Judgment and Decision Making, Vol. 17, No. 1, January 2022, pp. 449–486

Both better and worse than others depending on difficulty: Replication and extensions of Kruger's (1999) above and below average effects

Max Korbmacher* Ching (Isabelle) Kwan† Gilad Feldman‡

Abstract

Above-and-below-average effects are well-known phenomena that arise when comparing oneself to others. Kruger (1999) found that people rate themselves as above average for easy abilities and below average for difficult abilities. We conducted a successful pre-registered replication of Kruger's (1999) Study 1, the first demonstration of the core phenomenon (N = 756, US MTurk workers). Extending the replication to also include a between-subject design, we added two conditions manipulating easy and difficult interpretations of the original ability domains, and with an additional dependent variable measuring perceived difficulty. We observed an above-average-effect in the easy extension and below-average-effect in the difficult extension, compared to the neutral replication condition. Both extension conditions were perceived as less ambiguous than the original neutral condition. Overall, we conclude strong empirical support for Kruger's above-and-below-average effects, with boundary conditions laid out in the extensions expanding both generalizability and robustness of the phenomenon.

Keywords: above-average effect, below-average effect, bias, anchoring, egocentrism

*Co-first-author. Department of Health and Functioning, Western Norway University of Applied Sciences, Bergen, Norway, https://orcid.org/0000-0002-8113-2560.

†Co-first-author. Department of Psychology, University of Hong Kong, Hong Kong SAR.

‡Corresponding author. Department of Psychology, University of Hong Kong, Hong Kong SAR. https://orcid.org/0000-0003-2812-6599. Email: gfeldman@hku.hk.

We would like to thank Leo Chan for reviewing the materials during an early project stage and Raj Aiyer, Hirotaka Imada, Matan Mazor, Nicole Russel, Burak Tunca, and Meng-Yun Wang for reviewing the manuscript prior to submission. Their work led to many helpful comments, which improved the project output substantially. We would also like to thank Prasad Chandrashekar for his help with mixed modelling.

All materials, data, and code are available in the OSF supplement at https://osf.io/7yfkc/.

1 Introduction

1.1 Background

The above-average effect refers to the tendency to perceive oneself as better than the average person across different aspects. Kruger (1999) was the first to present instances of the opposite – a below-average effect – the tendency to view oneself as worse than the average person, and he proposed that this opposing effect depends on the difficulty of the ability domain. The above-average effect was observed when self-perceived skills in an ability domain were high, whereas the below-average effect occurred when self-perceived skills were low. Hence, Kruger identified the two effects' underlying mechanism to be the egocentric nature of comparative ability judgments and suggested an anchoring-and-adjustment account. Individuals anchor onto their own skills and then adjust away from their own anchor when judging the skill of others. Therefore, when considering easy activities, people perceive their ability/skill as high and display the above-average effect, thus failing to account for the "true" distribution curve of such abilities/skills which includes others who are also highly skilled. When activities are difficult and hence absolute domain ability is generally low, a below-average effect results from the failure to consider that others are also not highly skilled.

This result was first operationalized in Study 1 in Kruger (1999) using a questionnaire in which participants first compared themselves with their peers on four relatively easy and four relatively difficult ability domains (or activities). Participants then answered a series of questions concerning: 1) estimates of their own and classmates' absolute abilities (termed "comparative ability"); 2) desirability; 3) ambiguity of each ability; and 4) past experience of each ability. A strong negative correlation between domain difficulty and participants' comparative ability judgments supported both above and below-average effects (Kruger, 1999). The study demonstrated correlational evidence for the egocentric nature of comparative ability judgments, in the form of a strong positive correlation between participants' ratings of their own and their comparative abilities. For all ability domains, participant judgments of their own absolute abilities better predicted their comparative ability judgments than did participants' judgments of their peers' skills. Additional experimental studies (2 and 3 in Kruger, 1999) used a situation in which participants received either a very easy or a difficult test, leading to similar results as in Study 1. The anchoring-and-adjustment account was deemed consistent with the fact that cognitive load increased bias during comparative ability judgments.

We conducted a close replication and extensions of Kruger (1999) with two main goals; 1) test the robustness of above- and below-average effects, and 2) examine extensions to test whether ambiguities regarding domain difficulty may moderate this effect. Two between-subject conditions were added to the original design to test whether an easier or more difficult version of Kruger's original ability domains would moderate the effects. Furthermore, we added an additional dependent variable to assess the phenomenon using

ratings of perceived domain difficulty more directly. We begin by introducing the literature on above-and-below-average effects and the choice of target article for replication, then provide information on the original findings, and outline our added extensions.

1.2 Above-and-below-average effects

In the 1980s, researchers began to assess subjects' self-evaluations in relationship to their peers with the results showing over-estimations of own chances for positive outcomes compared to the average population (e.g., Weinstein, 1980, 1983). Focusing on comparisons with others, the phenomenon became later known as above or better-than-average effect (Kruger, 1999). Research picked up quickly on the above-average effect, testing boundary conditions such as culture (Heine & Lehman, 1997) or self-appraisal (Wilson & Ross, 2001). Kruger (1999) was the first to add that there is not only an above- but also below-average effect.

1.2.1 Underlying mechanisms

Throughout the last decades, a range of different underlying mechanisms was proposed to explain the above-average effect (less research focused on the below-average effect), such as informational differences (i.e., knowing more about oneself than others), focalism (i.e., focussing on oneself during comparative judgments), naïve realism, and egocentrism (Brown, 2012). The final mechanism was also used in the chosen study for replication (Kruger, 1999); when people assess how they compare with their peers, they may focus egocentrically on their own skills and insufficiently account for the skills of the comparison group. However, Kruger (1999) reported not only an above-average effect, but also a below-average effect, both explained by egocentrism.

1.2.2 Theoretical grounding

Originally, the above-average effect has been described as motivated by self-enhancement needs (i.e., to induce positive affect towards oneself) or a byproduct of motivated reasoning (Alicke, 1985; Brown, 1986; Kunda, 1990; Taylor & Brown, 1988). Self-enhancement enables the maintenance of a global self-concept allowing for both positive attributes under personal control and negative attributes resulting from factors beyond personal control (Alicke, 1985).[1] Self-verification can be used as another explanation for the above-average effect (Zell et al., 2020). Expanding on self-enhancement, the self-verification theory describes that both self-enhancement and exposure to information which creates and strengthens a biased view of oneself can lead to phenomena such as the above-and-below-average

[1]See Ziano et al. (2021) for a recent successful direct replication of Alicke (1985), showing that people rate more desirable traits to be more descriptive of themselves than of others, and extending that the effect was stronger for more controllable traits. This study was different from the current work as it focused on traits whereas the focus here is on skills.

effects (Zell et al., 2020). In that sense, higher self-esteem has been linked with stronger above-average effects (e.g., Bosson et al., 2000; Chung et al., 2016). Support for the motivational perspective and the ubiquity of the above-average effect was provided by those objectively being below-average in certain characteristics displaying the above-average effect (e.g., Sedikides et al., 2014). For instance, prisoners comparing themselves with non-prisoners on pro-social characteristics rated themselves as above-average in most characteristics (Sedikides et al., 2014). Another explanation can be found in social comparisons during which people evaluate their social position compared to relevant peers – with the tendency of positioning oneself as higher-standing (Gerber et al., 2018). An example of both effects applying during social comparisons is when Democrats and Republicans compare their own warmth and competency with the average person of their in- and outgroup (Eriksson & Funcke, 2013). In-group comparisons lead to below-average ratings for warmth among Democrats and above-average effects among Republicans, which reversed for outgroup comparisons (Eriksson & Funcke, 2013). Above-and-below-average effects have also been found to vary across ages, with egocentrism accounting for age differences (Zell & Alicke, 2011). Young, middle-aged, and older adults displayed an above-average effect for most ability and trait dimensions, whereas a below-average effect was observed for older adults with clear deficiencies (Zell & Alicke, 2011).

1.2.3 Follow-up research

Due to the large number of citations of Kruger's (1999) findings, it is difficult to generalize the publication's impact. However, focusing on follow-up research on the above and below-average effects', more recent studies provided information about the effects' wide applicability and boundary conditions, with a large body of work supporting the original findings (e.g., Aucote & Gold, 2005; Burson et al., 2006; Johansson & Allwood, 2007; Sweeny & Shepperd, 2007). For example, building on the original findings, Giladi and Klar (2002) demonstrated that individual items within a positive group tend to be rated as above-average and individual items within a negative group tend to be rated as below-average. These effects can be reversed depending on the timing of the denotation of the target item, which affects the direction and size of the comparative biases (Windschitl et al., 2008b).

Much subsequent research also continued to explore underlying mechanisms, such as motivations and debiasing factors influencing egocentrically biased comparative judgments. Epley and Caruso (2004) discussed how unconscious, automatic features of human judgment result in egocentric judgments that appear objective to the judges themselves. Windschitl et al.'s (2008a) experiments attempting to debias over-optimism for easy tasks and under-optimism for hard tasks through feedback was only successful under restrictive conditions. Yet, their results support the pervasiveness of egocentric biases as participants failed to generalize non-egocentric tendencies to new contexts.

1.3 Choice of study for replication

Kruger's (1999) work made an important contribution to the field by introducing the below-average effect and conditions in which occurs, which adds to the understanding of a highly prevalent effect with importance to daily reasoning. A recent meta-analysis of better-than-average-effect studies found the effect to be robust across studies, yet, with the effect being smaller for abilities compared to personality traits (Zell et al., 2020). Problematically, definitions and measurement of skill are incongruent which leads to biased assessment and operationalizations differ strongly between studies testing above-and-below average-effects, generally (Zell et al., 2020), and in specific contexts such as drivers' overconfidence in their driving skills (Sundström, 2008). Hence, despite the prolific literature that followed, the above-average effect's robustness has been repeatedly called into question (Sundström, 2008; Zell et al., 2020).

However, some studies failed to conceptually replicate mechanisms and boundary conditions originally reported by Kruger, such as the relationship of estimates about others in relationship to estimates about oneself. For example, Moore and Kim (2003) found mixed evidence for the relationship between comparative ability and the evaluations of others' ability. This was also shown in a practical context by Walsh and Ayton (2009). After presenting an imaginary scenario in which a doctor provides information about a serious diagnosis applying to the participant and how that affects others', own happiness estimates by participants were indeed influenced by information about others' happiness.

We chose Kruger's (1999) study for replication based on the following factors: impact, open questions about boundary conditions of the above and below-average effects, and absence of direct replications. To the best of our knowledge, no direct replications of Kruger (1999) have been publshed. Yet, the article has had a significant impact on several scientific and practical fields, including management (Bazerman & Moore, 2012), economy (DellaVigna, 2009; Koellinger et al., 2007), medicine (Stewart et al., 2013), education, or the workplace in general (Dunning et al., 2004). At the time of writing (May 2021), there were 1178 Google Scholar citations of the article and many important follow-up theoretical and empirical articles (Chambers & Windschitl, 2004; Moore, 2007; Moore & Cain, 2007; Moore & Small, 2007; Whillans et al., 2020; Windschitl et al., 2008b). We chose Study 1, as it was the first demonstration of the core phenomenon. We aimed to revisit this classic phenomenon in a well-powered preregistered close replication (e.g., Brandt et al., 2014).

1.4 Original hypotheses in target article

In the original study, participants compared themselves to their peers on eight ability domains of varying difficulty. Kruger proposed that (H_{orig1}:) compared to judgments of their peers' abilities, people's judgments of their own abilities account for more variance in their comparative ability judgments.

Past research on reasons for people's tendency to focus on their own ability when comparing themselves to others offers insight on why comparative ability judgments are egocentric in nature. One's own skills are more likely to be assessed first when comparing the self to others (Srull & Gaelick, 1983), are easier to conceptualize than skills of the average person (Higgins et al., 1982; Higgins & Bargh, 1987; Srull & Gaelick, 1983), and have a larger database to refer to than others' skills (Ross & Sicoly, 1979). These explanations formed the basis of Kruger's primary hypothesis. When comparing one's own ability to peers' ability, assessments are predominantly based on the perception of one's own skills and less on the perceptions of peers' skills, and therefore, perceptions of one's own absolute ability better predict comparative ability judgments.

Based on that, Kruger proposed that (H_{orig2}:) people tend to perceive themselves as above average when considering easy abilities, and that (H_{orig3}:) people tend to perceive themselves as below average when considering difficult abilities. We merged the dichotomized hypotheses to propose that the more difficult the ability domain is perceived to be, the more likely a person is to shift from perceiving oneself as above average to perceiving oneself as below average.

1.5 Original findings in target article

Kruger (1999) used a combination of correlational studies, one-sample t-tests, and multiple regression and found support for all hypotheses (Table 1). Above and below-average effects were prevalent for all but one difficult item: telling jokes. He observed an inverse association between the domain difficulty and comparative ability: as ability domains increased in difficulty, the perception of their comparative ability decreased. Participants believed to be above average for easy abilities and below average for difficult abilities.

