

Assessing Raters: What Factors Predict Discernment in Novice Creativity Raters?

Journal of Creative Behavior, 56(1), 41–54.

<https://doi.org/10.1002/jocb.515>

Journal of Creative Behavior に掲載されている論文（レビュー論文を除く）の文献引用数の平均値を 95%信頼区間で推定した

- Web of Science で Journal of Creative Behavior の文献 1000 件を抽出し、不備のある（論文のキーワードが空白の）文献を除いた 298 件を標本とした
- 298 件の標本データを基に、母平均が 95%の確率で存在する範囲を推定した
 - ◇ その範囲（信頼区間）は、 $56.87 \leq \text{母平均} \leq 62.42$

創造的なものと創造的でないものを識別することができることは、何の要因が影響しているのか？

- 本研究では、標準的な方法で創造性評価のパフォーマンスを調査し、創造性の基準の評価を確保し、評価パフォーマンスと多様な創造性、性格特性との関連をみることで、先行研究の知見を拡張した

拡散的思考課題（DT）は、創造性研究の重要な役割を担ってきた(Runco, 2010; Runco et al., 2016)

- 拡散的思考は、解答のない問題の独自の解決策を生成するものとして理解されている(e.g., Yoruk & Runco, 2014)
 - ◇ 例えば、代替使用課題(AUT; Guilford, 1967)は、日常的な物（ペーパークリップなど）の新奇で独創的な使用方法を生成するよう求める課題である
 - ◇ 拡散的思考 (DT) 課題は、創造性のポテンシャルを測る指標として確立している(Runco & Acar, 2012)
 - しかし、教示の方法(Acar, Runco, & Park, 2020; Forthmann et al., 2016)やスコアリングの方法(Reiter-Palmon, Forthmann, & Barbot, 2019)に応じて、結果が変動してしまう問題がある
- DT 課題のパフォーマンスは、量的にスコアリングする方法もあれば（流暢性など）、創造性を質的にスコアリングする方法もある(Forthmann, Szardenings, Dumas, & Feist, 2020)
 - ◇ 大抵の場合、独創性や独自性に焦点が当てられている(e.g., Runco et al., 2016; Torrance, 1974)
 - というのも、回答の中での希少性を求める客観的な方法で測定することができるから
 - ◇ しかし、独創性は創造性の重要な側面のうちの 1 つ（新奇性）しか考慮しない
 - 創造性には、新奇性と適切性(実用性)の両方が必要であると言われている(Diedrich, Benedek, Jauck, & Neubauer, 2015; Runco & Jaeger, 2012)
 - ◇ さらに、得られた回答サンプルに基づいて、独創性に相当する希少性を客観的に算出する方法では、同様な回答のわずかな違いが生む独自性を評価することができない(Silvia et al., 2008)
 - ◇ 以上の問題を回避するために、回答者とは別の、訓練されたノービスの評定者による創造性

の主観評価がよく行われる

- この DT 課題のパフォーマンスのスコアリング方法は、人の日常的な創造性の判断を反映し (Harrington, 1975), DT 課題の指示内容「できるだけ創造的なものを生成するように」と適合する (Reiter-Palmon et al., 2019)

📌 これまで、評価者の評価精度を測定するために様々な方法が取られてきた

- 伝統的には、ヒット率 (Runco & Dow, 2004; Runco & Smith, 1992) や不一致スコア (Grohman, Wodniecka, & KŁsak, M., 2006) などの観点で測定されてきた
- Silvia (2008) は、創造性の評価精度を測るには真の創造性評価値が必要であるが、創造性では利用できないので、代わりに評価者間の合意の程度に着目することを提案した
 - ✧ この方法では、絶対的な評価者間の合意ではなく、評価者の一貫性の観点から、評価者間信頼性を求めている (Shrout & Fleiss, 1979)

- DT 課題のスコアを計算するときに、評価者の評価の甘さを無視することになる

📌 Silvia (2008) は、アイデア生成者自身と外部の評価者との評価を比較した

- 生成者は、自分の生成した中で最も創造的なアイデアを選択できる傾向にあり、さらに、創造的で開放性の高い生成者は、外部の評価者が創造性を高く評価したアイデアを選択することができていた
 - ✧ 創造性のポテンシャルを独立して測定していないため、創造的な生成者のアイデアには、他の者よりもより創造的なアイデアが含まれていたから、選択精度が高かった可能性がある

