implicit" criteria listed by Group 3 participants into the nine CPSS subscales (see the Appendix for details).

2.4 Materials

2.4.1 Summary ratings

C*reativity* summary ratings were obtained by asking participants to "Indicate, by marking on the line below, how creative you think this greeting card is." Using a similar prompt, *Purchase Appeal* summary ratings were obtained by asking participants to indicate "how willing you think the average university student in a buying group will be to buy the greeting card you have been presented with." Scores for creativity and purchase appeal were derived as the distance the respondent's mark was from the left-end of the scale (labelled "not at all").

2.4.2 Divergent thinking tests (DTTs)

The *Unusual Uses* subtest from the Torrance Tests of Creative Thinking (TTCT: Torrance, 1974) was selected to derive estimates of creative abilities. The DTT was scored for fluency (number of responses) and flexibility (number of distinct categories a participant's responses fell into). This test is an established measure of divergent thinking (e.g., Hocevar, 1979; Bossomaier, Harre, Knittel & Snyder, 2009).

2.4.3 IPIP personality measure

A 50 item (25 reversed coded) five factor personality measure obtained from the International Personality Item Pool (IPIP) was used in this study. There were 10 items for each factor (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness) with reliabilities ranging from $\alpha = .77 - .82$ (IPIP). Responses were made using a visual-analogue scale ranging from 'very inaccurate' to 'very accurate'.

2.5 Procedure

2.5.1 Evaluation structure (Groups 1 to 3)

Participants in the three Evaluation Structure conditions started the experiment by completing a basic demographic survey and personality measures. They then familiarized themselves with the Tux Paint program before being randomly assigned to either a birthday or thank-you card. Participants were instructed to aim their designs at typical university student consumers and were given 10 minutes to complete the activity. Divergent thinking tests and two card evaluation tasks were given next. Half of the sample evaluated their own card first before rating a random peer's design, while the other half evaluated their peer's first. Participants rated within card type (i.e., those who created a birthday card evaluated birthday cards, those who created a thank-you card evaluated thank-you cards).

2.5.2 Consensus condition (Group 4)

Participants in the consensus group first provided summary ratings of creativity and purchase appeal and then sorted the criteria listed by Group 3 (semi-structured) participants

into CPSS subscales (see Appendix for details). Finally, the consensus participants completed the personality measure, however these data are not considered in the current analyses.

2.5.3 Accuracy scores

Evaluative accuracy is operationalized by the absolute value of the standardized difference between ratings provided by the individual evaluator (self or peer) and the mean consensus rating for that particular card[1]. We chose to *standardize* the difference to adjust for the level of agreement between consensus evaluators for a given card and the general nature of creative products (that some are unambiguously un/creative whereas others inspire more controversy). Thus, deviation from the consensus group when creative merit is unambiguous (i.e., consensus SD is low) is penalized more than when the consensus group have less agreement (i.e., consensus SD is high). We took the *absolute* value because both positive and negative deviations from the consensus mean reflect lower accuracy, albeit in opposite directions. High scores indicate greater misalignment (less accuracy).

3 Results

Comprehensive analyses are reported in two sections. The preliminary analyses considers validation and manipulation checks of the methods and evaluation criteria used, the main analyses consider the hypotheses more specifically.

3.1 Preliminary Analyses

Table 1 presents the descriptive statistics and zero-order correlations for the self, peer, and consensus ratings of creativity and purchase appeal across the 90 greeting cards. Descriptive statistics for accuracy scores are also included. To evaluate the degree of consistency in the ratings of the consensus group participants, intra-class correlations (ICC)

[1]Accuracy was calculated as follows: $accuracy = \left|\frac{participant\ rating - consensus\ mean}{consensus\ standard\ deviation}\right|$

were derived. Overall, there is good consensus-rater agreement in terms of creativity and purchase appeal across the two sets of greeting cards (*Happy belated birthday* cards: creativity ratings ICC = .72, appeal ratings ICC = .73; *Thank you for your support* cards: creativity ratings ICC = .71, appeal ratings ICC = .65). Validation is approached in terms of (1) expected correlations between appeal- and creativity-, and self-, peer- and consensus-evaluations, and (2) the criterion used to rate cards. Inspection of the correlations in Table 1 suggests that across evaluation methods ($N = 90$), ratings of creativity are highly predictive of purchase appeal for both self- and peer-ratings ($r_{SC,SA} = .74$, $p < .001$; and $r_{PC,PA} = .65$ $p < .001$, respectively) – cards that were rated as creative tended to have higher purchase appeal (and vice versa). However for a given card, there was little overall alignment between the creativity and appeal ratings of one's own card and the rating provided by a product-involved peer of the same card ($r_{SC,PC} = .11$, $p = .33$ and $r_{SA,PA} = .08$, $p = .47$). That is, while the product-involved creators align purchase appeal with creativity, there was limited agreement between self- and peer- ratings of the same card. This finding is consistent with Chambers and Windschitl's (2004) account of disproportionate weighing of self-referent information when comparing product-involved and product-uninvolved raters. However, that the consensus participants were aligned but creators were not, suggests that lack of specific evaluation experience is at play (the consensus group rated 45 cards of different types, whereas the creators evaluated only two cards – their own and a peer's – of the same type).