To examine the relationship between one's own absolute ability and comparative ability judgments, we conducted multiple regressions predicting comparative ability from their own ability, and others' ability for each of the eight abilities. Participants' perception of their own ability better predicted their comparative ability judgments. Participants anchored onto their own absolute ability, as opposed to their peers' absolute ability when comparing themselves to others across ability domains. Here we summarize effect sizes and power analysis for the original study results in the sections "effect size calculations of the original study effects" and "power analysis of original study effect to assess required sample for replication" in the OSF supplement.

1.6 Extensions to the Original Study Design

1.6.1 Extension 1: Manipulating domain difficulty

We aimed to extend the replication study by considering the ambiguities in the definitions of easy and difficult used in the domains of the original study. The ability domains in the target article were only succinctly described (see Table 2). Each ability domain may connote

TABLE 1: Kruger's (1999) findings: Mean comparative ability estimates and judgmental weight of own and peers' abilities.

Ability	Domain difficulty[1]	Comparative ability[2]	Judgmental weight of Own ability[3]	Judgmental weight of Others' ability[3]
Easy				
Using a mouse	3.1	58.8**	0.21	0.06
Driving	3.6	65.4****	.89****	–.25*
Riding a bicycle	3.9	64.0****	.61****	–0.02
Saving money	4.3	61.5**	.90****	–.25***
Difficult				
Telling jokes	6.1	46.5	.91****	–0.03
Playing chess	7.1	27.8****	.96****	–.22**
Juggling	8.3	26.5****	.89****	–0.16
Programming	8.7	24.8****	.85****	–0.1

[1] Higher numbers reflect greater difficulty.

[2] Mean percentile estimates above 50 reflect an above-average effect, estimates below 50 reflect a below-average effect.

[3] Standardised betas from multiple regressions predicting participants' comparative ability (percentile) estimates from their estimates of their own absolute ability and the absolute ability of their peers, respectively.

*$p < .05$. **$p < .01$. *** $p < .001$. ****$p < .0001$.

different meanings, depending on how participants interpret the domains. For instance, the ability "saving money" was categorized as an easy ability. Yet, the amount of money saved was not specified, and that may matter for perceived difficulty, as saving 3% of income per month is likely to be perceived as easier than saving 20% of income per month.

Therefore, we manipulated domain difficulty. In our replication, we randomly assigned participants to one of the three conditions receiving different definitions of the ability domains, either: 1) original domain condition (replication); 2) easy domain condition (extension) with an easy reinterpretation of the original domains; or 3) difficult domain condition (extension) with a difficult reinterpretation of the original domains (Table 2).

For the two extension groups, the extension domains aim to be specifically defined in measurable terms. More context is provided for the domains to be more specific, such as the hand used (dominant versus non-dominant hand) for using a mouse, the location and type of car (home country and automatic gear car versus foreign country and manual gear car) for driving, and the help received for computer programming (someone very knowledgeable versus someone not very knowledgeable), which is an ability domain most participants may not have experience with. Additionally, an objective measure should be

TABLE 2: Extension: Manipulation of perceived domain difficulty in target's domains.

Original domain group (replication)	Easy domain group (extension)	Difficult domain group (extension)
Easy domains		
Using a mouse	Using a mouse with your dominant hand	Using a mouse with your non-dominant hand
Driving	Driving a car with automatic gear in your home country	Driving a car with manual gear in a foreign country where people drive on the opposite side of the road
Riding a bicycle	Riding a bicycle for 10 minutes on a flat road	Riding a bicycle for an hour up a road with an upwards incline slope
Saving money	Saving 3% of your income each month	Saving 20% of your income each month
Difficult domains		
Telling jokes	Telling a joke to one person you know well (e.g., friend, family member, etc.)	Telling a joke in front of a live audience in an improv stand-up comedy club
Playing chess	Win a game of chess against an AI (computer) in beginners' mode	Win a game of chess against an AI (computer) in advanced mode
Juggling	Juggling 2 balls	Juggling 4 balls
Programming	Programming guided by someone very knowledgeable in programming	Programming guided by someone not knowledgeable in programming

quantitatively determined in units that can be measured (e.g., length of time, amount of money) or counted (e.g., number of people; Roth et al., 2008). Therefore, the extension domains also use criteria such as time (10 minutes versus 1 hour), number of people (one person versus a live audience in an improv stand-up comedy club), and difficulty (beginner mode versus advanced mode).

1.6.2 Extension 2: Measuring domain difficulty

For the second extension, we added an additional dependent variable measuring domain difficulty. In the original study, domain difficulty was determined in a pretest by a separate group of participants (n = 39). They rated their absolute ability – the extent of how skilled they are – on the eight abilities on a 10-point scale (higher number indicates higher

skill level): "For this ability, please rate your own ability from 1 (very unskilled) to 10 (very skilled)". The ratings were then reverse-scored and higher numbers indicated greater domain difficulty. The four ability domains lower than the midpoint of the scale were categorized as easy domains, whereas the four ability domains higher than the midpoint of the scale were categorized as difficult domains.

Due to problems associated with categorizing the continuous variable of the difficulty level of ability domains into easy domains or difficult domains, in the current replication, we measured domain difficulty on a continuous scale: "Please rate the difficulty of this ability from 1 (very easy) to 10 (very difficult)". Details on the adjustment can be found in the section below "adjustments to the original study". In contrast to the original study, domain difficulty ratings were scored on a similar scale as comparative ability, (own and others') comparative ability, desirability, and ambiguity.

We examined difficulty ratings across all domains to assess whether perceived difficulty was as expected in the original and conditions in which difficulty was manipulated. For the easy domain condition, we hypothesized that easy interpretations of the original domains would result in lower domain difficulty ratings across all abilities compared to ratings of the original domain group. For the difficult domain condition, we hypothesized that difficult interpretations of the original domains would result in higher domain difficulty ratings across all abilities compared to original domain group ratings. We expected the ambiguity ratings for both easy and difficult conditions to be lower than that in the original's domains. Additionally, we tested whether comparative ability would be influenced by our easy/difficult manipulations.[2]

1.7 Hypotheses

Based on the original study and the current extension hypotheses, this replication aims to test four central hypotheses (Table 3).

1.8 Adjustments to the original study

In the original study, the eight ability domains were divided into two categories: four easy domains and four difficult domains. On a 10-point scale from very easy to very difficult, easy domains had domain difficulty ratings below 5 (the midpoint of the scale), and difficult domains above 5, respectively. The above-average effect was tested for the easy domains, whereas the below-average effect was tested for the difficult domains.

Yet, several issues may arise from treating continuous variables as categorical. First, the categorization of continuous variables, especially dichotomization of placing variables into two groups, might lead to misclassifications, loss of information and power (Naggara

[2]Although this test was the reason for the preregistration, due to an error, neither hypotheses or tests related to the core questions of the extensions were part of the preregistration. Hence, analyses connected to this question in the extension will be treated as exploratory.

TABLE 3: Summary of the hypotheses.

Hypothesis	Statement	Variables	Conditions
H1	Compared to judgments of others' abilities, participant judgments of their own abilities better predict their comparative ability judgments.	Own absolute ability; others' absolute ability; comparative ability	Replication and extension conditions
(Original)			
H2	The more difficult the ability domain, the more likely a person is to shift from perceiving oneself as above average to perceiving oneself as below average.	Comparative ability; domain difficulty; desirability; ambiguity	Replication and extension conditions
(Original reframed)			
H3 (Extension)	Compared to the replication condition participants, the easy domain condition participants assign lower domain difficulty and ambiguity ratings to abilities.	Domain difficulty; ambiguity	Replication and easy domain conditions
H4 (Extension)	Compared to the replication condition, the difficult domain condition participants assign higher domain difficulty and lower ambiguity ratings to abilities.	Domain difficulty; ambiguity	Replication and difficult domain conditions

et al., 2011). Second, the loss of power by dichotomizing variables at the median is equal to discarding one-third of the data (Cohen, 1983; MacCallum et al., 2002). Third, variation between categorized groups may be underestimated as close response scores divided into different groups are defined as being very different instead of very similar. It has thus been suggested to keep variables continuous using methods such as linear regressions instead of t-tests (Altman & Royston, 2006).

For the above reasons, we did not assign ability domains to specific dichotomic easy/difficult categories. The above- and below-average-effects were tested on a continuous scale: instead of using one-sample t-tests, correlations were used to test the relationship between domain difficulty and comparative ability in three different ways: item-wise, compiled items in a vector (but not averaging across them), and row-wise averaged for the three conditions. Applying this method is a more direct assessment of perceived difficulty with the same sample. For a full overview of differences between the current and the original study see the OSF supplement, section "Comparisons and deviations".

1.9 Pre-registration and open science

Before data collection, the experiment was pre-registered (see the OSF supplement). Pre-registrations, power analyses, materials, data, exclusions, manipulations, power analyses, and other details and disclosures are available in the OSF supplement. Data collection was completed before analyses.

2 Method

2.1 Participants and power analyses

We conducted power analyses in R using the pwr package (Champely et al., 2018). The power analyses suggested a sample size of 160 to be sufficient for reaching 95% power with an alpha-level = .05 assuming an effect size of $f^2 = 0.099$ (informed by Kruger, 1999) for a 2-factor multiple linear regression analysis (see OSF supplement, section "Power analysis of original study effect to assess required sample for replication"). We tried to exceed this estimate (following replication recommendation such as Simonsohn, 2015) and added extensions thereby leading to the recruitment of 756 Amazon MTurkers. A total of 65 participants failed to meet the pre-registered inclusion criteria and were excluded, resulting in a total of 691 included participants (see Table1 in the OSF supplement for sample comparison and exclusion details).

2.2 Design

The original study used a within-subject design with one-sample analyses conducted for each condition (easy versus difficult domains), yet in the current replication, we used a 3 (between difficult conditions: original, easy, difficult) x 2 (within difficulty conditions: easy, difficult) mixed-design. All participants were presented with eight items (within-subjects; see Table2). We used the same methods as in the original study for within-group analyses and added additional analyses for the between-group comparisons (see the OSF supplement for more details and full measures).

2.3 Procedure

Participants were recruited through MTurk on TurkPrime/CloudResearch (Litman et al., 2017) and completed questionnaires via a provided "Qualtrics" link after giving consent. Participants were randomly assigned to one of three conditions: 1) Original domains (8 original domains; 4 easy and 4 difficult domains), 2) Easy domains extension (easy reinterpretations of the 8 original domains), or 3) Difficult domains extension (difficult reinterpretations of the 8 original domains).

TABLE 4: Comparison of original study and replication's samples.

	Kruger (1999)	MTurk sample (pre-exclusion)	MTurk sample (post-exclusion)
Sample size	37	756	691
Geographic origin	US American	US American	US American
Gender	8 males, 29 females	442 males, 307 females, 7 unspecified	397 males, 288 females, 6 unspecified
Medium (location)	Questionnaire (Cornell University)	Computer (online)	Computer (online)
Compensation	Course credit	Nominal payment	Nominal payment
Year	1999	2020	2020

Based on the categorization in the original study, of the eight ability domains, four were categorized as easy and the other four as difficult (see Table2), presented in randomized order.

2.4 Measures

The original study had six dependent variables and the current study added an additional dependent variable of perceived domain difficulty. Across all conditions, the dependent variables were measured as participant ratings for each of the eight ability domains (Table 2). We computed Cronbach's α-scores for the original and extension eight-item scales, first for all domains together, and then divided using the original's categorization of easy and difficult domains, being α_{all} >.63, α_{all} >.46, α_{all}>.47 (see the OSF supplement section "Reliability for domains across conditions").

2.5 Exclusion criteria

The following exclusion criteria were pre-registered: 1) low proficiency of English (less than 5 on a scale of 1 to 7); 2) not being serious (less than 4 on a scale of 1 to 5); 3) correctly guessing one of the hypotheses; 4) having seen or done the survey before; 5) failure to complete the survey; and 6) not in or from the United States, to keep sample characteristics as close to the original study as possible.

2.6 Evaluation criteria for replication findings

We compare the replication effects with the original effects in the target article using the criteria set by LeBel et al. (2019) (See the OSF supplement sections "Criteria for evaluation of replications" and "Replication evaluation").

We categorized the current replication as a "close replication" and provided details in Table5. Variables and questions were the same as in the original, with the addition of extensions and adjustments to fit the MTurk sample, instead of Cornell university students.

3 Results

We analyzed the data using R v3.6.3 (R Core Team, 2020), with analyses conducted both on a participant- and an item-level. To allow for a broader assessment of the data, we conducted preprocessing by both calculating mean scores (Table 6 for correlation matrices for each condition), and compiling the values for variables' eight items (abilities) in their raw form, resulting in 8 rows per participant (see "Correlations per condition" subsection in the OSF supplement for correlation matrices for each condition). For analyses conducted on an item level, participant ratings for each of the eight abilities were examined.