📌 Grohman et al. (2006) は、評価者内の創造性評価と評価者間の創造性評価について調査した

- 評価者の創造性ポテンシャルと評価の精度に有意な関連は確認されなかった
 - ✧ 先に行った自分のアイデアの創造性評価が、後の他者のアイデアの創造性評価に影響していた可能性が明らかになっていない

📌 Benedek et al. (2016) は、創造的な評価者は他者のアイデアの評価精度が高いことを示した

- 評価者の創造性と評価精度を独立して測定しても、この関連が確認される

📌 拡散的思考能力以外で測定できる創造性のポテンシャルに関しても、この関連が確認される

- 創造的活動に頻度良く従事している者 (Tan et al., 2015), 高い創造的成果がある人 (Benedek et al., 2016)
- ノービスよりも専門家の方が評価の合意の程度が高いが、互いの評価には正の相関関係が見られ (Kaufman, Gentile, and Baer, 2005), 十分な評価者数とアイデア数を用いるとノービスでも合意の程度が高い (Hass, Rivera, & Silvia, 2018)
 - ✧ まとめると、ノービスは一般的に拡散的思考課題の評価者としてうまく機能するかもしれないが、中でも創造的な人はさらに評価精度が高いため、ある程度の人数で評価者間信頼性が高くなるかもしれない

📌 別系統の研究では、創造性を過小評価する一般的な傾向が見られることが示唆されている (e.g., Mueller, Melwani, & Goncalo, 2012)

- それによって、信頼できる創造性評価を損ねているのかもしれない
- このネガティブな評価者バイアスは、自分のアイデアの評価よりも他者のアイデアの評価の際

に観察されやすい (Grohman et al., 2006)

- ✧ 評価の合意が高い、創造的活動をよくする人ほど創造性評価が甘い (Tan et al., 2015)
- ✧ 評価精度が高い、経験への開放性が高い人ほど創造性を高く評価する (Benedek et al., 2016; Silvia, 2008)
- ✧ 一方、ナルシズムが高い人ほど自分の創造性のパフォーマンスが（他者と相対的に）低いにもかかわらず、自分の創造的才能を高く評価する (Jonason, Abboud, Tome ´, Dummett, and Hazer, 2017)
 - 同様の関連がアイデアの創造性評価にも観察されているが、有意な差はみられていない (Goncalo, Flynn, & Kim, 2010)
 - ナルシズムは、自分の過大評価だけでなく、他者のアイデアの過小評価も引き起こす可能性も考えられる

📊 DT 課題に関して、評価者の特性を系統的に調査した研究はほとんどなく、評価精度の定義や考慮した性格特性によっても大きく異なっている

- その上、創造性評価を行う状況で評価精度を測定した研究はほとんどない

📊 そこで、本研究は、大量の DT 課題の回答に対するノービスの創造性評価と、評価者の創造性や性格特性を多様に測定し、創造性評価の精度を予測する要因を調査する

- 期待する結果は、評価精度の高さが、創造性のポテンシャルで測定される高い創造性 (Benedek et al., 2016; Silvia, 2008)、創造的活動 (Tan et al., 2015)、創造的成果 (Benedek et al., 2016) に関連するということ
- さらに、評価精度のための評価者バイアスの役割と、それが評価者の創造性や性格特性とどのように関連するのかを検討する
- 加えて、評価のモチベーションや評価時間が評価精度に与える影響や、評価者間信頼性が評価者数にどのように依存しているのかについても調査した

方法

参加者

📊 166 名のうち、6 名を除外した

- 分析対象者は 160 名（77%女性、年齢 $M = 25.92$, $SD = 9.59$, range: 18–60 years）

手続き

📊 参加者は、LimeSurvey を通してオンライン調査を完了した

- 創造性のポテンシャル検査、創造的活動・成果、創造性の自己評価に回答した
- 他者のアイデアの評価に対するモチベーションに回答した
- AUT の回答全 180 個の評価課題に取り組んだ
- 最後に、BFI（ビッグファイブ性格検査）と SD3（short dark triad personality）に回答した