In terms of accuracy, higher self- and peer-ratings tended to be less accurate (were associated with greater deviation from the consensus group), although this effect was stronger for appeal ratings ($r_{acc,self} = .55$, $p < .001$; $r_{acc,peer} = .51$, $p < .001$) than for creativity ratings ($r_{acc,self} = .27$, $p = .01$; $r_{acc,peer} = .24$, $p = .03$).

[Insert Table 1 here]

3.1.1 Evaluation criteria

The CPPS sub-sample (Group1) was considered to investigate the evaluation criteria used during self- vs. peer-evaluation. Differences in the importance of the three higher-level CPSS dimensions, *Novelty*, *Resolution*, and *Style* (see Appendix) can be identified by regressing these dimensions on self- and peer-ratings of creativity and purchase appeal separately.

Overall, the three CPSS dimensions accounted for a large proportion of variance in self- and peer-ratings of creativity[2] (R^2_{self} = .58, $F_{3,26}$ = 11.88, p < .001, and R^2_{peer} = .63, $F_{3,27}$ = 15.27, p < .001), and self- and peer- ratings of purchase appeal (R^2_{self} = .68, $F_{3,26}$ = 18.02, p < .001, and R^2_{peer} = .50, $F_{3,27}$ = 8.95, p < .001). Looking more closely, *Novelty* was significantly and uniquely implicated in evaluations of self- and peer-ratings of creativity ($\beta_{novelty\text{-}self}$ = .61, sr^2 = .36, p < .001; $\beta_{novelty\text{-}peer}$ = .62, sr^2 = .26, p < .001); and self- and peer-ratings of appeal ($\beta_{novelty\text{-}self}$ = .40, sr^2 = .15, p = .002; $\beta_{novelty\text{-}peer}$ = .65, sr^2 = .28, p = .001). Second, whereas *Resolution*, positively and significantly predicted self-ratings of creativity and appeal (creativity: $\beta_{resolution\text{-}self}$ = .48, sr^2 = .09, p = .026; appeal: $\beta_{resolution\text{-}self}$ = .65, sr^2 = .28, p = .001), *Resolution* did not contribute uniquely to peer-ratings of creativity or appeal (creativity: $\beta_{resolution\text{-}peer}$ = -.004, sr^2 ≈ .00, p = .986; appeal: $\beta_{resolution\text{-}peer}$ = .52, sr^2 = .07, p = .062). Finally, although there was significant systematic variability in the degree to which a card's *Style* score was related to creativity and appeal ratings, this variability was "covered" almost entirely by the *Novelty* and *Resolution* scales (although the effect for peer-ratings of creativity approached significance, creativity: $\beta_{style\text{-}peer}$ = .44, sr^2 = .05, p = .057; sr^2 for other ratings were effectively 0).

These results suggest first, that there is systematic variability in the creativity and purchase appeal of the cards, providing the empirical foundation for further conceptual investigations, and second, that participants are systematically sensitive to these criteria differences. In alignment with findings of other research groups, we are able to replicate the

[2] One of the CPSS participants did not provide CPSS ratings for their own card

importance of *Novelty* regardless of perspective (i.e., involvement). Finally, the results suggest differences in self versus peer evaluations in terms of *Resolution* (and possibly the degree of *Style*), encouraging us to explore such differences further.

3.2 Main Analyses

The main focus of our hypotheses is in relation to evaluation accuracy. The basis of all psychometric calculations of accuracy is the raw evaluation and we thus begin our examination of the hypotheses by exploring effects on ratings. We then pursue the follow-on question of accuracy and the extent to which effects are complexly determined by a combination of moderating factors.

3.2.1 Creativity and purchase appeal ratings

Creativity Ratings. To investigate the effect of evaluation structure and involvement on creativity ratings, a Structure (structured, semi-structured, unstructured) x Involvement (Self, Peer, Consensus) x Order (self-first, peer-first) mixed repeated-measure ANOVA was conducted. There was a significant main-effect for involvement, $F_{2,168} = 5.69$, $MSE = 2045.91$, $\eta^2 = .063$, $p = .003$. Planned contrasts indicated that product-involved evaluators (self: $M = 48.10$, $SE = 2.49$, peer: $M = 45.59$, $SE = 2.60$) rated significantly higher than consensus-raters ($M = 38.88$, $SE = 1.15$), $F_{1,84} = 18.65$, $MSE = 5712.10$, $\eta^2 = .182$, $p < .001$. Although self-ratings of creativity overall tended to be higher than peer-ratings, this did not reach statistical significance $F_{1,84} = 0.552$, $MSE = 567.51$, $\eta^2 = .007$, $p = .460$. There were no other significant main-effects or interactions.