3.1 Domain difficulty comparisons by conditions

We conducted paired-sample Wilcoxon tests comparing difficulty *ratings* between the grouped 4 easy and 4 difficult replication/original and extension domain *items* and found domain difficulty ratings to be higher for difficult abilities across all comparisons (summarized in Table7, $ps < .001$), supporting Kruger's (1999) original categorization.[3] Hence, all conditions were analyzed as in the original study, including correlations between the variables across the eight domains, and one-sample Wilcoxon-tests testing for the above-average effect in easy ability domains and the below-average effect in difficult ability domains (Tables 8.1–8.3 in the OSF supplement).

3.2 Replication: original domain condition

We conducted all analyses in this section on the original domain condition ($n = 240$).

3.2.1 H_1: Relationship between absolute and comparative ability

In a linear regression model, own and others' absolute ability ratings predicted mean comparative ability judgments ($F(2, 237) = 323.9$, $p < .001$, $R_{adj}^2 = .73$, 95% CI [0.68, 0.79]).[4] However, we found support only for participants' judgments of their own absolute ability as predictors of their comparative ability judgments ($\beta = 0.90$, $t(239) = 19.93$, $p < .001$).

On an item level, we conducted multiple regressions for each of the eight abilities to examine how participants' estimates of both own and others' absolute abilities predict comparative ability estimates (see Table8 for standardized betas). Own absolute abilities

[3] T-statistics for the distinction of ability items into easy and difficult were not reported in Kruger (1999).

[4] See "Additional Tables and Figures" in the OSF supplement for regression plots and tables.

TABLE 5: Classification of the replication, based on LeBel et al. (2018).

Design facet	Replication	Details of deviation
IV operationalization	Same	
DV operationalization	Same	
IV stimuli	Similar, with an added extension	IV1 ability domains is changed from one condition of 4 easy and 4 difficult abilities, to 3 conditions of the replication group, the easy domain group, and the difficult domain group. Participants were presented with either the original ability domains, easy interpretations of the original ability domains, or difficult interpretations of the original ability domains.
DV stimuli	Similar, with an added extension	An additional dependent variable, DV1 (domain difficulty), is added.
		For DV2 (comparative ability judgment), the scale was changed from 0–99 to 0–100 for easier comprehension.
		For DV2 (comparative ability judgment) and DV4 (Judgmental weight of others' absolute abilities), the comparison group was changed from "other students from the course" to "other MTurk workers" to ensure applicability for all Mturk participants.
DV stimuli	Similar, with an added extension	For DV7 (experience in the ability domain), the scale used to measure prior experience was unspecified in the original study. Similar to the majority of other dependent variables, it is measured using a scale of 1 (no experience at all) to 10 (very experienced).
Procedural details	Similar, with an added extension	Participants are all assigned to the same condition in the original study. In the replication, they are randomly assigned to one of the three conditions.
Physical settings	Different	From a questionnaire to filling out an online Qualtrics survey.
Contextual variables	Different	From Cornell University undergraduates to American MTurk workers as participants.
Replication classification	Close replication	With two added extensions.

were generally better in explaining changes in comparative ability judgments than others' skills, which supports H_1.

TABLE 6: Mean ratings across all abilities for the three conditions.

	Original domains (n = 240)		Easy domains (n = 225)		Difficult domains (n = 226)	
Variable	Mean	SD	Mean	SD	Mean	SD
Mean domain difficulty	6.05	1.15	5.22	1.63	7.39	1.19
Mean comparative ability	53.29	14.5	58.86	14.92	46.97	18.86
Mean own absolute ability	6.04	1.37	6.64	1.44	4.77	2.02
Mean others' absolute ability	6.22	1.14	6.59	1.33	5.06	1.79
Mean desirability	8.15	0.99	7.9	1.15	7.54	1.35
Mean ambiguity*	3.00	1.24	2.68	1.23	2.76	1.43

* Ambiguity scores were reversed to indicate increasing ambiguity from 1 to 10.

TABLE 7: Asymptotic Wilcoxon-Mann-Whitney tests comparing perceived domain difficulty ratings between easy and difficult abilities (within conditions).

Condition	T-statistic	df	Mean difference	p-value	Effect size r	95% CI
Original (replication)	668.5	238	2.78	<.001	0.82	[0.79, 0.85]
Easy domain (extension)	1416	223	1.99	<.001	0.75	[0.69, 0.80]
Difficult domain (extension)	1917	224	1.22	<.001	0.69	[0.62, 0.75]

For the relationship between absolute and comparative ability ratings across all abilities (240 participants * 8 items), we found a strong relationship between comparative ability estimates and others' ability ratings ($r(6) = 0.94$, $p < .001$, 95% CI [0.71, .99]); and between comparative ability estimates and own ability ratings ($r(6) = 0.99$, $p < .001$, 95% CI [0.96, .99]). Hotelling's (1940) t indicated these correlations to be different from each other ($t(5) = 4.66$, $p = .006$).

3.2.2 H_1: Additional correlation analyses for the relationship between absolute and comparative ability

When adding two modes of analysis, namely, *vector-compiled scores* and *inventory mean scores*[5], Pearson's *rs*, calculated for *vector-compiled scores* of comparative ability estimates

[5] *Vector-compiled scores* were each participant (in the replication condition) scores in all 8 domains lined up in one vector with 8 (domains) * 240 (participants) = 1920 rows. *Inventory mean scores* were calculated by

TABLE 8: Replication condition: Mean comparative ability estimates and judgmental weight of own versus peers' abilities.

Ability	Domain difficulty1	Percentile estimate2	Judgment weight: Own ability3	Judgment weight: Others' ability3
Using mouse	2.70 (2.63)	71.2*** (17.90)	0.29***	0.04
Driving	5.19 (2.41)	65.2*** (22.08)	0.85***	–0.11**
Riding bicycle	4.14 (2.44)	61.0*** (20.48)	0.76***	–0.06
Saving money	6.63 (2.08)	62.9*** (21.10)	0.79***	–0.05
Telling jokes	6.10 (2.06)	52.4 (22.63)	0.75***	0.04
Playing chess	7.74 (1.75)	41.0*** (27.00)	0.82***	–0.03
Juggling	7.64 (1.97)	32.0*** (27.67)	0.59***	0.18**
Programming	8.29 (1.74)	40.7*** (29.22)	0.83***	–0.06

Note: Tablepresented as in original study (Kruger, 1999, Table 2) encompassing descriptive statistics, one-sample t-tests, and regressions.

[1] Mean (SD) scores for item-wise domain difficulty. Higher numbers reflect greater difficulty.

[2]Mean (SD) scores for item-wise comparative ability/percentile estimates. Scores above 50 reflect an above-average effect, estimates below 50 reflect a below-average effect. See supplementary tables 8.1 and 9.1 for test statistics and CI's.

[3]Standardised betas from multiple regressions predicting participants' comparative ability (percentile) estimates from own absolute ability and peers' absolute ability, respectively.

$^{**}p < .01$. $^{***}p < .001$.

and other's absolute ability, were $r(1918) = 0.50$ (95% CI [0.46, 0.53]); and between comparative ability estimates and own absolute ability were $r(1918) = 0.81$ (95 % CI [0.79, 0.82]); with these correlations being different from each other (Hotelling's (1940) $t(1917) = 27.61$, $p < 0.001$). For *inventory mean scores*, correlations between comparative ability estimates and other's absolute ability were $r(238) = 0.53$ ($p < .001$, 95% CI [0.43, 0.62]); and between own and comparative ability $r(238) = 0.85$ ($p < .001$, 95% CI [0.82, 0.89]); with these correlations being different from each other (Hotelling's $t(237) = 11.75$, $p < 0.001$).

However, when using a mixed-effects model with random intercepts at the level of participants to explain comparative ability, positive changes in own ability explained positive changes in comparative ability and the relationship between others' and comparative ability being the opposite (Table 9). The findings from both replicated and the new analyses present strong support for H_1.

averaging the 8 domains for each participant (row-wise), resulting in 240 rows. P-values for vector-compiled scores correlations are not provided as those do not account for repeated responses of the same person.

TABLE 9: Estimated fixed-effects coefficients of the mixed-effects regression model with changes in Comparative Ability explained by Others' and Own Ability.

Predictors	B	S.E.	CI	p
(Intercept)	12.56	1.33	[9.95, 15.18]	< 0.001
Own Ability	7.18	0.16	[6.86, 7.50]	< 0.001
Others' Ability	–0.42	0.21	[–0.84, –0.01]	0.04

Note. The table presents the fixed-effects coefficients with all the model predictors. See supplementary section "Mixed Models" for step-wise regression results.

3.2.3 H_2: Relationship between comparative ability, domain difficulty, and desirability.

We conducted one-sample t-tests to examine domain-wise comparative ability ratings using the 50th percentile estimates of comparative ability to classify above and below average effects (as in Kruger, 1999). Similar as in Kruger's (1999) findings, participants indicated to be above-average for all easy ability domains (*ps* < .001) and below-average for three of the four difficult ability domains (*ps* < .001; see Table 8 column 2 for descriptive statistics, and tables 8.1 and 9.1 in the OSF supplement for test statistics and CI's). For the above and below-average effects across all abilities, we found a strong negative correlation between comparative ability estimates and domain difficulty ($r(6) = –0.85$, $p = .0073$, 95% CI [–0.97 -0.37]).[6] Item-wise comparative-ability-domain-difficulty correlations are provided in the supplementary under 'Replication condition: Item-wise correlations between domain difficulty and comparative ability ratings for each ability domain'.

When comparing desirability ratings between easy ($M = 8.731$, $SD = 1.01$) and difficult ability domains ($M = 7.58$, $SD = 1.40$), a paired-samples Wilcoxon test revealed easy abilities to be more desirable ($M_{\text{difference}} = 1.16$, $Z(238) = 9.42$, $p < .001$, $r = 0.66$, 95% CI [0.59, 0.73]). One-sample Wilcoxon tests revealed that all domain-specific desirability scores were higher than the scale midpoint (*ps* < .001; supplementary Table 9.4). That corresponded with a strong positive relationship between comparative ability and desirability ($r(6) = 0.72$, $p = .0448$, 95% CI [0.03, 0.95]).

3.2.4 H_2: Additional Analyses for the relationship between comparative ability, domain difficulty, and desirability.

Similarly, we found a negative association between comparative ability and domain difficulty ratings when using *vector-compiled scores* ($r(1918) = –0.35$, 95% CI [–0.39, –0.31]).[7]

[6]See Table 11 in the OSF supplement: equivalence tests 1–2.

[7]See Table 11 in the OSF supplement: equivalence tests 2–3. The presented correlation on vector compiled scores is not an optimal measure as these do not account for dependence in several measures provided by the same individual. Hence, p-values are not informative and therefore not reported.

However, when using *inventory mean scores*, opposite to the original study, we found a positive association between comparative ability and mean domain difficulty ratings ($r(238) = 0.16$, $p = .013$, 95% CI [0.04, 0.28]).[8] As this *inventory mean scores* correlation did not correspond to the other results, we conducted an exploratory analysis[9], revealing a small positive correlation between comparative ability and domain difficulty ratings in easy ($r(238)= 0.03$, 95% CI [–0.10, 0.15], $p = .70$); and a small negative correlation in difficult ability domains ($r(238)= –0.10$, 95% CI [–0.23, 0.02], $p = .11$). Using mixed models with random intercepts at the participant level, H_2 was not supported as difficulty did not predict changes in comparative ability (Table 10).

TABLE 10: Estimated fixed-effects coefficients of the mixed-effects regression model with changes in Comparative Ability explained by Others' and Own Ability in the Replication Condition.

Predictors	B	S.E.	CI	p
(Intercept)	8.72	2.4	[4.01, 13.43]	< .001
Own	7.07	0.18	[6.73, 7.42]	< .001
Other	–0.48	0.21	[–0.90, –0.06]	0.025
Difficulty	–0.04	0.16	[–0.36, 0.28]	0.817
Desirability	0.57	0.21	[0.16, 0.99]	0.007
Ambiguity	0.12	0.17	[–0.21, 0.45]	0.48

Note. The table presents the fixed-effects coefficients with all the model predictors. See supplementary section "Mixed Models" for step-wise regression results.

The original analysis' methods provided support for H_2. Additionally, a Simpson's paradox can be observed when averaging all eight domains into one score over various manipulated factors for each participant and then correlating them.[10]

3.3 Extension: Easy domain and difficult domain conditions

3.3.1 Comparative ability for easy and difficult items by conditions

We conducted paired-sample Wilcoxon tests comparing difficulty ratings between the easy and difficult replication/original and extension domains and found comparative ability to be estimated higher for easy abilities across all comparisons (summarized in Table 7, all $p <$ .001).