材料

創造性評価

📊 参加者は、全 180 個の AUT の回答の創造性を 4 件法で評価した（0：全く創造的でない～3：非常

に創造的である)

➤ AUT の回答は、タイヤ、ペン、缶の 3 種類で各種類 60 個あった

✧ 創造性検査でも上記 3 種の AUT 課題に取り組むことで、課題の理解を深めさせた

🎨 180 個の回答は、大量の回答プールの中から抽出された

➤ 6 人の非専門家によって、創造性を 4 件法で評価されていた (0: 全く創造的でない~3: 非常に創造的である)

✧ 得られた創造性評価値 (6 人の平均値) を 6 セグメントに分割 (0-0.49, 0.5-0.99, 1-1.49, 1.5-1.99, 2-2.49, 2.5-3) し、各セグメントに該当する AUT の回答を 10 個ずつ抽出した

- 創造性評価は右に裾が長い分布のため (Jauk et al., 2014), これによって回答の創造性がフルスペクトラムのものに直面させることを可能とする

🎨 基準評価として、参加者 160 名の創造性評価の平均値を使用し、それに対する参加者個人の評価パフォーマンスを測った

➤ 評価パフォーマンスの指標として、評価者の評価精度と評価者のバイアスの 2 つを設定した
<評価者の評価精度>

AUT の全 180 個の回答に対する、個人の創造性評価と全体平均の創造性評価の相関

→ 相関が高いほど、個人の参加者と参加者全体の間の創造性評価の合意があることを意味する
<評価者のバイアス>

AUT の全 180 個の回答の、個人の創造性評価平均値から全体の創造性評価平均値を差し引いた値
→ 負の値は全体平均よりも厳しい評価であることを意味し、正の値は全体平均よりも甘い評価であることを意味する

創造性検査

1. 創造性のポテンシャル (CP) を測定

🎨 創造性評価に用いたタイヤ、ペン、缶の 3 種類に関する AUT 課題に取り組んだ (Guilford, 1967)

➤ 創造的なものを生成するように教示 (Nusbaum, Silvia, & Beaty, 2014)

🎨 参加者の回答の創造性を 6 名の評価者が 4 件法で評定した (0: 全く創造的でない~3: 非常に創造的である) $\alpha=.80\sim.81$

➤ 拡散的課題 (DT) の創造性スコアとして、参加者が 3 種類の各 AUT 課題で生成した中で最も創造的な回答 3 つの創造性平均値を求めた 内的整合性 $\omega h=.55$

➤ 拡散的課題 (DT) の流暢性スコアとして、参加者が 3 種類の AUT 課題で生成した総数の平均値を求めた 内的整合性 $\omega h=.87$

2. 創造的活動・成果を測定

<創造的活動>

🎨 Inventory of Creative Activities and Achievements (ICAA; Diedrich et al., 2018) の短縮版を使用

➤ ここ 12 カ月間の創造的活動 (音楽, 料理, 科学など 9 ドメイン) の頻度を 5 件法で回答した (0: 全くしていない~4: 日常的に行っている)

🎨 全ドメインをまとめた平均値を創造的活動のスコアとした 内的整合性 $\omega h=.63$

<創造的成果>

- ✚ 今までで最も創造的な成果を3つ挙げさせ、6名の評価者がその創造性を0～5の6件法で評価し、平均値を創造的成果のスコアとした 信頼性係数 $\alpha = .88$

3. 創造性の自己評価

- ✚ 自分のことをどれほど創造的だと思うかを尋ね、10件法で回答した(1:全く創造的でない～10:非常に創造的である)

タスク要因

- ✚ 創造性評価課題に対するモチベーションに関して10件法で回答した(1:非常に低い～10:非常に高い)
- ✚ 1種の創造性評価課題の完了に要した時間の平均値を評価時間とした

性格特性

- ✚ ビッグファイブ性格検査のドイツ語版(BFI; Rammstedt & Danner, 2017)を使用し、性格特性の構造を測定した
 - 45項目の質問に5件法で回答した
- ✚ マキャベリズム(政治目的のためには反道徳的手段も許されるという考え方)、ナルシズム、サイコパシーの個人差を測定するために、short dark triad 性格検査のドイツ語版(SD3; Malesza et al., 2019)を用いた
 - 27項目の質問に5件法で回答した

結果

記述統計量

TABLE 1. Descriptive statistics and intercorrelations of rating accuracy, creativity, and personality measures.