Purchase Appeal Ratings. A similar mixed repeated-measure ANOVA was conducted on purchase appeal. There was a significant main-effect for involvement, $F_{2,168} = 13.46$, $MSE = 5222.93$, $\eta^2 = .138$, $p < .001$. Planned contrasts indicated that appeal ratings from product-involved participants (self: $M = 47.52$, $SE = 2.22$, peer: $M = 43.63$, $SE = 2.65$) were significantly higher than for product-uninvolved raters (consensus: $M = 32.82$, $SE = 1.20$),

$F_{1,84}$ = 41.89, *MSE* = 14647.97, η^2 = .333, $p < .001$. However, there were no differences between self and peer-based appeal ratings $F_{1,84}$ = 1.25, *MSE* = 1361.11, η^2 = .015, p = .266. Again, as for the creativity ratings, evaluation structure and order main-effects were not significant and there were no statistically significant interactions.

3.2.2 Accuracy of creativity and purchase appeal ratings

A core hypothesis of the current study is that a structured method of evaluation would lead to more accurate evaluations than a natural, intuitive evaluation due to the more conscious and deliberate evaluation process required. The results of the aforementioned analyses indicate that product-uninvolved consensus raters tend to evaluate products on average significantly lower in both creativity and purchase appeal than the product-involved participants (creators). The deviation from the consensus group was on average approximately one standard deviation (Table 1). Two Structure x Involvement x Order mixed repeated-measure ANOVA were conducted. For accuracy of creativity ratings, none of the main-effects or interactions reached statistical significance. Although the results were largely the same for purchase appeal accuracy, the Structure x Order interaction effect size was sufficiently large to warrant further investigations ($F_{2,84}$ = 2.87, *MSE* = 2.06, η^2 = .064, p = .062). Interaction contrasts indicated that evaluation accuracy was poorest when evaluating the purchase appeal of one's own card first (M_{self_first} = 1.35, SE = 0.15) rather than second (M_{peer_first} = .82, *SE* = 0.15) using a natural, unstructured evaluation method, $F_{1,86}$ = 5.62, *MSE* = 1.98, η^2 = .061, p = .020. When a structured or semi-structured method is used, the order of evaluation did not have a statistically significant effect, $F_{1,86}$ = 0.87, *MSE* = 0.31, η^2 = .010, p = .352 (M_{self_first} = 0.91, *SE* = 0.11; M_{peer_first} = 1.00, *SE* = 0.11). There were no differences in the effect of evaluation order on accuracy of appeal ratings as a function of structure.

In sum, overall, accuracy of creativity evaluations tended not to be influenced by the

structure or the order of evaluation. Consistent with expectations, there was however some evidence that accuracy of purchase appeal evaluations was enhanced by a structured evaluation method, but only when participants were required to make their evaluations without seeing the work of others' first. In other words, a structured approach to evaluation might mitigate potentially detrimental effects of not having the chance of seeing the work of others' before judging the purchase appeal of one's own work. In the final set analyses, presented in the next section, we investigate the extent to which these effects are qualified by individual differences.

3.2.3 Individual differences: Person-Task-Situation effects

Descriptive statistics and zero-order correlations (collapsed across all experimental conditions) between individual characteristics and consensus creativity and appeal ratings are reported in Table 2. None of the personality traits predicted consensus creativity ratings or appeal ratings. The flexibility dimension of divergent thinking was significantly associated with consensus creativity ratings (r = .20, p = .027) and consensus appeal ratings (r = .29, p = .003). With regard to the link between individual characteristics and accuracy of judgement, conscientiousness is significantly associated with accuracy of creativity evaluations of a self-created product (r = .18, p = .048); evaluators with higher levels of trait conscientiousness tended to be less accurate in ratings (i.e., greater divergence from the consensus). Both flexibility (r = -.19, p = .033) and fluency (r = -.20, p = .031) were significantly linked to accuracy of both ratings of self-created products. Higher divergent thinking performance tended to be associated with less divergence from the consensus mean (i.e., higher levels of accuracy). No other significant relationships were observed.

[Insert Table 2 here]

To explore the potential moderating effects of personality, divergent thinking (DT), and structure and order of evaluation on accuracy of creativity and appeal ratings, moderator

analyses extending on the analyses of Section 3.2.2 were run[3]. In addition to *Divergent Thinking* and consistent with Silvia (2008), *Openness*, *Conscientiousness*, and *Agreeableness* were considered as interacting between-subject variables.

For the DT variables and controlling for the separate effect of each, although there was a main-effect for flexibility in that higher scores tended to be associated with more accurate evaluations of creativity ($F_{1,81}$ = 5.48, *MSE* = 2.88, η^2 = .063, p = .022), there were no significant interaction effects. On the other hand, although fluency was not associated with a main effect, it did significantly interact with evaluation structure to influence accuracy of creativity ratings ($F_{2,81}$ = 6.17, *MSE* = 3.25, η^2 = .132, p = .003). In the structured condition (Group 1), *higher* levels of DT fluency was associated with *more* accurate self-evaluations of creativity (β = -.124, t = -2.29, η^2 = .061, p = .025), but this was not the case for the less structured conditions. In sum, flexibility was important for more accurate creativity evaluations regardless of structure, order and product-involvement. Fluency had a more specific effect in that it was facilitative in structured situations of one's own card but had no moderating effect under other conditions.