[8]See Table 11 in the OSF supplement: equivalence tests 4–5.

[9]Not included in the preregistration.

[10]For an overview of all correlations between mean scores across inventories for the replication condition see Tables 3.1 and 3.2 in the OSF supplement.

3.3.2 Relationship between absolute and comparative ability

We conducted multiple linear regression analyses to test how ratings of both own and others' ability predict comparative ability judgments across all abilities. Models in both conditions predicted variance in comparative ability judgments ($F_{easy}(2, 222) = 246.6$, $p < .001$, $R_{adj}^2 = .69$, 95% CI [0.62, 0.76]; and $F_{difficult}(2, 223) = 342.9$, $p<.001$, $R_{adj}^2 = .75$, 95% CI [0.70, 0.81]). Yet, the only significant predictors of participants' own absolute ability were comparative ability judgments in both the easy ($\beta = 0.86$, $t(222) = 17.32$, $p < .001$) and the difficult domain condition ($\beta = 0.90$, $t(223) = 15.61$, $p < .001$).

TABLE 11: Extension conditions: Mean comparative ability estimates and judgmental weight of own and peers' abilities by domain difficulty.

	Easy domain condition		Difficult domain condition	
Ability	Judgmental weight of own ability[1]	Judgmental weight of others' ability[1]	Judgmental weight of own ability[1]	Judgmental weight of others' ability[1]
Using mouse	0.48***	0.03	0.58***	0.15*
Driving	0.75***	–0.1	0.78***	–0.02
Riding bicycle	0.65***	0.06	0.79***	0.06
Saving money	0.81***	–0.03	0.78***	–0.07
Telling jokes	0.70***	0.10*	0.70***	0.14**
Playing chess	0.79***	0.02	0.75***	0.01
Juggling	0.78***	0.05	0.68***	0.05
Programming	0.85***	–0.03	0.79***	0.03

[1] Standardised betas (β) from multiple regressions predicting participants' comparative ability (percentile) estimates from their estimates of their own absolute ability and the absolute ability of their peers, respectively.

*$p < .05$. **$p < .01$. ***$p < .001$.

Item-wise multiple linear regression analyses showed, consistent with the original study and replication condition, that extension condition participants weighted own ability estimates stronger than others' ability estimates when assessing their comparative abilities (Table 11). All standardized betas (β) of own absolute abilities were positive and *ps* <.001 (for all abilities), whereas βs of others' absolute abilities were bi-directional and smaller.

For the *easy* domain condition, the correlation between own ability and comparative ability was $r(6) = 0.99$ ($p < .001$, 95% CI [0.97, 0.999]); and the correlation between others' and comparative ability was $r(6) = 0.96$ ($p < .001$, 95% CI [0.78, 0.99]); and these correlations were different from each other (Hotelling's (1940) $t(5) = 2.85$, $p = 0.037$). For the *difficult* domain condition, the correlation between own ability and comparative ability was $r(6) = 0.97$ ($p < .001$, 95% CI [0.85, 0.995]); and the correlation between others' and

comparative ability was $r(6) = 0.92$ ($p = .001$, 95% CI [0.60, 0.99]); with weaker support found for these correlations as being different from each other (Hotelling's $t(5) = 2.24$, $p = 0.075$).

3.3.3 Additional Analyses: Relationship between absolute and comparative ability

The vector-compiled score correlation for the *easy* domain condition between own and comparative ability was $r(1798) = 0.78$ (95% CI [0.76, 0.80]); and between others' and comparative ability was $r(1798) = 0.47$ (95% CI [0.43, 0.51]). For the *difficult* domain condition correlations between own and comparative ability was $r(1806) = 0.78$ (95% CI [0.76, 0.80]); and between others' and comparative ability was $r(1806) = 0.45$ (95% CI [0.41, 0.48]).

Additionally, also mixed models indicated that own ability was a better predictor of comparative ability than others' ability (Table 12).

TABLE 12: Estimated fixed-effects coefficients of the mixed-effects regression model with changes in Comparative Ability explained by Others' and Own Ability in the Extension Conditions.

Predictors	B	S.E.	CI	p
Easy condition extension				
(Intercept)	15.48	1.3	[12.94, 18.02]	<0.001
Own	6.56	0.16	[6.25, 6.88]	<0.001
Other	–0.04	0.2	[–0.43, 0.36]	0.861
Difficult condition extension				
(Intercept)	16.53	1.49	[13.61, 19.44]	<0.001
Own	6.4	0.16	[6.09, 6.72]	<0.001
Other	–0.02	0.22	[–0.44, 0.41]	0.94

Note. Fixed-effects coefficients with all model predictors. Participants represented the random effect. See supplementary section "Mixed Models" for step-wise regression results.

Inventory mean score correlations for the *easy* domain condition between own and comparative ability was $r(223) = 0.83$ ($p < .001$, 95% CI [0.78, 0.87]); and between others' and comparative ability was $r(223) = 0.52$ ($p < .001$, 95% CI [0.42, 0.61]). In the *difficult* domain condition the correlation between own and comparative ability was $r(224) = 0.87$ ($p < .001$, 95% CI [0.83, 0.90]); and between others' and comparative ability $r(224) = 0.70$ ($p < .001$, 95% CI [0.62, 0.76]).

TABLE 13: Extensions: Mean domain difficulty and mean comparative ability estimates tested against the average (scale midpoint).

Ability	Easy domain condition		Difficult domain condition	
	Domain difficulty	Percentile estimate[1]	Domain difficulty	Percentile estimate[2]
Using mouse	3.13 (2.90)	71.27 (20.51)***	5.77 (2.36)	55.79 (21.12)***
Driving	4.58 (2.77)	66.32 (22.74)***	7.16 (2.15)	40.63 (29.28)***
Riding bicycle	3.88 (2.77)	65.76 (21.92)***	7.85 (2.12)	48.90 (27.54)
Saving money	5.31 (2.76)	63.63 (25.55)***	6.32 (2.60)	62.68 (25.77)***
Telling jokes	4.64 (2.67)	59.74 (20.55)***	7.67 (1.98)	40.82 (27.03)***
Playing chess	6.71 (2.56)	47.81 (27.55)	8.34 (1.89)	41.86 (27.22)***
Juggling	6.07 (2.60)	46.76 (27.75)	7.81 (2.00)	39.67 (27.66)***
Programming	7.46 (2.13)	49.57 (25.71)	8.15 (1.84)	45.36 (26.02)**

*p<.05, **p<.01, ***p<.001.

Note: Scores are displayed with the following structure: Mean (SD).

[1] Scores above 50 reflect an above-average effect, estimates below 50 reflect a below-average effect. See Table 9.2 in supplementary for test statistics and CI's.

[2] See Table 9.2 in supplementary for test statistics and CI's.

3.3.4 Relationship between domain difficulty and comparative ability.

As indicated above, one-sample t-tests indicated above-average-effect for the easy and below-average effect for the difficult condition (Table 13 for mean scores and SD's,and Tables 9.2–9.3 in the OSF supplement for test statistics). However, the below-average-effect was not expressed in the easy extension condition, and the above-average-effect was not clearly expressed in the difficult extension condition. Item-wise correlations between comparative ability and domain difficulty for each ability are provided in the OSF supplement under 'Extension conditions: correlations between comparative ability and domain difficulty ratings for each ability domain'. The easy domain condition contains mixed results of medium to no associations (p <.936), whereas the difficult domain condition contains negative associations for all abilities (p <.001). Congruent with original and replication findings, there were negative relationships between domain difficulty and comparative ability in the *easy* $r(6) = –0.90$ (p = .002, 95% CI [–0.982, –0.537])[11]; and *difficult conditions* ($r(6) = -0.75$, p = .033, 95% CI [–0.951, –0.092]).[12]

[11] See OSF supplement: equivalence tests 7–8.

[12] See OSF supplement: equivalence tests 9–10.

3.3.5 Additional analyses for the relationship between domain difficulty and comparative ability.

Congruent with both original and replication findings, correlations between comparative ability and mean domain difficulty were negative for *vector-compiled score* in the easy ($r(1798) = –0.27$, 95% CI [–0.31, –0.22]) and difficult ($r(1798) = –0.31$, 95% CI [–0.35, –0.27]) conditions. When averaging across the inventory (*inventory mean scores*), this relationship changes to $r(223) = 0.32$ ($p < .001$, 95% CI [.19, .43]) in the easy condition and $r(223) = –0.13$ ($p = .0498$, 95% CI [–0.26, –0.0002]) in the difficult condition – showing the possibility of a Simpson's paradox, just as in the replication condition.[13] Different from the replication data, in both easy and difficult conditions, with decreasing difficulty, comparative ability increases (Table 14).

TABLE 14: Estimated fixed-effects coefficients of the mixed-effects regression model with changes in Comparative Ability explained by Others' and Own Ability in the Extension Conditions.

Predictors	B	S.E.	CI	p
Comparative Ability Easy Condition				
(Intercept)	18.6	2.35	[13.99, 23.21]	<0.001
Own	6.37	0.18	[6.02, 6.71]	<0.001
Other	–0.13	0.21	[–0.54, 0.28]	0.546
Difficulty	–0.41	0.16	[–0.72, –0.11]	0.008
Desirability	0.15	0.21	[–0.26, 0.56]	0.468
Ambiguity	–0.1	0.19	[–0.47, 0.27]	0.6
Comparative Ability Difficult Condition				
(Intercept)	25.57	2.66	[20.36, 30.79]	<0.001
Own	6.11	0.17	[5.78, 6.45]	<0.001
Other	–0.16	0.22	[–0.59, 0.26]	0.451
Difficulty	–1.04	0.2	[–1.43, –0.64]	<0.001
Desirability	0.16	0.2	[–0.22, 0.55]	0.405
Ambiguity	–0.17	0.19	[–0.55, 0.21]	0.37

Note. The table presents the fixed-effects coefficients with all the model predictors. Participants represented the random effect. See supplementary section "Mixed Models" for step-wise regression results.

[13] See OSF supplement Tables 5.1, 5.2, 7.1 and 7.2 for correlations between mean scores across inventories in the extension conditions.

3.3.6 Comparisons of ambiguity and difficulty ratings between the three conditions

As parametric assumptions were not met[14], to test whether different domain definitions from the original domains would result in different domain difficulty and ambiguity ratings, we first conducted a Kruskal-Wallis test that showed differences in difficulty scores across conditions ($H(2) = 237$, $p < .001$, $\eta^2 = 0.34$; Figure 1). Supporting the first part of $H_{3–4}$, post-hoc Bonferroni corrected Mann-Whitney tests showed that, compared to the replication condition ($Mdn_{replication} = 6.00$, $M_{\text{replication}} = 6.05$, $SD = 1.15$), participants in the easy domain condition ($Mdn_{\text{easy}} = 5.00$, $M_{\text{easy}} = 5.22$, $SD = 1.63$) rated lower domain difficulty ($p < .001$). Participants in the difficult domain condition ($Mdn_{difficult} = 7.78$, $M_{difficult} = 7.39$, $SD = 1.19$) rated higher domain difficulty than in the other conditions ($ps < .001$; Figure 1A). We conducted a second Kruskal-Wallis test and found differences in participants' ambiguity ratings between the three conditions ($H(2) = 11.47$, $p = .003$, $\eta^2 = 0.014$; Figure 1B). As predicted in the second part of $H_{3–4}$, post-hoc Bonferroni corrected Mann-Whitney tests showed replication condition ambiguity ratings ($Mdn_{replication} = 2.88$ $M_{replication} = 3.00$, $SD = 1.24$) to be lower than both the easy extension condition ($Mdn_{easy} = 2.38$, $M_{easy} = 2.68$, $SD = 1.23$; $p_{adj} = 0.01$) and the difficult extension condition ambiguity ratings ($Mdn_{difficult} = 2.38$, $M_{difficult} = 2.76$, $SD = 1.43$; $p_{adj} = 0.01$). We found no support for differences between easy and difficult extension conditions' ambiguity ratings, ($p_{adj} \approx 1.00$).

3.3.7 Relationship between comparative ability, and domain difficulty and desirability (examining H_2 in the extension conditions)

In the following section, the easy ($n = 225$) and difficult ($n = 226$) extension conditions results are analyzed in the same way as reported above for the replication condition. For the above- and below-average effects across all abilities, we found a strong negative correlation between comparative ability estimates and domain difficulty in both extension conditions (see above). Item-wise comparative-ability–domain-difficulty correlations are provided in the OSF supplement 'Extension conditions: correlations between comparative ability and domain difficulty ratings for each ability domain'.