		<i>M (SD)</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Discernment	(1)	0.55 (0.15)																		
Rater bias	(2)	0.00 (0.34)	.39																	
Sex	(3)	0.77 (0.42)	.13	.03																
Age	(4)	25.92 (9.59)	-.11	-.17	-.22															
Rating motivation	(5)	6.97 (2.03)	.21	.10	.13	-.08														
Creativity self-rep.	(6)	6.63 (1.95)	.17	.09	.08	.09	.31													
Creative activities	(7)	2.44 (0.58)	.14	.15	.13	-.13	.21	.51												
Creative achievement	(8)	1.95 (0.55)	.29	.08	.00	.01	.24	.33	.24											
DT fluency	(9)	4.95 (2.13)	.07	-.03	-.02	.03	.11	.12	.20	-.01										
DT creativity	(10)	1.52 (0.25)	.28	.08	.00	-.07	.24	.31	.34	.24	.55									
Rating time	(11)	236.67 (74.94)	.25	.20	-.17	.17	.08	.06	.06	.24	-.25	.01								
Openness	(12)	3.69 (0.58)	.24	.21	-.06	.08	.32	.62	.50	.45	.20	.37	.27							
Conscientiousness	(13)	3.56 (0.63)	.02	.03	.10	.15	-.02	.10	.05	-.08	-.06	-.08	.03	-.02						
Extraversion	(14)	3.30 (0.72)	.12	.10	.31	.07	.18	.23	.26	.10	.13	.09	.09	.22	.15					
Agreeableness	(15)	3.67 (0.55)	.11	.14	-.02	-.01	.07	.16	.09	.15	-.03	.04	.18	.12	.12	.20				
Neuroticism	(16)	2.99 (0.81)	.11	.03	.17	-.15	-.10	-.18	-.06	-.02	-.08	-.03	-.17	-.15	-.27	-.34	-.26			
Machiavellianism	(17)	2.78 (0.65)	-.07	-.02	.03	-.11	.03	-.04	.03	-.07	.05	.07	-.20	.00	-.13	-.12	-.44	.23		
Narcissism	(18)	2.61 (0.54)	-.03	.01	-.02	.04	.13	.29	.29	.05	.07	.06	.07	.33	.13	.43	-.14	-.27	.30	
Psychoticism	(19)	2.04 (0.49)	.01	-.06	-.20	.00	-.03	-.04	.03	-.07	.06	.07	-.06	.06	-.01	.04	-.49	.09	.45	.41

Notes. $N = 160$; self-rep. = self-reported; DT = divergent thinking; rating time = average rating duration across three tasks in seconds; sex 0 = male, 1 = female; $|r| > .154$: $p < .05$, $|r| > .201$: $p < .01$, $|r| > .253$: $p < .001$.

評価者の評価精度の予測

- ✚ 評価者の評価精度は、創造性の指標（創造的成果、拡散的思考（DT）創造性）、経験への開放性、つく要因（評価のモチベーション、評価時間）と正の相関関係にあった
- ✚ 基準変数（評価者の評価精度）との相関が有意な変数を予測変数とし、すべての予測変数を投入した重回帰分析を行った
 - 誤差項の分布に正規分布を仮定できず、誤差分散の不均一性が見られたため、ブートストラップ法に基づく重回帰分析を行った
 - ✧ モデルは統計的に有意だった ($F_{5,154} = 6.77, p < .001, R^2 = .180, 95\% \text{ CI } [.11, .32]$; see Table 2)
 - ✧ 以上の分析から、評価者の評価精度は、独立して DT 創造性、創造的成果、評価時間によって予測できることが明らかになった
 - 開放性、評価のモチベーションは独立して評価者の評価精度を予測できなかった
 - 2つ目のステップで、モデルに、評価者の評価精度と正の相関がある評価者バイアスを追加した
 - ✧ 拡張したモデルは、評価者の評価精度の説明量を有意に増加させた ($F_{1,153} = 21.51, p < .001; R^2 = .281, 95\% \text{ CI } [.22, .42]$; see Table 2)
 - 評価者バイアスは他の予測変数よりも評価者の評価精度を予測できていた
 - 評価が甘い評価者ほど評価精度が高いと解釈できる
 - 評価者バイアスを予測因子に含めても、他の予測因子に影響しなかった