For the personality variables, the null main-effects and interactions for structure and order reported in Section 3.2.2 persisted, however *Openness* and *Conscientiousness* but not *Agreeableness*, moderated these null-effects as three-way interactions.

Overall, *Openness* interacted with structure of evaluation and product-involvement, $F_{2,78}$ = 5.13, *MSE* = 1.785, η^2 = .116, p = .008. Unpacking this interaction revealed that under highly structured conditions (Group 1 – CPSS), *Openness* was associated with marginally higher accuracy in creativity rating of a peer's card (β = -.192, t = -1.74, η^2 = .037, p = .086), but had no effect (tending toward lower accuracy, but not significantly so) for creativity rating of one's own card (β = .126, t = 1.26, η^2 = .02, p = .213). Although both simple-effects

[3] Personal characteristics were included as interacting continuous variables in a GLM model of the mixed between/repeated measures ANOVA reported in Section 3.2.2. To ensure sufficient power, the DT variables were considered in separate analyses from the personality variables.

are not statistically significant, the significant overall interaction indicates a trend toward lower accuracy as the comparison moves to the rating of one's own card. That is, openness to externally set structure for judging creativity (e.g., CPSS), facilitated accurate peer- but not self-ratings of creativity.

Conscientiousness interacted with involvement and structure to affect the accuracy of both creativity ratings, $F_{2,78} = 3.77$, $MSE = 1.31$, $\eta^2 = .088$, $p = .027$, and appeal ratings, $F_{2,78} = 5.91$, $MSE = 2.70$, $\eta^2 = .131$, $p = .004$. When evaluations were unstructured, *higher* levels of conscientiousness tended to be associated with *more* accurate peer-evaluations of creativity, $\beta = -.329$, $t = -3.22$, $\eta^2 = .117$, $p = .002$, but not in structured (structured and semi-structured) conditions or when evaluating the creativity of one's own product. Similar results were observed for ratings of appeal. Higher levels of conscientiousness tended to be associated with *more* accurate peer-evaluations of appeal in unstructured evaluations, than under structured, self-evaluation conditions, $\beta = -.261$, $t = -2.22$, $\eta^2 = .060$, $p = .029$.

In summary, we investigated self- and peer- creativity and appeal rating accuracy when evaluating self-created versus peer-created products first, under structured, semi-structured, and unstructured evaluation conditions. Our findings suggest that flexibility – as one dimension of creativity – was important for accuracy under all conditions, and that fluency was important when self-evaluating creativity in structured situations. *Openness* was important for creativity-evaluations of the work of product-involved peers in structured situations. *Conscientiousness* was also important for the evaluation of both the creativity and appeal of peers' work. High levels of conscientiousness can compensate for lack of structure in the evaluation process.

4 Discussion

Individuals are generally poor evaluators of creative products (Elsbach & Kramer, 2003).

Although the current study replicates this result once again, we also showed that accurate evaluation is complexly influenced by a number of person- and situation-related factors in predictable ways. Our core hypothesis was that structured evaluation would subsequently facilitate accurate summary ratings. We further investigated if accuracy depended on whether one evaluated a peer's card before evaluating one's own card, or vice versa, and whether differences in person-related characteristics and product involvement impacted evaluation accuracy interactively. The overall results are clear in that a complex array of person-task-situation factors is at play.

Although there are studies which have reported no effects of a rater's perspective on evaluations (e.g., Alicke, 1985), research within (e.g., Runco & Smith, 1992) and outside of creativity (e.g., above-average effects) has largely found differences between self and peer ratings. Following Dailey and Mumford (2006), we expected differences in creativity and purchase appeal ratings between creators and their peers. Specifically, we expected creators to be more optimistic (i.e., positively biased) in their self-ratings. We found only partial support. Consistent with expectations, self-peer differences were observed in the CPSS criteria used, however these differences did not translate to overall differences in summary ratings. Furthermore, there was no support for the notion of a general self-bias when judging creativity or purchase appeal of products created by oneself or one's peers. There were also no general effects of evaluation structure on accuracy. These null results on a general level were, however, qualified by a range of significant interactions that we argue can be understood through a tripartite framework of person-task-situation interactions.

4.1 Involvement in product creation and order of evaluation

Raters who were involved in the creation of the product under scrutiny, provided consistently higher creativity and appeal evaluations of their own and their peers' work than product-uninvolved participants. When using the aggregated consensus judgment as a

benchmark, product-involved raters were inaccurate by a margin of about one standard deviation. Such inaccuracy is consistent with a social comparison account which proposes that task involvement leads to a disproportionate weighing of self-referent information based on one's own limited experiences (Chambers & Windschitl, 2004; Klar & Giladi, 1999). When judging similar others, people tend to use more self-relevant information compared to when rating dissimilar others (Pollmann, Funkenauer & van Dijk, 2008). Our results seem to resonate with these findings.