When comparing desirability ratings between easy and difficult ability domains via Wilcoxon signed ranks test, in the easy extension condition easy ($M = 4.23$, $SD = 2.13$) abilities to be more desirable than difficult abilities ($M = 6.22$, $SD = 1.56$; $Z(223) = –10.62$, $p < .001$, $r = 0.75$, 95% CI [0.70, 0.80]), as well as in the difficult extension condition easy abilities (M = 6.78, SD = 1.44), difficult ($M = 7.99$, $SD = 1.30$; $Z(224) = –9.26$, $p < .001$, $r = 0.69$, 95% CI [0.62, 0.75]). One-sample Wilcoxon tests revealed that all domain-specific desirability scores were higher than the scale midpoint ($ps < .001$; OSF supplement Tables 9.5–9-6). Moreover, correlations between comparative ability and desirability in easy ($r(6)$

[14]See "Statistical assumptions and normality Tests" section in the detailed supplementary on OSF for parametric tests.

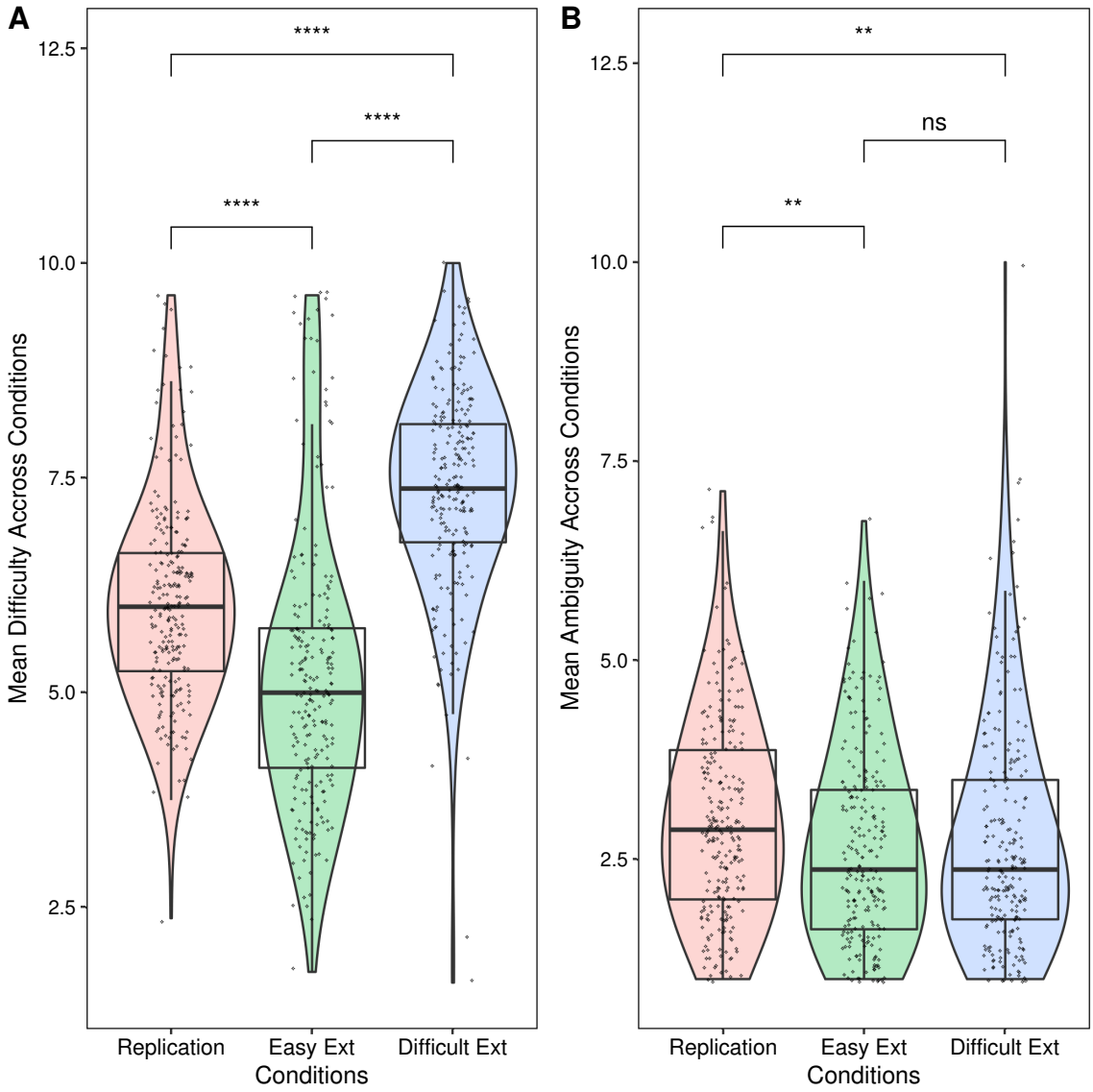


FIGURE 1: Box and violin plots of domain difficulty and ambiguity ratings across replication, easy extension, and difficult extension conditions with uncorrected p-values for group-wise comparisons and overall models. Panel A: Mean difficulty across conditions. Panel B: Mean ambiguity across conditions. $^{ns}p>.05$, $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$, $^{****}p<.0001$.

= 0.66, p = .074, 95% CI [–0.08, 0.93]) and difficult extension conditions ($r(6)$ = 0.15, p =.72, 95% CI [–0.62, 0.77]) remain uncertain.

3.3.8 Extension H_2: Additional Analyses for the relationship between comparative ability, and domain difficulty and desirability

Similarly, we found a negative association between comparative ability and domain difficulty ratings when using *vector-compiled scores* in the easy extension condition ($r(1798) = –0.27$,

95% CI [–0.31, –0.22])[15] as well as in the difficult extension condition (r(1806) = -0.31, 95% CI [–0.35, –0.27])[16]. Similar to our findings for the replication condition, when using *inventory mean scores*, we found a positive association between comparative ability and mean domain difficulty ratings in the easy extension condition (r(223) = 0.32, p < .001, 95% CI [0.19, 0.43])[17] and a negative association in the difficult extension condition (r(223) = –0.13, p = .05, 95% CI [–0.26, –0.0002])[18].

3.3.9 Exploratory Analysis: comparative ability across conditions

In an exploratory analysis using a 3 (Condition) x 2 (Difficulty) mixed design, an aligned rank-transform nonparametric factorial ANOVA showed both main effects of condition (F(2, 1376) = 47.03, p < .0001, η^2_G = 0.064) and difficulty (F(1, 1376) = 302.17, p < .0001, η^2_G = 0.169), as well as the interaction effect (F(1, 1376) = 15.23, p < .0001, η^2_G = 0.022), were significant.[19]

Post-hoc multiple comparisons revealed significant differences between all comparisons at Bonferroni corrected *ps* <.001, except the comparison between easy items in replication compared to easy items in the easy extension, difficult items in the replication compared to difficult extension, and difficult easy-extension compared to easy difficult-extension (as expected from power-simulations), with *ps* ≈ 1.00.

3.4 Replication Evaluation

The following section compares the original study and current replication based on the replication evaluation criteria by LeBel et al. (2019). We found clear support for replication hypotheses H_1 and H_2. Both correlations between own absolute ability and comparative ability across all abilities displayed as conducted in the original study and additional analyses detected strong effects in the same direction as the original, but we found no support for difficulty as a predictor of comparative ability in a mixed-effects model using the replication data (Table 15). Positive and significant standardized betas for all own absolute abilities, and predominantly negative and non-significant standardized betas for others' absolute abilities were replicated (Table 16). The strong evidence bolsters Kruger's research on egocentrism as comparative ability judgments are based on participants' own levels of ability instead of their perceptions of others' level of ability (Kruger, 1999; Kruger & Burrus, 2004). An underlying mechanism might be focalism, a complementary bias on people's tendency to place more judgmental weight on the target (self) and less weight on the referent (others)

[15]See OSF supplement: equivalence tests 11–12.

[16]See OSF supplement: equivalence tests 13–14.

[17]See OSF supplement Table 11: equivalence tests 15–16.

[18]See OSF supplement Table 11: equivalence tests 17–18.

[19]As this analysis was an oversight in our preregistration, an additional power simulation was executed, showing excellent power for observing main and interaction effects of a 3x2 mixed ANOVA. See OSF supplement "Power Simulation for Exploratory Analysis" for more information.

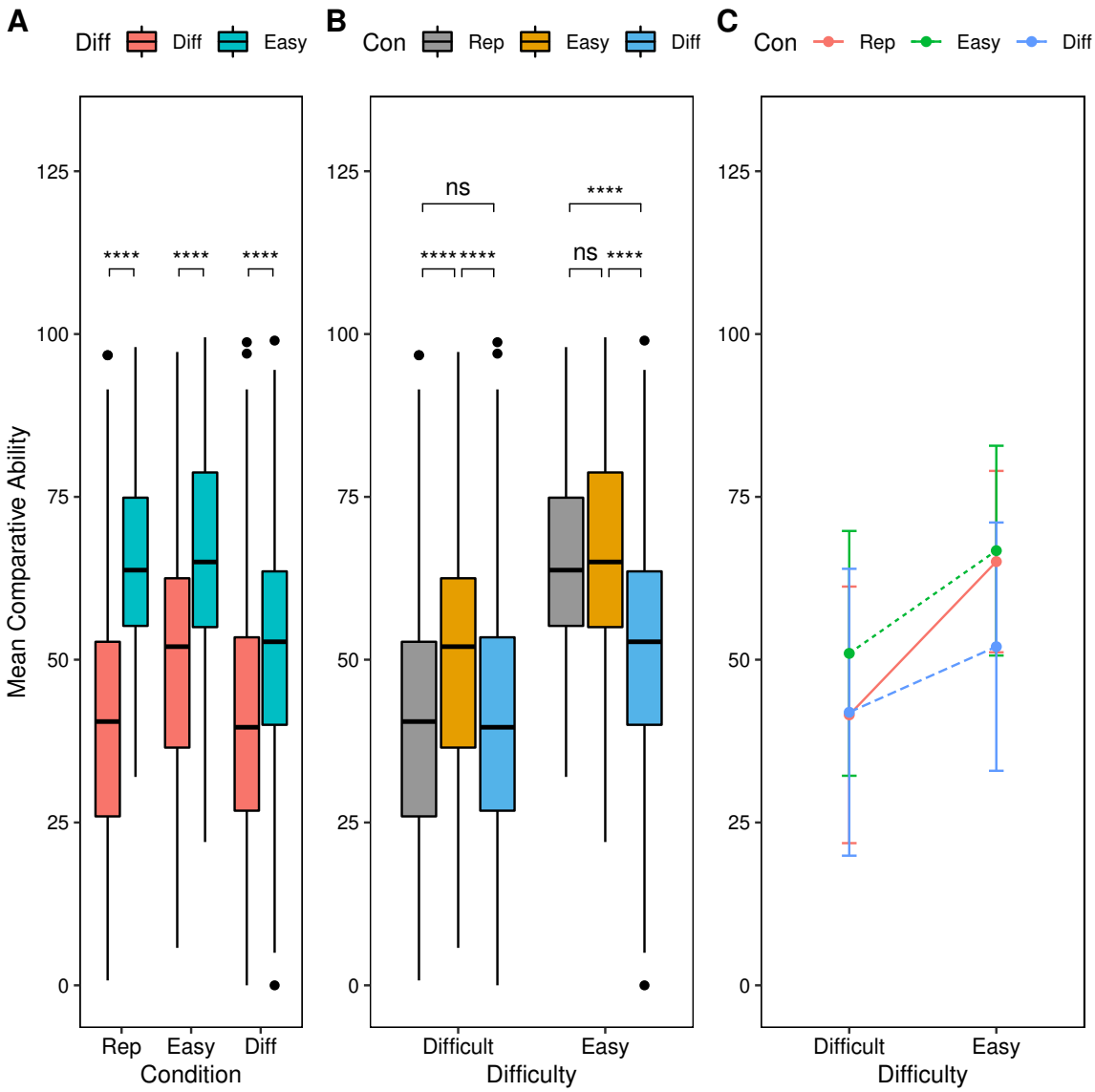


FIGURE 2: Comparative ability across conditions. Panel A. Mean easy and difficult mean comparative ability ratings by condition. Panel B. Mean comparative ability ratings by difficulty. Panel C. Mean easy and difficult mean comparative ability ratings by condition with SD. $^{ns}p>.05$, $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$, $^{****}p<.0001$.

when making direct comparisons between the two (Krizan & Suls, 2008). An alternative explanation is that people simply have more information about themselves than they do about others. Paired with expectations about distributions of values of luck and skills, participants might have rationally judged, based on their best guess, that their own abilities are higher compared to others' abilities when tasks were easy and vice versa when tasks were difficult (Moore & Healy, 2008).

Above and below-average effects (H_2) replicated with a slightly smaller effect. Additional analyses revealed a smaller effect in the same direction, but when averaging the

TABLE 15: Comparison of correlational study effect sizes between the original article and replication based on criteria created by LeBel et al. (2019).