TABLE 2. Multiple hierarchical linear regression analysis (enter method; bootstrapped with 2000 samples) predicting rater discernment by variables showing significant zero-order correlations with the criterion.

Predictor	β	β 95% CI	p	sr^2	sr^2 95% CI	r (p)	Fit
Step 1							$R^2 = .180^{**}$ 95% CI [.11, .32]
(Intercept)			.138				
Rating motivation	.11	[-.05, .26]	.173	.01	[.00, .06]	.21 (.007)	
Creative achievements	.16	[.00, .32]	.050	.02	[.00, .08]	.29 (<.001)	
DT creativity	.22	[.09, .35]	.008	.04	[.01, .11]	.28 (<.001)	
Mean rating time	.20	[.07, .35]	.009	.04	[.00, .11]	.25 (.001)	
Openness	-.01	[-.21, .19]	.923	.00	[.00, .03]	.24 (.003)	
Step 2							$R^2 = .281^{**}$ 95% CI [.22, .42]
(Intercept)			.012				
Rating motivation	.09	[-.07, .25]	.199	.01	[.00, .05]	.21 (.007)	
Creative achievements	.18	[.03, .33]	.022	.02	[.00, .08]	.29 (<.001)	
DT creativity	.21	[.08, .35]	.006	.04	[.00, .10]	.28 (<.001)	
Mean rating time	.15	[.01, .30]	.045	.02	[.00, .08]	.25 (.001)	
Openness	-.06	[-.28, .13]	.455	.00	[.00, .05]	.24 (.003)	
Rater bias	.33	[.20, .48]	<.001	.10	[.04, .20]	.39 (<.001)	

Notes. CI, 95% percentile confidence interval based on 2000 bootstrapping samples; r , zero-order correlation; sr^2 , squared semi-partial correlation; β , standardized regression weight. ** $p < .01$

評価者バイアスの予測

✚ 評価者バイアスは経験への開放性や評価時間との間に正の相関関係があった

- 重回帰分析の結果、開放性 ($\beta = .17$) と評価時間 ($\beta = .16$) は独立して評価者バイアスを予測していた

信頼できる評価に必要なノービスの評価者数

✚ ブートストラップ法で、評価者の数を 2~160 まで変化させたときの評価者間信頼性を調べた

- どの評価者数においても、1000 のランダムサンプル数で描き、クロンバック α による評価者間信頼性 (IRR) を求めた
- ✧ IRR が平均的に .80 を超えるのは 10 名以上の評価者数のときで、.90 を超えるのは 21 名以上の評価者数のとき