Taken together, it was argued that product-involvement would be the basis of an order of evaluation effect, where a self-referent bias would be tempered through the experience of first evaluating the card of a peer. Our results suggested that although order of evaluation was implicated in subtle interactive ways with personality and DT (which we discuss shortly), there was no clear evidence for a general order of evaluation effect. The role of experience, which was proposed to drive the order of evaluation effect, was however observed in the consensus group. Recall that consensus participants provided summary ratings for 45 randomly selected cards of different types. Thus while they were not product-involved in the same way as peer- and self-raters, the experimental procedure ensured consensus raters had evaluation experience and emerging domain-specific knowledge of a broad range of products (cf. Ward, 2008) that creators did not have. The advantage of this type of experience was demonstrated in the data in two ways. First, consensus raters showed significant and substantial intra-class correlations, indicating there was clear agreement amongst consensus raters in terms of the creativity and appeal of the cards. On the other hand, the rating experience of creators was limited to only two cards (their own and a peer's of the same type), and in contrast to the consensus group, there was little convergence (low correlations when rating the same card). It may be the case that the experience of rating only a single card before rating one's own was simply not sufficient to impact summary evaluations overall.

Lack of *specific* experience is often a characteristic of situations where evaluations are required, and all else equal this supports a case for imposing structure under such circumstances (e.g., Beckmann & Schumacher, 2004).

4.2 Structuring the evaluation process

Evaluations do not occur in a social vacuum. Amongst other things, domain knowledge (Ward, 2008), professional training (Runco & Bahleda, 1996), group identity, and norms (Adarves-Yorno, Postmes & Haslam, 2007) can all impose contextual pressures that bias evaluation and impacts accuracy. It was hypothesised that the structured *CPSS* method would lead to the most accurate evaluations of creativity and purchase appeal. Similarly, the *Implicit*, semi-structured method was expected to lead to more accurate summary evaluations than a *Naturalistic*, unstructured method because, as for the CPSS, it was thought to prevent oft-criticized snap judgments (Besemer, 2000). Our findings suggest that whereas there was partial evidence that accuracy of purchase appeal ratings was enhanced by a structured evaluation when participants were required to make their ratings without seeing the work of others' first, overall, structured methods did not lead to increases in accuracy.

While it may be possible that the different evaluation methods did not effectively guide summary ratings of creativity and purchase appeal, this seems unlikely given that the *Novelty* and *Resolution* dimensions of the CPSS were significant predictors of creativity and appeal ratings. Furthermore, findings from the current study show that the CPSS categories accounted for a large amount of variance in summary ratings and was highly representative of laypersons' criteria for creative products.

4.3 Moderating effects of individual characteristics

Against a backdrop of overall misalignment in accuracy of about one standard deviation, and a lack of general effects of evaluation structure and order of evaluation, there were significant person-task-situation interactions. First, in terms of creative ability, previous

research suggests that divergent thinking abilities are positively associated with evaluative accuracy (Runco & Smith, 1992; Silvia, 2008). Accuracy is typically defined in terms of the difference between participants' perception of the statistical rarity (Originality in DT tests) and actual statistical rarity. In our study we defined accuracy as standardized differences from the mean consensus group rating. Under these conditions, flexibility facilitated accurate creativity evaluations regardless of structure, order and product-involvement. Fluency facilitated accuracy in structured situations only when evaluating one's own card.

According to previous research, Openness, Conscientiousness, and Agreeableness were expected to differentially predict accuracy of creativity and appeal ratings (Silvia, 2008). Our findings emphasise the necessity to take more complex person-task-situation interactions into considerations in our efforts to better understand the processes that underpin evaluation accuracy. Under structured CPSS conditions but not other conditions, openness to experience, a preference for variety and open-mindedness, facilitated *more* accurate peer-ratings of creativity. Conscientiousness, a sense of purpose, responsibility and achievement orientation, has been found to be *negatively* related to creative behaviours in the past (George & Zhou, 2001) and our findings were consistent with this, but only when collapsing across all conditions. Closer analyses indicated that conscientiousness might actually be *facilitative* in providing accurate evaluations of creativity, particularly in the absence of structure for evaluations. In unstructured conditions, *higher* levels of conscientiousness tended to be associated with *more* accurate peer-evaluations of appeal than in structured, self-evaluation conditions. Although Agreeableness, being tolerant, trusting and accepting of others, has been found to be deleterious to creativity (King, Walker & Broyles, 1996), in the current study it was not significantly associated with accuracy either overall or as part of a person-task-situation interaction.

4.4 Conclusion

The current research suggests that closer attention should be given to the interactive effects of situation, task, and person characteristics on the evaluation of creativity. The findings suggest overall situation characteristics might obscure more subtle but important interactive effects. We hypothesised that such moderating effects would largely be driven by unfavourable conditions (e.g., lack of experience, lack of structure). That is, when person-task-situation characteristics were challenging, such as when divergent thinking or conscientiousness is low, when participants are more product-involved, and when self-evaluation is conducted in isolation (i.e., without having the opportunity to appraise the work of peers). We have found partial evidence for this to be the case and that the characteristics that facilitate creativity are not necessarily those that make for accurate evaluation of creative products. As such, we encourage researchers to further explore evaluation accuracy within the context of a person-task-situation framework. Although methodologically challenging, such an approach is likely to be needed to fully unpack the complexity of processes underlying accurate evaluation of creativity.