Variables (across all abilities)	p Original study	Correlation coefficient (r) and 95% CI	p Replication condition	Correlation coefficient (r) and 95% CI	Replication evaluation
Own ability and comparative ability	<.001	r(6) = .95 [0.90, 0.97]	<.001	r(6) = 0.99, [0.96, 1.00]	Signal-consistent
Inventory mean and absolute own ability and comparative ability	/	/	<.001; <.001	r(238) = .85 [0.82, 0.89]; r(1918) = 0.50, [0.46, 0.53]; Own (B = 7.18) vs others' ability (B = –0.42)	Additional analyses
Domain difficulty and comparative ability	<.001	r(6) = –.96, [–0.98, –0.92]	0.007	r(6) = –0.85, [–0.97 -0.37]	Signal-consistent, smaller
Inventory mean and absolute domain difficulty and comparative ability	/	/	.013; <.001	r(238) = 0.16 [0.04, 0.28]; r(1918) = –0.35, [–0.39, –0.31]; Difficulty as predictor of comparative ability B = -0.04	Additional analyses

entire inventory for each participant and thereby reducing the variability in responses, a Simpson's paradox seems to occur. Additionally, we found no support for difficulty as a predictor of comparative ability in a mixed regression model using the replication data, but we found support in both extensions. Participants tended to indicate higher rather than lower comparative ability in both the replication and the easy conditions, where difficulty ratings were normally distributed. This was not the case for the difficult condition, where

TABLE 16: Comparison of mean comparative ability estimates and judgmental weight of own versus others' abilities by domain difficulty between the original study and replication condition.

	Original study		Replication condition		Replication outcome
Ability	Judgmental weight of own ability1	Judgmental weight of others' ability1	Judgmental weight of own ability1	Judgmental weight of others' ability1	
Using mouse	0.21	0.06	0.29***	0.04	Replicated, own absolute abilities are all positive (same direction) and significant (all p <.001)
Driving	.89****	–.25*	0.85***	–0.11**	
Riding bicycle	.61****	–0.02	0.76***	–0.06	
Saving money	.90****	–.25***	0.79***	–0.05	
Telling jokes	.91****	–0.03	0.75***	0.04	
Playing chess	.96****	–.22**	0.82***	–0.03	
Juggling	.89****	–0.16	0.59***	0.18**	
Programming	.85****	–0.1	0.83***	–0.06	

Note. The original study only provided the standardized betas and *p*-values. The transformed R^2 and F^2 values would only represent the effect size of one predictor instead of the overall regression, so only the *p*-values and directions were compared.

[1]Standardised betas from multiple regressions predicting participants' comparative ability (percentile) estimates from their estimates of their own absolute ability and the absolute ability of their peers, respectively.

*$p < .05$. **$p < .01$. ***$p < .001$. ****$p < .0001$.

difficulty ratings were right-skewed. In other words, the Simpson paradox was produced by the above-average-effect being stronger than the below-average-effect in the replication and the easy conditions. Overall, this shows the contextual effects of the inventory's difficulty on participants' ratings of tasks difficulty and comparative ability. Using both one-sample Wilcoxon and t-tests, both above-and-below-average effects replicated with smaller effects, whereas above-average effect sizes replicated closer to the original study (Table 17). Despite smaller effect sizes, the observed results support above-and-below-average effects. The prevalence of the below-average-effect also demonstrates that motivated reasoning to

see oneself as superior fails to account for certain situations, such as for difficult abilities in the replication.

TABLE 17: Comparison of one-sample t-test effect sizes between the original article and replication based on criteria created by LeBel et al. (2019).

	Cohen's d and 95% CI	Replication outcome
Original study (n=37)		
Each easy ability	0.90 [0.22, 1.57]	
Each difficult ability (excluding telling jokes)	–1.44 [–2.17, –0.72]	
Replication condition (n=240)		
Easy abilities		
Using mouse	1.18 [1.02, 1.35]	Signal-consistent
Driving	0.69 [0.55, 0.83]	Signal-consistent, smaller
Riding bicycle	0.54 [0.40, 0.67]	Signal-consistent, smaller
Saving money	0.61 [0.47, 0.75]	Signal-consistent, smaller
Difficult abilities		
Telling jokes	0.11 [–0.02, 0.23]	No signal
Playing chess	–0.33 [–0.46, –0.20]	Signal-consistent, smaller
Juggling	–0.65 [–0.79, –0.51]	Signal-consistent, smaller
Programming	–0.32 [–0.45, –0.19]	Signal-consistent, smaller

4 Discussion

We replicated and extended the findings in Kruger's (1999) Study 1. Both the replication and the extension results provide strong support for above- and below-average effects, depending on difficulty. In addition, we present important boundary conditions. *First*, above-and-below-average effects appear stronger the more difficult the domain abilities are (compare Tables 8 and 11). *Second*, the difficulty of different activities (ability domains) might provoke or suppress below -or above-average-effects; we observed a below-average-effect when the presented abilities were difficult, and vice versa, an above-average effect when the presented abilities were easy. In that context, we observed an interaction effect between manipulations (making the original scale easier or more difficult) and item-group difficulty (easy vs difficult items), looking at comparative ability. Ambiguity was low across conditions with additional information introduced in the extensions decreasing ambiguity.

4.1 Replication outcomes

Egocentrism is a compelling, yet only one of many explanations for above-and-below-average-effects (Zell et al., 2020). Alternatively, judgments might be rationally based on differential access to information influencing predictions (Moore & Small, 2007). In other words, by having more information about the own than others' performance in different activities, others' performance is evaluated less extremely than the own performance (Moore & Healy, 2008).

Moreover, the replication advances our understanding of the conditions in which the above or below-average effects are more pronounced, i.e., when abilities' difficulty and supplied information about them differ. It complements a recent meta-analysis on the above-average-effect (Zell et al., 2020), showing a larger effect when using the direct (compare oneself to others on a single scale with the midpoint defined as average) rather than indirect testing method (assess oneself and the comparison group independent from each other, with the average being defined as the difference between the two values). Fewer research center on the below-average-effect, yet success in replicating the effect suggest that the same conditions may also be applicable in strengthening the below-average effect.

On the other hand, the replication's smaller effect sizes challenge the influence of certain established factors on the effects. For instance, people showed the strongest biases in comparative ability judgments when the comparison group was abstract instead of concrete, and no specific information and contact with the comparison group contributes to that abstractness (Alicke et al., 1995).

A notable discrepancy between the original and replication is the comparison group: original study participants compared themselves to other students from their psychology course, which was much more concrete than replication participants comparing themselves to others of the same age, gender, and socioeconomic background. The replication's smaller effects suggest that in contrast to past explanations, people may not display tendencies to choose vulnerable comparison targets to compare themselves with when given an abstract referent group (Chambers & Windschitl, 2004). As people display preferences in selecting representative targets, they might choose comparison targets of varying ability depending on task difficulty, and the availability of information and cognitive resources (Nisbett et al., 1983). This may have been the case for the current replication and is a promising direction for future research.

4.2 Outcomes of the extensions to the original study

Both H_3 and H_4 were supported. We found lower domain difficulty ratings in the easy domain condition than the replication condition (d = 0.59) and higher domain difficulty ratings in the difficult domain condition than the replication condition (d = 1.15) supporting the first part of the extension hypotheses ($H_{3–4}$) on differences in domain difficulty. As interpretations of easy or difficult abilities contribute to different perceptions of domain

difficulty, the observed results provide insight on how this affects participant interpretation of "average" ability. In a study by Kim et al. (2017), people construed below-median averages and showed above-average effects for abilities perceived as easy, and construed averages at or above the median for abilities perceived as difficult. For accurate assessments of comparative ability judgments, researchers not only need to ascertain how people interpret "average" ability, but also place efforts in lowering variations in the perceived difficulty of abilities. Hence, the original domain definitions may have been open to interpretation, influencing the results.

Moreover, we found support for the second part of $H_{3–4}$, that ambiguity was lower in the replication conditions. Eventually, more information provided might have led to clarification and hence decreased perceptions of ambiguity. Previous research showed a tendency to view oneself as above-average for ambiguous abilities (Dunning et al., 1989), and to select favorable, self-serving definitions amongst ambiguous traits describing a wide variety of behaviors (Gilovich, 1983; Kunda, 1987), which could not be reflected from our data. Finally, comparing comparative ability scores across conditions (replication vs extensions) and by the difficulty of the items (easy vs difficult), show an interaction effect. That indicates that both domain difficulty and ambiguity might influence comparative ability ratings and thereby above-and-below-average-effects. However, despite the presented extensions potentially presenting the influence of abilities' difficulty and their definitions' ambiguity on the effects, more research is needed to address above-and-below-average-effects' boundary conditions.

4.3 Limitations and future directions

Deviating from the original study, in our replication we measured the continuous relationship between variables and analyzed data on participant and item levels. Moreover, possible inferences from comparisons between added and original study correlations between domain difficulty and comparative ability are limited. Our tests supported original ability categorizations as easy or difficult, all original study tests (including one-sample tests and correlations of ratings across all abilities) were also carried out for the replication condition. While we recommend future replications testing the continuous relationship between variables to avoid limitations in performing study comparisons, misclassification, and issues in categorizing continuous variables, we also caution of low reliability when using the presented scale and particularly the suggested (easy and difficult ability) subscales (Table 5).

Furthermore, the replication's ability domain definitions are all based on Kruger's (1999) original domains. Yet, these domains may not be as accurate and widely applicable at present. For example, a recent survey indicated that the easy ability "saving money" is challenging for the majority, with 69% of Americans having less than $1000 in their savings accounts (Huddleston, 2019). For future tests, the current ability domains can be updated and pretested. Although Kim et al. (2017) found the above-average-effect, most of

the 14-items they used were general abilities such as written or spoken expression. More relevant and comprehensive items can be included in future studies and bigger pretest samples (original study: $n = 39$) used to select ability domains and validate the instrument.

How do people assess task difficulty? This question goes beyond the scope of the current investigation yet is a critical open question if difficulty serves as a moderator between the above and below average effects. Difficulty has been described in previous research to increase as a function of cognitive and/or physical load, with those loads being rather additive than interactive components in making difficulty (Feghhi & Rosenbaum, 2019). Different factors might be linked to such perceptions, such as error probability, weights of errors (one error is worse than another), attention demands or potentially a cost-benefit calculation determining judgments of difficulty (Feghhi & Rosenbaum, 2019).

The underlying mechanisms of task difficulty judgments remain unclear, yet in our extension's stimuli, we attempted to embed quantitative numerical information regarding load constructed to be perceived as more and less difficult. We found that these were indeed rated as more and less difficult by the participants. This allows for the use of a quantifiable latent concept such as load as a predictor of difficulty. The operationalization of such latent concepts requires systematic testing in future research.

Together with many past studies, the present replication only establishes the ubiquity of the above and below-average effects. Much less is known about the effects' impacts, especially for the below-average effect. The directionality of the above-average effect's impacts is still debated. Tendencies to see oneself as better than others can serve a wide variety of affective, cognitive, and social functions such as temporary boosts in task performance, longer life expectancy, and well-being (Bopp et al., 2012; Ehrlinger & Dunning, 2003; Taylor & Brown, 1988; Zell et al., 2020). But it can also result in harmful long-term consequences of having unrealistic expectations, heightened disengagement, and decreased self-esteem (Polivy & Herman, 2000; Robins & Beer, 2001). In contrast, less research has been conducted on the below-average-effect's impacts, predominantly focusing on its negative consequences, such as lower grades (Mattern et al., 2010), or worse subjective well-being (Goetz et al., 2006). Other research suggested that the below-average-effect can also induce positive motivational and behavioral consequences in the long run (Whillans et al., 2020). This highlights the need for continued research on the below-and-above-average-effects' consequences.

5 Conclusion

We closely replicated Kruger's (1999) study, showing the above- and below-average effects to be robust. Manipulating the difficulty of (easy and difficult) ability domains participants were to compare themselves with others, which showed that easier items might provoke the above-average effect but dampen the below-average effect and vice versa for more difficult items.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. http://dx.doi.org/10.1037/0022-3514.49.6.1621.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68*(5), 804–825. https://psycnet.apa.org/doi/10.1037/0022-3514.68.5.804.

Altman, D. G., & Royston, P. (2006). The cost of dichotomizing continuous variables. *BMJ*, *332*(7549), 1080. http://dx.doi.org/10.1136/bmj.332.7549.1080.

Aucote, H. M., & Gold, R. S. (2005). Non-equivalence of direct and indirect measures of unrealistic optimism. *Psychology, Health and Medicine*, *10*(2), 194–201. http://dx.doi.org/10.1080/13548500512331315443.

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in Managerial Decision Making*. John Wiley & Sons.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77. http://dx.doi.org/10.1037/0022-3514.90.1.60.