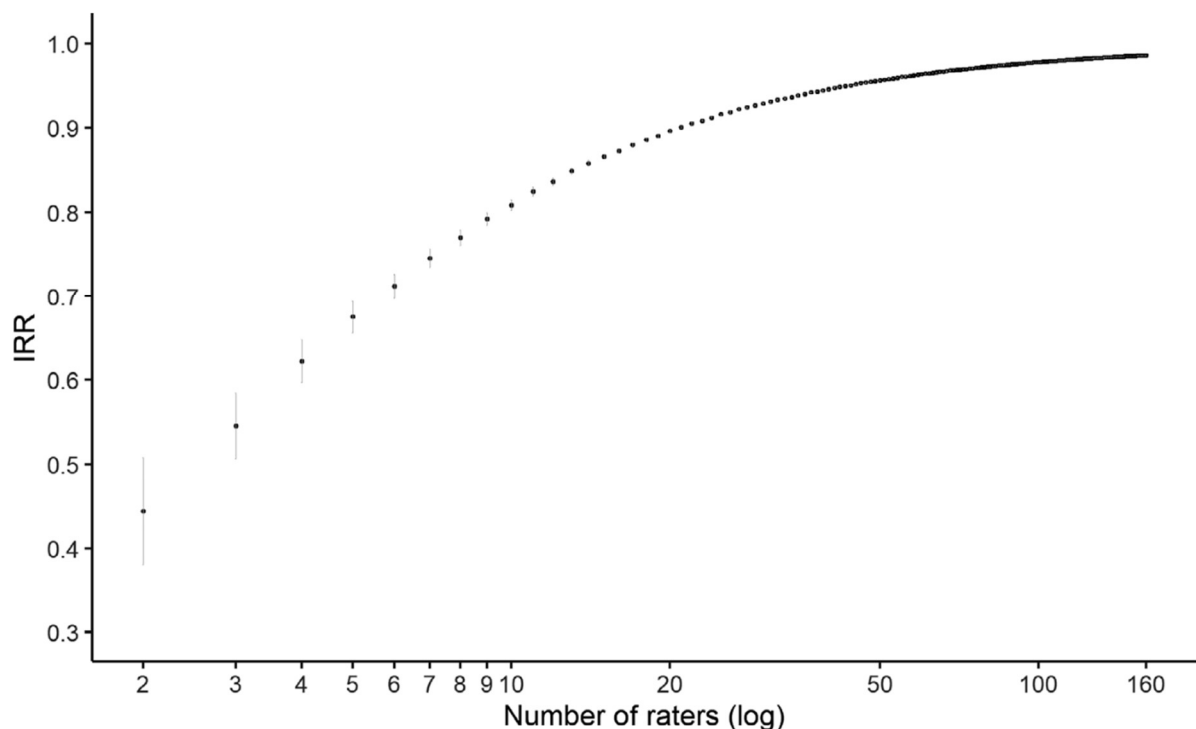


FIGURE 1. The relationship between the number of novice raters (x -axis) and inter-rater-reliability (y -axis). *Note.* IRR = inter-rater-reliability; x -axis is logarithmic; each dot represents the average inter-rater-reliability for a given number of raters after bootstrapping with $N = 1000$; grey vertical lines reflect the standard error.

予備調査のアイデア選択時の評価と本調査の評価の比較

- ✚ 全 180 個の回答に対する、本調査の参加者による評価と元々の予備調査の評価との間には、強い正の相関が見られた ($r = .83, p < .001$)
- ✚ 評価の平均値は、本調査の方が ($M = 1.16, SD = 0.54$) 予備調査よりも低かった ($M = 1.43, SD = 0.85, t = 3.62, p < .001, d = 0.38$)
- この効果は主に、創造性が高い回答に対する創造性評価が、本調査の方が予備調査よりも低いことによって生じている (Figure 2)

- 対照的に、創造性が低い回答に対する創造性評価は、本調査の方が予備調査よりもやや高い
- ✧ 結果的に、一様分布だった予備調査の評価の分布は、本調査では正規分布に近づいた
- ✧ このパターンは、平均への回帰を反映している
 - Figure S3 から、人数の少ない予備調査の評価者による評価平均とその分散との間には逆 U 字型曲線の関係が見られる
 - 極端な評価には、評価者間の合意が必要であることが解釈できる
 - 人数の多い本調査の評価者では、創造的でない回答に対する評価には合意している（分散が小さい）が、創造的な回答に対してはあまり合意していない
 - ✧ 創造的な回答において、創造的でない回答よりも平均への回帰が起こりやすい原因を説明している
 - ✧ 他の説明では、予備調査の評価では、（本調査よりも評価する回答の数が多く）創造的な回答が（他の回答に比べて相対的に）稀であるため、創造的な回答を識別しやすいのかもしれない