5 Acknowledgements

This research was supported under Australian Research Council's Linkage Projects funding scheme (project LP0669552) and Discovery Projects funding scheme (project DP140101147). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. All required ethics approvals were obtained through the USYD human ethics review committee.

6 References

Adarves-Yorno, I., Postmes, T., & Haslam, S. A. (2007), Creative innovation or crazy irrelevance? The contribution of group norms and level of identity to innovative behavior and perception of creativity, *Journal of Experimental Social Psychology, 43*, 410-416.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621-1630.

Amabile, T. M. (1982) The Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology, 43*, 997-1013.

Amabile, T. M. (1988). A Model of Creativity and Innovation in Organizations. In B. M. Staw, & L. L. Cummings (Eds.), *Research in Organizational Behavior* (pp. 123-167). Greenwich, CT: J.A.I. Press.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*, 113-117.

Barron, F. (1988). Putting creativity to work. In R. J. Sternberg (Ed.), *The Nature of Creativity* (pp. 76-99). Cambridge, England: Cambridge Univ. Press.

Beckmann, J. F. (2010). Taming a beast of burden - On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction 20*(3), 250-264.

Beckmann, N., Beckmann, J.F., Birney, D.P., & Wood, R.E. (2015). A problem shared is learning doubled: Deliberate processing in dyads improves learning in complex dynamic decision making tasks. *Computers in Human Behavior, 48,* 654-662.

Beckmann, J. F., & Schumacher, J. (2004). Urteilsbildung und Entscheidung [Judgement and Decision Making]. In B. Strauß, U. Berger, J. von Troschke & E. Brähler (Eds.),

Lehrbuch Medizinische Psychologie und Medizinische Soziologie (pp. 403-425). Göttingen: Hogrefe.

Besemer, S. P. (2000). To buy or not to buy: predicting the willingness to buy from creative product variables. *Korean Journal of Thinking and Problem-Solving*, *10*, 5-18.

Besemer, S. P., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal*, *12*, 287–96.

Bleakley, A. (2004). Your creativity or mine? A typology of creativities in higher education and the value of a pluralistic approach. *Teaching in Higher Education, 9*, 463-475.

Bossomaier, T., Harre, M., Knittel, A., & Snyder, A. (2009). A semantic network approach to the Creativity Quotient (CQ). *Creativity Research Journal, 21*, 64-71.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130*, 813-838.

Charles, R., & Runco, M. A. (2001). Developmental trends in the evaluative and divergent thinking of children. *Creativity Research Journal*, *13*, 415-435.

Csikszentmihalyi, M. (1990) The domain of creativity. In M.A. Runco and R.S. Albert (Eds.). Theories of Creativity (pp. 190–212). Newbury Park, C.A: Sage

Dailey, L.R., & Mumford, M.D. (2006). Evaluative aspects of creative thought: Errors in appraising the implications of new ideas. *Creativity Research Journal, 18*, 367-384.

Demirkan, H., & Hasirci, D. (2009). Hidden Dimensions of Creativity Elements in Design Process. *Creativity Research Journal*, *21*, 294-301.

Elsbach, K. D., & Kramer, R. M. (2003). Assessing creativity in Hollywood pitch meetings: Evidence for a dual-process model of creativity judgments. *Academy of Management Journal, 46,* 283-301.

Ferguson, G. A. (1956). On transfer and human ability. *Canadian Journal of Psychology, 10*, 121-130.

Gary, M. S., Birney, D. P. & Wood, R. E., (submitted). Intuitive versus semi-structured managerial reasoning by analogy.

George, J. M., Zhou, J. (2001). When openness to experience and conscientiousness are related to creative behavior: An interactional approach. Journal of Applied Psychology, *86*, 513-524.

Groborz, M., & Necka, E. (2003). Creativity and cognitive control: Explorations of generation and evaluation skills. *Creativity Research Journal, 23*, 183-197.

Hackman, J. R. (1969). Towards understanding the role of tasks in behavioural research. *Acta Psychologica, 31*, 97-128.

Hirst, G., van Knippenberg, D., & Zhou, J. (2009). A Multi-Level Perspective on Employee Creativity: Goal Orientation, Team Learning Behavior, and Individual Creativity. *The Academy of Management Journal, 52*, 280-293.

Hocevar, D. (1979). Ideational fluency as a confounding factor in the measurement of originality. *Journal of Educational Psychology, 71*, 191-196.

International Personality Item Pool (n.d.). *A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences*, Retrieved March 20, 2010, from http://ipip.ori.org/

Kaufman, S. B., Christopher, E. M., & Kaufman, J. C. (2008). The genius portfolio: How do poets earn their creative reputations from multiple products?, *Empirical Studies of the Arts, 26*, 181-196.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal, 20*, 171-178.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*, 96-106.

King, L. A., Walker, L., & Broyles, S. J. (1996). Creativity and the five-factor model. *Journal of Research in Personality, 30*, 189-203.