Bopp, M., Braun, J., Gutzwiller, F., & Faeh, D. (2012). Health risk or resource? gradual and independent association between self-rated health and mortality persists over 30 years. *PLoS ONE, 7*(2), 1-10. http://dx.doi.org/10.1371/journal.pone.0030795.

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631–643. https://dx.doi.org/10.1037/0022-3514.79.4.631.

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., . . . Veer, A. V. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224. http://dx.doi.org/10.2139/ssrn.2283856.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition*, *4*(4), 353-376. http://dx.doi.org/10.1521/soco.1986.4.4.353.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, *38*(2), 209–219. http://dx.doi.org/10.1177/0146167211432763.

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*(5), 813–838. http://dx.doi.org/10.1037/0033-2909.130.5.813.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2018). *Package 'pwr'* (1.3-0). https://cran.r-project.org/web/packages/pwr/pwr.pdf.

Chung, J., Schriber, R. A., & Robins, R. W. (2016). Positive illusions in the academic context: A longitudinal study of academic self-enhancement in college. *Personality and Social Psychology Bulletin*, *42*(10), 1384–1401. http://dx.doi.org/10.1177/0146167216662866.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. http://dx.doi.org/10.1177/014662168300700301.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates. http://dx.doi.org/10.1177/014662168300700301.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*(1), 5–17. https://psycnet.apa.org/doi/10.1037/0022-3514.84.1.5.

Epley, N., & Caruso, E. M. (2004). Egocentric ethics. *Social Justice Research, 17*(2), 171–187. https://doi.org/10.1023/B:SORE.0000027408.72713.45.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature, 47*(2), 315–72. http://dx.doi.org/10.1257/jel.47.2.315.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*(6), 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x.

Eriksson, K., & Funcke, A. (2013). A below-average effect with respect to American political stereotypes on warmth and competence. *Political Psychology*, *36*(3), 341–350. http://dx.doi.org/10.1111/pops.12093.

Feghhi, I., & Rosenbaum, D. A. (2019). Judging the subjective difficulty of different kinds of tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(8), 983–994. http://dx.doi.org/10.1037/xhp0000653.

Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological bulletin*, *144*(2), 177–197. http://dx.doi.org/10.1037/bul0000127.

Giladi, E. E., & Klar, Y. (2002). When standards are wide of the mark: Nonselective superiority and inferiority biases in comparative judgments of objects and concepts. *Journal of Experimental Psychology: General, 131*(4), 538–551. https://psycnet.apa.org/doi/10.1037/0096-3445.131.4.538.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*(6), 1110–1126. https://psycnet.apa.org/doi/10.1037/0022-

3514.44.6.1110.

Goetz, T., Ehret, C., Jullien, S., & Hall, N. C. (2006). Is the grass always greener on the other side? Social comparisons of subjective well-being. *The Journal of Positive Psychology*, *1*(4), 173–186. http://dx.doi.org/10.1080/17439760600885655.

Heine, S. J., & Lehman, D. R. (1997). The cultural construction of self-enhancement: An examination of group-serving biases. *Journal of Personality and Social Psychology, 72*(6)*,* 1268–1283. http://dx.doi.org/10.1037/0022-3514.72.6.1268.

Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology, 43*(1), 35–47. https://psycnet.apa.org/doi/10.1037/0022-3514.43.1.35.

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology, 38*(1)*,* 369–425. https://doi.org/10.1146/annurev.ps.38.020187.002101.

Hotelling, H. (1940). The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, *11*(3), 271–283. http://dx.doi.org/10.1214/aoms/1177731867.

Huddleston, C. (2019). Survey: 69% of Americans have less than $1,000 in savings. Retrieved July 27, 2020, from https://www.gobankingrates.com/saving-money/savings-advice/americans-have-less-than-1000-in-savings/.

Johansson, M., & Allwood, C. M. (2007). Own-other differences in the realism of some metacognitive judgments. *Scandinavian Journal of Psychology*, *48*(1), 13–21. http://dx.doi.org/10.1111/j.1467-9450.2007.00565.x.

Kim, Y., Kwon, H., & Chiu, C. (2017). The better-than-average effect is observed because "average" is often construed as below-median ability. *Frontiers in Psychology, 8,* 898. http://dx.doi.org/10.3389/fpsyg.2017.00898.

Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior and Human Decision Processes, 67*(2), 229–245.https://doi.org/10.1006/obhd.1996.0076.

Koellinger, P., Minniti, M., & Schade, C. (2007). "I think I can, I think I can": Overconfidence and entrepreneurial behavior. *Journal of Economic Psychology*, *28*(4), 502–527. https://doi.org/10.1016/j.joep.2006.11.002.

Krizan, Z., & Suls, J. (2008). Losing sight of oneself in the above-average effect: When egocentrism, focalism, and group diffuseness collide. *Journal of Experimental Social Psychology, 44*(4), 929–942. http://dx.doi.org/10.1016/j.jesp.2008.01.006.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221–232. https://doi.org/10.1037/0022-3514.77.2.221.

Kruger, J., & Burrus, J. (2004). Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology, 40*(3), 332–340. http://dx.doi.

org/10.1016/j.jesp.2003.06.002.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53*(4), 636–647. https://psycnet.apa.org/doi/10.1037/0022-3514.53.4.636.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389-402.https://doi.org/10.1177/2515245918787489.

LeBel, Vanpaemel, Cheung, & Campbell. (2019). A brief guide to evaluate replications. *Meta-Psychology,3*. https://dx.doi.org/10.15626/MP.2018.843

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. https://psycnet.apa.org/doi/10.1037/1082-989X.7.1.19.

Mattern, K. D., Burrus, J., & Shaw, E. (2010). When both the skilled and unskilled are unaware: Consequences for academic performance. *Self and Identity*, *9*(2), 129–141. http://dx.doi.org/10.1080/15298860802618963.

Moore, D. A., & Kim, T. G. (2003). Myopic Social Prediction and the Solo Comparison Effect. *Journal of Personality and Social Psychology*, *85*(6), 1121–1135. http://dx.doi.org/10.1037/0022-3514.85.6.1121.

Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes, 102*(1), 42–58. https://doi.org/10.1016/j.obhdp.2006.09.005.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197–213. http://dx.doi.org/10.1016/j.obhdp.2006.09.002.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology, 92*(6), 972–989. http://dx.doi.org/10.1037/0022-3514.92.6.972.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://psycnet.apa.org/doi/10.1037/0033-295X.115.2.502.

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., & Altman, D. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, *32*(3), 437–440. http://dx.doi.org/10.3174/ajnr.a2425.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*(4), 339–363. http://dx.doi.org/10.1037/0033-295x.90.4.339.

Polivy, J., & Herman, C. P. (2000). The false-hope syndrome: Unfulfilled expectations of self-change. *Current Directions in Psychological Science, 9*(4), 128–131. https://psycnet.apa.org/doi/10.1111/1467-8721.00076.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80*(2), 340–352. https://psycnet.apa.org/doi/10.1037/0022-3514.80.2.340.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37*(3), 322–336. https://doi.org/10.1037/0022-3514.37.3.322.

Roth, A. V., Schroeder, R., Huang, X., & Kristal, M. (2008). *Handbook of metrics for research in operations management: Multi-item measurement scales and objective items*. London: SAGE.

Sedikides, C., Meek, R., Alicke, M. D., & Taylor, S. (2014). Behind bars but above the bar: Prisoners consider themselves more prosocial than non-prisoners. *British Journal of Social Psychology, 53*(2), 396–403. https://doi.org/10.1111/bjso.12060.

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341.

Srull, T. K., & Gaelick, L. (1983). General principles and individual differences in the self as a habitual reference point: An examination of self-other judgments of similarity. *Social Cognition, 2*(2), 108–121. https://psycnet.apa.org/doi/10.1521/soco.1983.2.2.108.

Stewart, M., Brown, J. B., Weston, W., McWhinney, I. R., McWilliam, C. L., & Freeman, T. (2013). *Patient-centered medicine: transforming the clinical method*. CRC Press.

Sundström, A. (2008). Self-assessment of driving skill. A review from a measurement perspective. *Transportation Research Part F: Traffic Psychology and Behaviour, 11*(1), 1–9. https://psycnet.apa.org/doi/10.1016/j.trf.2007.05.002.

Sweeny, K., & Shepperd, J. A. (2007). Do people brace sensibly? Risk judgments and event likelihood. *Personality and Social Psychology Bulletin*, *33*(8), 1064–1075. http://dx.doi.org/10.1177/0146167207301024.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. http://dx.doi.org/10.5116/ijme.4dfb.8dfd[

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210.http://dx.doi.org/10.1037/0033-2909.103.2.193.

Walsh, E., & Ayton, P. (2009). My imagination versus your feelings: Can personal affective forecasts be improved by knowing other peoples' emotions? *Journal of Experimental*

Psychology: Applied, *15*(4), 351–360. http://dx.doi.org/10.1037/a0017984.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806–820. http://dx.doi.org/10.1037/0022-3514.39.5.806.

Weinstein, N. D. (1983). Reducing unrealistic optimism about illness susceptibility. *Health Psychology*, *2*(1), 11–20. http://dx.doi.org/10.1037/0278-6133.2.1.11.

Whillans, A. V., Jordan, A. H., & Chen, F. S. (2020). The upside to feeling worse than average (WTA): A conceptual framework to understand when, how, and for whom WTA beliefs have long-term benefits. *Frontiers in Psychology*, *11*, 642. http://dx.doi.org/10.3389/fpsyg.2020.00642.

Wilson, A. E., & Ross, M. (2001). From chump to champ: People's appraisals of their earlier and present selves. *Journal of Personality and Social Psychology, 80*(4), 572–584. https://psycnet.apa.org/doi/10.1037/0022-3514.80.4.572.

Windschitl, P. D., Rose, J. P., Stalkfleet, M. T., & Smith, A. R. (2008a). Are people excessive or judicious in their egocentrism? A modeling approach to understanding bias and accuracy in people's optimism. *Journal of Personality and Social Psychology, 95*(2), 253–273. https://psycnet.apa.org/doi/10.1037/0022-3514.95.2.253.

Windschitl, P. D., Conybeare, D., & Krizan, Z. (2008b). Direct-comparison judgments: When and why above- and below-average effects reverse. *Journal of Experimental Psychology: General, 137*(1), 182–200. https://doi.org/10.1037/0096-3445.137.1.182.

Zell, E., & Alicke, M. D. (2011). Age and the better-than-average effect. *Journal of Applied Social Psychology, 41*(5), 1175–1188. https://doi.org/10.1111/j.1559-1816.2011.00752.x.

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin, 146*(2), 118–149. https://psycnet.apa.org/doi/10.1037/bul0000218.

Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of Alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, *12*(6), 1005–1017. https://doi.org/10.1177/1948550620948973.

Kruger (1999): Replication and extensions
Supplementary

Contents

Open Science disclosures

Main manuscript, data and code

The manuscript, data and code are shared using the Open Science Framework. Review link for data and code of the study: osf.io/7yfkc

Final pre-registration is on: https://osf.io/byx4z
(there was one previous pre-registration, 10 minutes before the final one. Both were conducted before data collection, and the second one simply fixed a minor glitch regarding a default set in one of the comparative questions. Previous pre-registration is on: https://osf.io/nm568/

Procedure and data disclosures

Data collection

Data collection was completed before analysing the data.

Conditions reporting

All collected conditions are reported.

Data exclusions

Details are reported in the materials section of this document

Variables reporting

All variables collected for this study are reported and included in the provided data.

Exclusion criteria

The current replication's exclusion criteria are summarised in table 4.

Table 1.