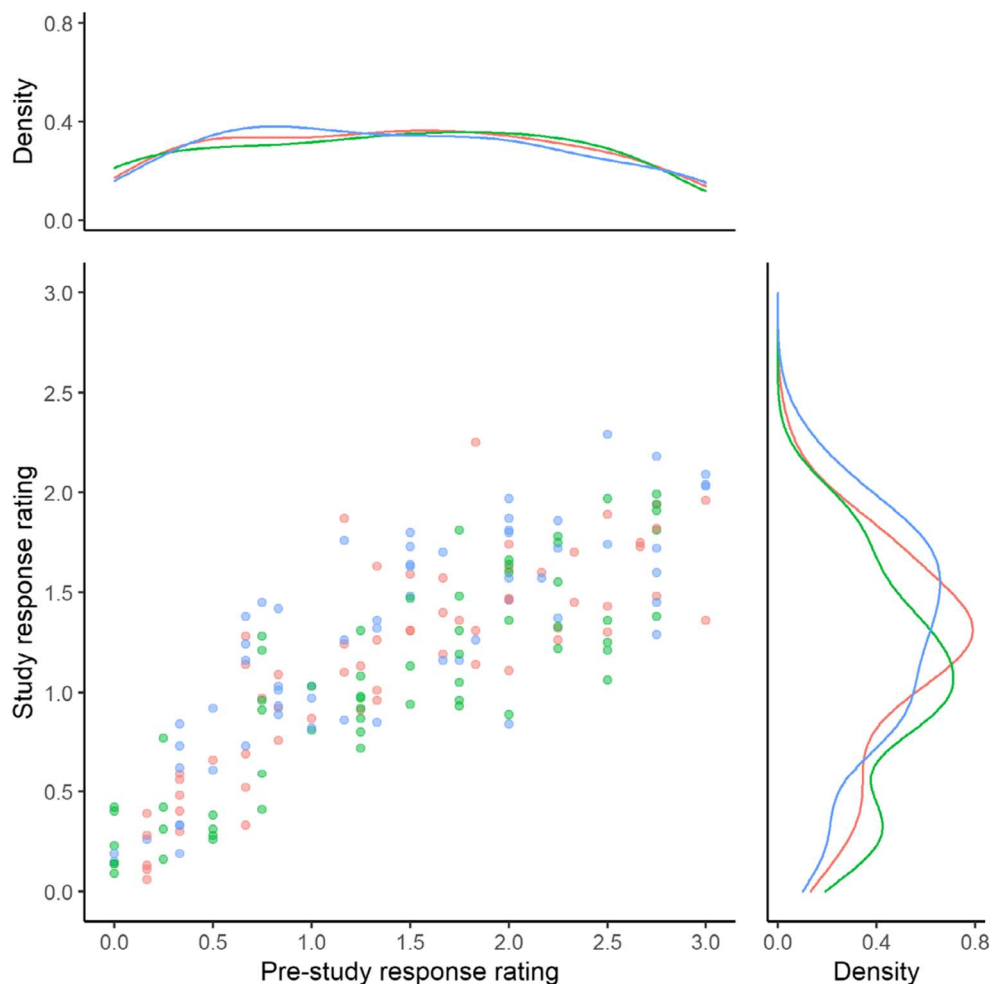
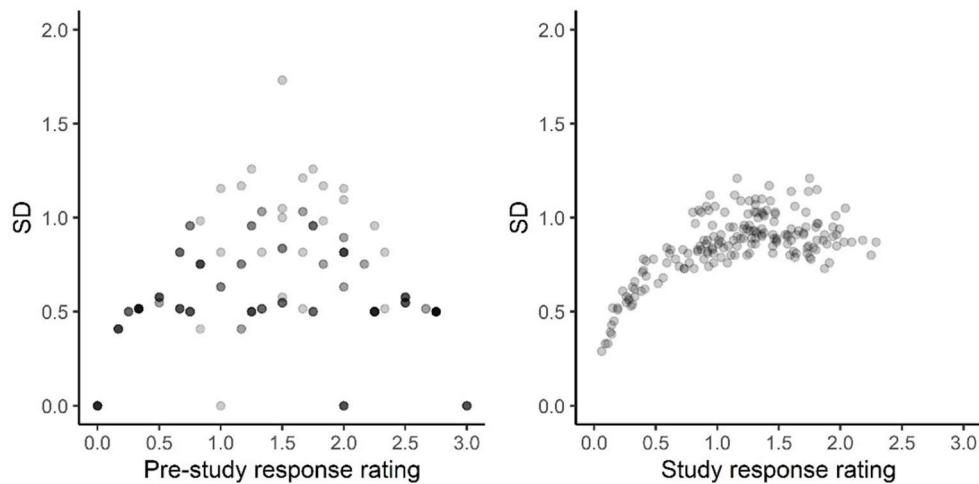