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*, 585–594.

Klein, K.J. & Knight, A.P. (2005). Innovation Implementation: Overcoming the challenge. *Current Directions in Psychological Science, 14*, 243-246.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*, 1258-1265.

McGrath, J. E. (1984). Groups: Interaction and performance. Englewood Cliffs, NJ: Prentice-Hall.

Meehl, P.E. (1954). *Clinical versus statistical prediction*. Minneapolis, MN: University of Minnesota Press.

Mumford, M. D. (1999). Blind variation or selective variation: Evaluative elements in creative thought. *Psychological Inquiry, 10*, 344 – 348.

Newell, A., & Simon, H. A. (1972). Human information processing. Englewood Cliffs, NJ: Prentice-Hall.

Nickerson, R. S. (1999). Enhancing creativity. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 392-430). New York: Cambridge University Press.

O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised Creative Product Semantic Scale. *Creativity Research Journal*, *2*, 267-278.

O'Quin, K., & Besemer, S. P. (2006). Using the creative product semantic scale as a metric for results-oriented business, *Creativity and Innovation Management, 15*, 31-41.

Pollmann, M. M. H., Finkenauer, C., & van Dijk, W. W. (2008). The Order Effect in Self-Other Predictions: Considering Target as a Moderator. *European Journal of Social Psychology, 38*, 315-332.

Runco, M. A. (2003). Idea evaluation, divergent thinking, and creativity. In M. A. Runco (Ed.), *Critical creative processes* (pp. 69-94). Cresskill, NJ: Hampton Press.

Runco, M. A., & Bahleda, M. (1986). Implicit theories of artistic, scientific, and everyday creativity. *Journal of Creative Behavior*, *20*, 93-98.

Runco, M. A., & Chand, I. (1994). Problem finding, evaluative thinking, and creativity. In M. A. Runco (Ed.), *Problem finding, problem solving, and creativity* (pp. 40-76). Westport, CT: Ablex.

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences, 13*, 295-302.

Ruscio, J., Whitney, D. M., & Amabile, T. M. (1998). Looking inside the fishbowl of creativity: Verbal and behavioral predictors of creative performance. *Creativity Research Journal*, *11*, 243-263.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*, 139–146.

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49*, 607–627.

Torrance, E. P. (1974). *The Torrance Tests of Creative Thinking–Norms—Technical Manual Research Edition—Verbal Tests, Forms A and B—Figural Tests, Forms A and B.* Princeton, NJ: Personnel Press.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 43-75). New York: Cambridge University Press.

Ward, T. B. (2008). The Role of Domain Knowledge in Creative Generation, *Learning and Individual Differences, 18*, 363-366.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39,* 806–820.

Wood, R.E., Beckmann, J.F., & Birney, D.P. (2009). Simulations, learning and real world capabilities. *Education + Training, 51*, 491-510.

7 Appendix

The CPSS, as described by O'Quin and Besemer (2006, p. 35) is underpinned by a three-dimensional model which composes *Novelty*, *Resolution* and *Style*. The novelty dimension with its subscales *Surprising* and *Original* refers to the degree of "newness in the product or the idea". The resolution dimension with its sub-scales *Logical*, *Useful*, *Valuable* and *Understandable* refers to the degree to which the product "does what it is supposed to do" or "meets the needs". The style dimension with its subscales *Organic*, *Well-crafted* and *Elegant* refers to "the degree to which the product combines unlike elements into a refined, developed, coherent whole". The appropriateness of the CPSS for its current use was determined in two ways, (1) for capturing variability in implicit criteria of creativity, and (2) psychometric indicators of reliability and validity.

7.1 Utility of the CPSS to capture implicit criteria of creativity

The CPSS is expected to capture the implicit criteria provided by participants in the semi-structured condition (Group 3). The *criteria* developed by these participants were investigated to assess the extent greeting cards fit with the existing CPSS subscales. The 30 participants in the semi-structured condition came up with 190 unique criteria of creativity that were collated into five *Implicit Criteria sets* (38 criteria in each). Participants in the Consensus condition were presented criteria from 3 of the 5 sets sequentially in a random order with the requirement to sort them into one of the 9 CPSS subscales (or to indicate whether the implicit criterion was "important but does not fit into the CPSS subscales', or that 'the criterion was irrelevant"). The random allocation resulted in each criterion being categorized by 18 consensus participants.

A total of 3420 categorizations were made. Of these, only 254 (7.22%) were sorted as important but not fitting the CPSS categories (see subscales in Table A1). The results of the sorting indicate that implicit criteria of creative products are well-captured by the nine CPSS

subscales, adding further support for the validity and current use of the CPSS scale. The categories used give insights into sorting values. The most commonly used criteria were first *Originality* (17.87%) and then *Usefulness* (13.1%), followed by *Valuable* (10.00%), *Elegant* (8.81%), *Well-crafted* (8.47%), *Surprising* (7.76%), *Logical* (6.53%), *Understandable* (6.48%), and *Organic* (2.59%). The remaining 11.19% of criteria were categorized as *Not Important*.