FIGURE 2. Comparison of pre-selected responses regarding their pre-study response creativity ratings (x-axis) by six independent raters (four for the task: pen) and study response ratings by 160 novice raters (y-axis). Notes. Colors refer to different AUT items (blue = tire, red = can, green = pen). Density plots indicate uniform distribution of pre-study response ratings (top) and regression to the mean in the study response ratings (right). [Color figure can be viewed at wileyonlinelibrary.com]

■ **Figure S3.** Distribution of standard deviations for the pre-study response ratings (left) and study response ratings (right).⁴



Note. Each dot represents one AUT idea. Darker points result from overplotting. SD = standard deviation.⁴

考察

<結果のまとめ>

- ✚ 評価者が創造的で開放性が高いほど評価精度が高いことを明らかにした
- ✚ さらに、評価精度の高さは評価のモチベーションが高いことや評価に時間をかけること、評価が甘いことに関連していた
 - このことから、評価者の創造性、性格特性、タスク要因は評価精度に貢献しているといえる

<評価者の創造性>

- ✚ 創造的な人は評価精度の高い評価者であるという先行研究の知見を再現した (Benedek et al., 2016; Silvia, 2008; Tan et al., 2015; but cf. Grohman et al., 2006)
 - 自分のアイデアの判断だけでなく (Silvia, 2008), 他者のアイデアの判断でも同様であることを示した
 - ✧ 生成者と評価者を分けることで、評価するアイデアが同一となるため重要
 - 創造的成果と拡散的思考課題 (DT) のパフォーマンスが高い人が評価精度も高かった
 - ✧ 逆にいうと、評価精度が高い人は創造的なパフォーマンスを発揮できるのかもしれない

<評価者の性格特性>

- ✚ 性格特性に関しては、新しい経験への開放性だけが評価者の評価精度と相関していた
 - これも先行研究と一貫する (Benedek et al., 2016; Silvia, 2008)
 - 追加の分析で、創造的でないアイデアの評価は評価者間で一貫するが、創造的なアイデアはあまり一貫しないことが示された
 - ✧ 創造性の評価は独創性の評価を強く反映しており (Diedrich et al., 2015), 独創性の高いアイデアは特に不適切とみなされる傾向があり、それによって創造性が過小評価されることが示されている (Benedek et al., 2016; Blair & Mumford, 2007; Mueller et al., 2012; Zhu, Ritter, & Dijksterhuis, 2020)
 - ✧ まとめると、開放性の高い人は、全ての回答をより創造的であると判断するのではなく、実

際に創造性の高い回答を明確に判断することができ、全体的に高い評価精度をもたらしていると考えられる

■ 一方、評価精度と dark triad 性格特性との有意な関連は確認されなかった

- 先行研究と同様に (Goncalo et al., 2010; Jonason et al., 2017; Jonason, Richardson, & Potter, 2015), ナルシズムが高い人ほど自己報告の創造性が高かった
- ✧ しかし、他者のアイデア評価をネガティブに評価することは確認されなかった

<評価者のバイアス、タスク要因>

■ 評価者の創造性評価が甘いことが評価精度を最も良く予測した

- 評価の甘さを予測したのは、経験への開放性と評価時間であり、評価時間は評価精度も予測した
- ✧ 創造的な言語的な回答は、大抵長文で複雑であるため (Forthmann, Oyeade, Oyeade, Ojo, Gu ¨ nther, & Holling, 2019), 十分な時間をかけて熟慮することが必要かもしれない

<評価者の人数と評価者間信頼性>

■ 分析で示したように、評価者間信頼性が低下してしまうのは、創造性が高いアイデアに対する評価の合意が低いためである

- とはいえ、評価者が評価の訓練を受けていれば、評価者間信頼性がより高くなると推測する