7.2 Psychometrics of the CPSS

Participants in the CPSS evaluation group (n = 30) used the 9 CPSS subscales to evaluate both their own and a peer's greeting card. Table A1 reports the descriptive statistics and correlations for the *CPSS* subscales. All subscales had sufficient levels of reliability, apart from 'Understandable' in peer evaluation. In terms of validity, subscales belonging to the same dimension for both self and peer evaluation were significantly correlated although several significant correlations across dimensions were also observed.

[Table A1 here]

8 Tables

8.1 Table 1

Descriptive statistics for creativity ratings, purchase appeal ratings and accuracy

Rating		Mean	SD	CC	CA	SC	SC_{Acc}	SA	SA_{Acc}	PC	PC_{Acc}	PA
Consensus	Creativity (CC)	38.88	10.82									
	Appeal (CA)	32.82	11.22	.85 **								
Self	Creativity (SC)	48.10	23.43	.26 **	.17							
	Accuracy (SC_{Acc})	0.98	0.66	-.29 **	-.28 **	.27 **						
	Appeal (SA)	47.52	22.69	.20	.09	.74 **	.12					
	Accuracy (SA_{Acc})	1.02	0.76	-.14	-.23 *	.33 **	.46 **	.55 **				
Peer	Creativity (PC)	45.59	24.57	.23 *	.23 *	.11	.04	.07	.03			
	Accuracy (PC_{Acc})	0.95	0.76	-.03	-.10	.12	.17	.06	.06	.24 *		
	Appeal (PA)	43.63	25.17	.19	.22 *	.03	.01	.08	.02	.65 **	.35 **	
	Accuracy (PA_{Acc})	0.97	0.83	-.19	-.25 *	-.02	.13	.05	.17	.30 **	.51 **	.51 **

Note. CC: Consensus-rated Creativity; CA: Consensus-rated Appeal; SC: Self-rated Creativity; SA: Self-rated Appeal; PC: Peer-rated Creativity; PA: Peer-rated Appeal; * $p < .05$ (two-tailed). ** $p < .01$ (two-tailed). $N = 90$.

8.2 Table 2

Descriptive statistics of psychological characteristics and correlations with ratings and accuracy

	Mean	SD	CC	CA	SC_{Acc}	PC_{Acc}	SA_{Acc}	PA_{Acc}
Neuroticism	3.88	1.48	.01	-.02	-.04	-.02	-.03	-.01
Extraversion	5.93	1.45	.11	.08	.04	-.05	.13	-.13
Openness	6.76	1.29	.13	.14	-.04	-.10	.08	-.15
Agreeableness	6.61	1.22	-.03	-.03	.07	-.10	.09	-.09
Conscientiousness	6.30	1.45	-.07	-.12	.18 *	-.14	.12	-.10
DT Fluency	10.08	3.96	.08	.16	-.10	.04	-.19 *	-.01
DT Flexibility	3.20	1.04	.20 *	.29 **	-.19 *	-.13	-.20 *	-.05

Note. CC: Consensus-rated Creativity; CA: Consensus-rated Appeal; SC: Self-rated Creativity; PC: Peer-rated Creativity; SA: Self-rated Appeal; PA: Peer-rated Appeal; * $p < .05$ (two-tailed). ** $p < .01$ (two-tailed). $N = 90$.

8.3 Table A1

Descriptive Statistics for CPSS subscales

Dimensions	Subscale	1	2	3	4	5	6	7	8	9
Novelty	1. Original		.70**	-.20	-.42*	-.18	.21	.32	.25	-.18
	2. Surprise	.71**		-.42	-.63**	-.26	.24	.28	.15	-.24
Resolution	3. Logical	.10	.02		.75**	.74**	.43*	.39*	.60**	.64**
	4. Understandable	-.11	-.22	.80**		.56**	.27	.28	.44*	.56**
	5. Useful	.12	.01	.74**	.74**		.48**	.48**	.66**	.72**
	6. Valuable	.14	.10	.71*	.69**	.86**		.67**	.71**	.35
Style	7. Well-crafted	.21	.13	.64**	.53**	.63**	.74**		.82**	.48**
	8. Elegant	.16	.07	.63**	.57**	.63**	.68**	.85**		.53**
	9. Organic	.17	.02	.76**	.57**	.65**	.67**	.75**	.70**	
	Self (Intra-personal)									
	Mean	3.76	3.53	5.06	5.02	4.77	4.50	4.06	4.23	4.47
	SD	1.12	0.96	1.05	1.08	0.99	1.03	0.99	0.95	1.27
	Cronbach α	.78	.79	.68	.75	.77	.74	.70	.74	.78
	Other (Inter-personal)									
	Mean	3.69	3.51	4.70	4.96	4.67	4.03	3.88	3.93	4.56
	SD	1.22	1.19	1.23	0.90	0.91	0.90	1.04	1.00	1.31
	Cronbach α	.86	.84	.76	.24	.74	.77	.80	.67	.86

N = 30. Each subscale score is composed of 5 items; Upper-triangle indicates correlations for peer evaluations; Lower-triangle indicates correlations for self evaluations.