Summary of exclusion criteria

Exclusion Criteria	Reason
Exclusion Criteria 1a Participants indicating low proficiency of English. Item: "On a scale from 1-7, what do you think is your proficiency of English?" (1 = being not proficient at all, 7 = being very proficient.) *Exclusion: if self-report less than 5 on a 1-7 scale*	Without a fair proficiency of English participants may not fully understand the questions and may affect results.
Exclusion Criteria 1b Participants who self-report not being serious about filling in the survey. Item: "On a scale from 1-5, what do you think is your seriousness in filling in the survey?" (1 = not serious at all, 5 = very serious.) *Exclusion: if self-report less than 4 on a 1-5 scale*	Nonserious answering behavior increases noise and reduces experimental power. Excluding their response can increase data validity.
Exclusion Criteria 1c Participants who correctly guessed any one of the hypotheses of this study in the funnelling section. Item: "What do you think the purpose of the last part was?" (If you are not sure please write "not sure") *Exclusion: if guessed correctly for both replication and extension*	Participants who could guess any of the hypotheses of the study may commit experimental bias and do not reflect the true nature of the investigated phenomenon

Exclusion Criteria 1d Participants who have already seen or done the survey before. Item: "Have you ever seen the materials used in this study or similar before? If yes - please indicate where." *Exclusion: answered 'yes'*	Experimental bias in responses
Exclusion Criteria 1e Participants who failed to complete the survey. (duration = 0, leave question blank)	Incomplete data
Exclusion Criteria 1f Participants not from the United States. Item: "Which country are you originally from? (Country of birth)" *Exclusion: if response is not part of the United States*	Does not fit our targeted population criteria

Additional Tables and Figures

Replication condition

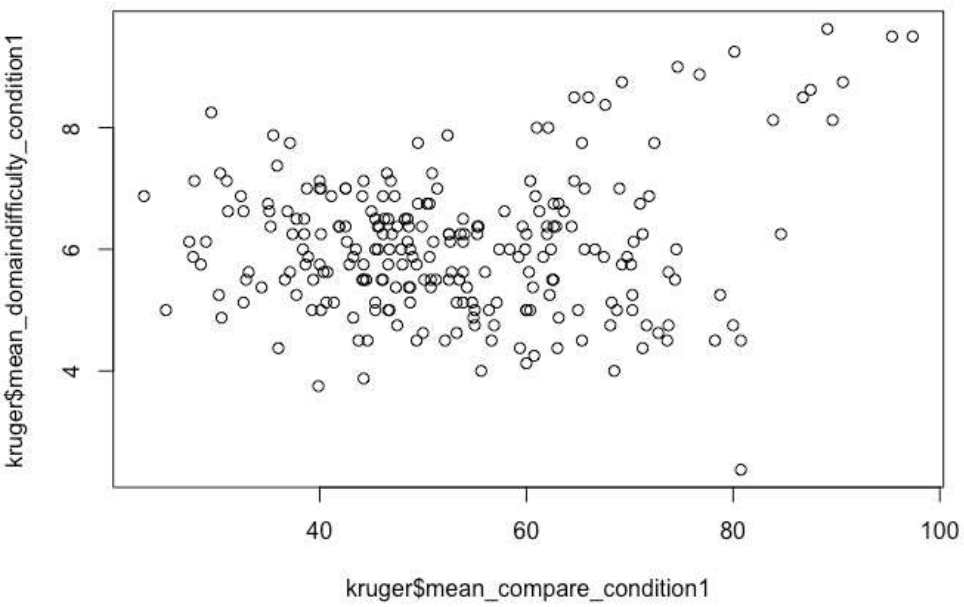


Figure 1.1 Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the replication group.

Table 2

Replication condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	1.27	[-4.13, 6.67]						
kruger$mean_own_condition1	9.51**	[8.57, 10.44]	0.90	[0.81, 0.99]	.45	[.36, .54]	.85**	
kruger$mean_other_condition1	-0.87	[-2.00, 0.26]	-0.07	[-0.16, 0.02]	.00	[-.00, .01]	.53**	
								R^2 = .732**
								95% CI[0.68, 0.79]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

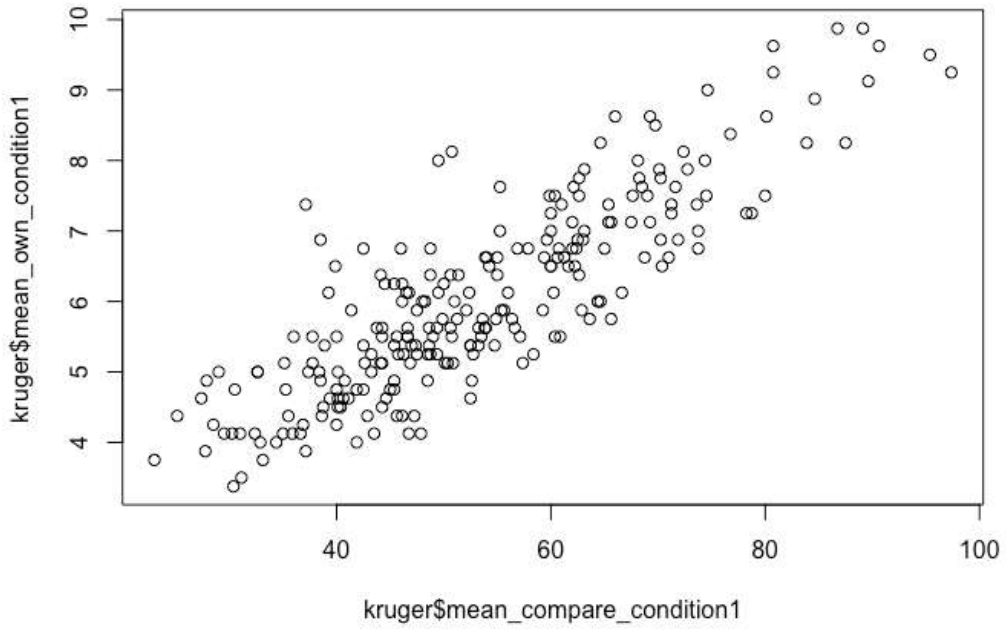


Figure 1.2. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the replication group.

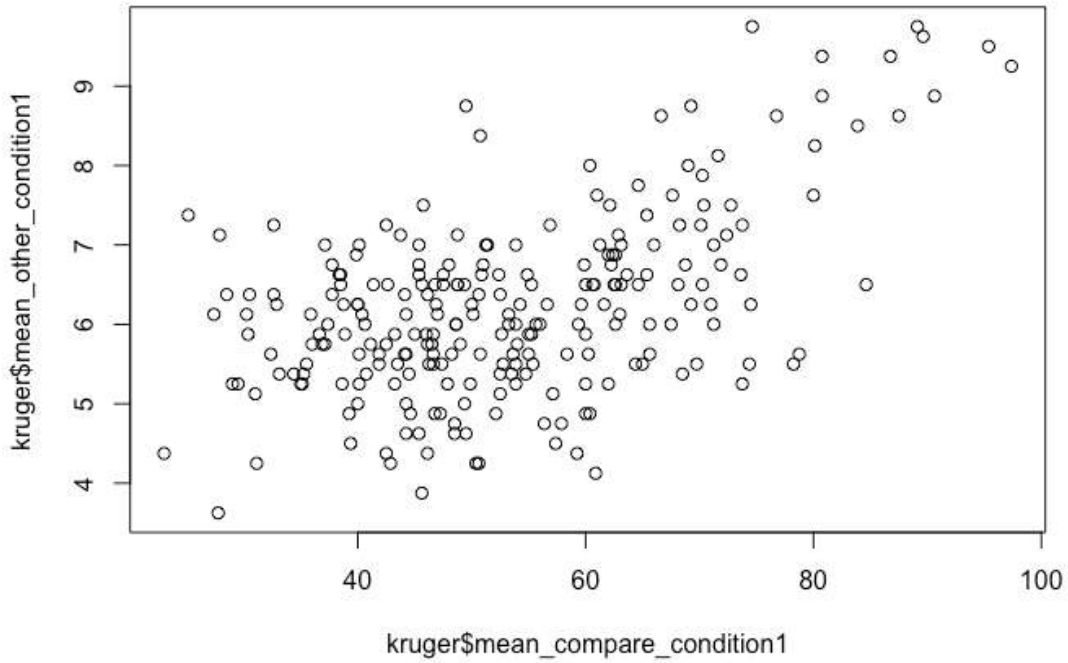


Figure 1.3. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the replication group.

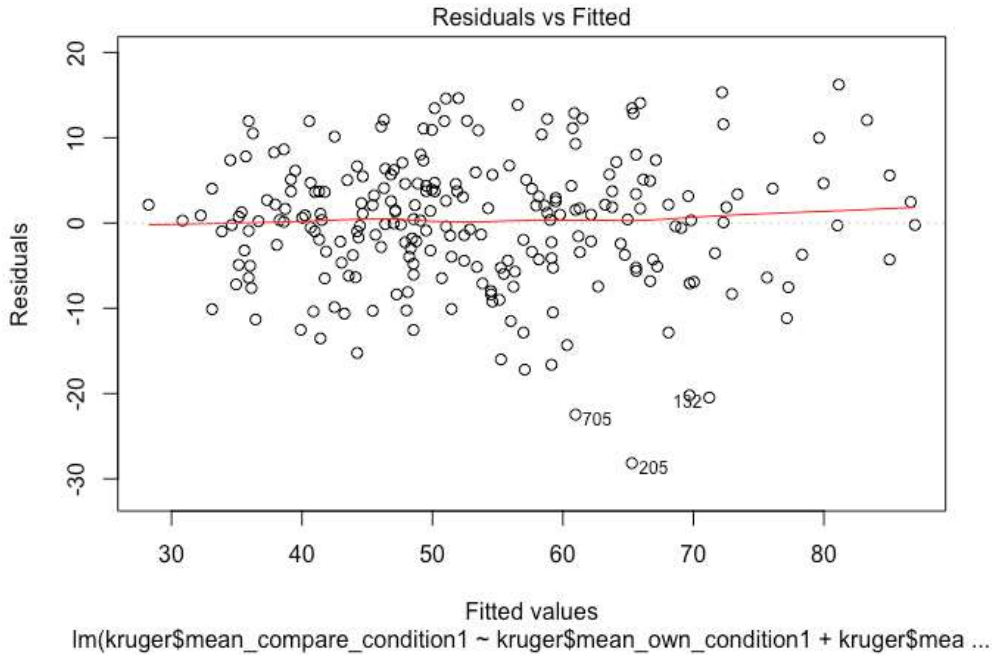


Figure 1.4. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the replication group.

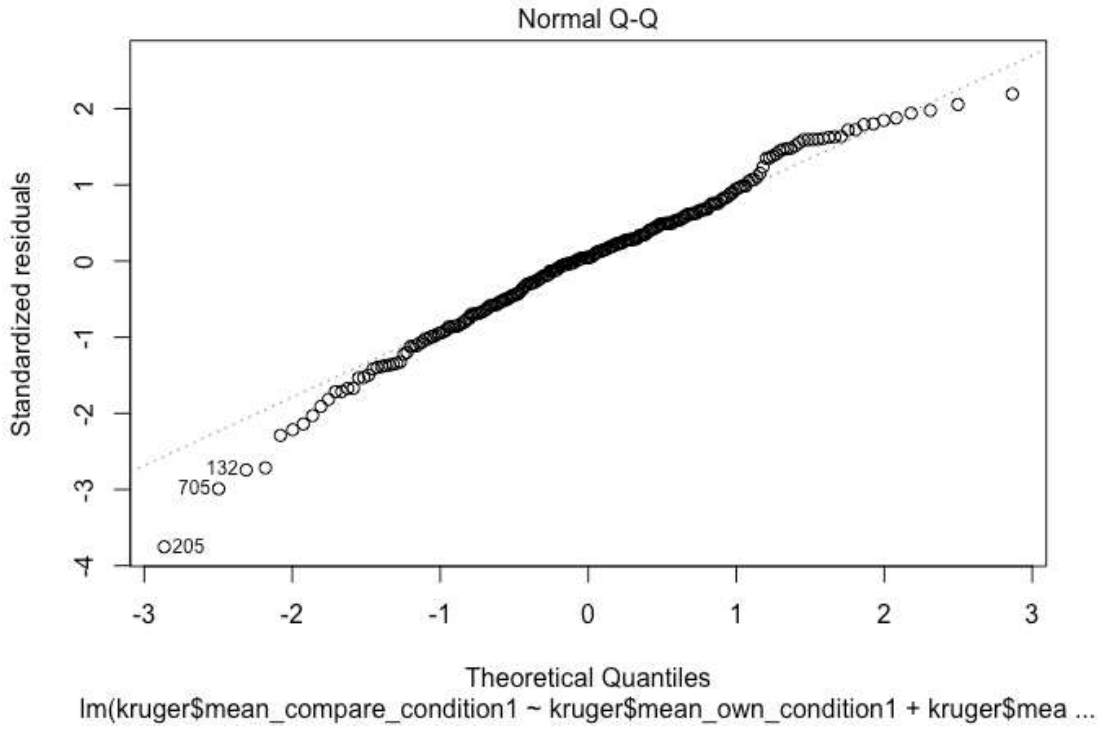


Figure 1.5. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the replication group.

Correlation Matrix Replication Condition

Table 3.1
Person's r for mean values (across abilities) in the replication (original) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.16	1.00				
Own ability	0.85	0.21	1.00			
Peers' ability	0.53	0.34	0.67	1.00		
Desirability	0.11	0.06	0.14	0.20	1.00	
Ambiguity	-0.06	-0.01	-0.11	-0.13	-0.35	1.00

Table 3.2
P-values for correlations between mean values (across abilities) in the replication (original) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	.012	0				
Own ability	<.001	.001	0			
Peers' ability	<.001	<.001	<.001	0		
Desirability	0.090	0.362	.025	.002	0	
Ambiguity	.340	.829	.093	.039	<.001	0

Easy domain condition

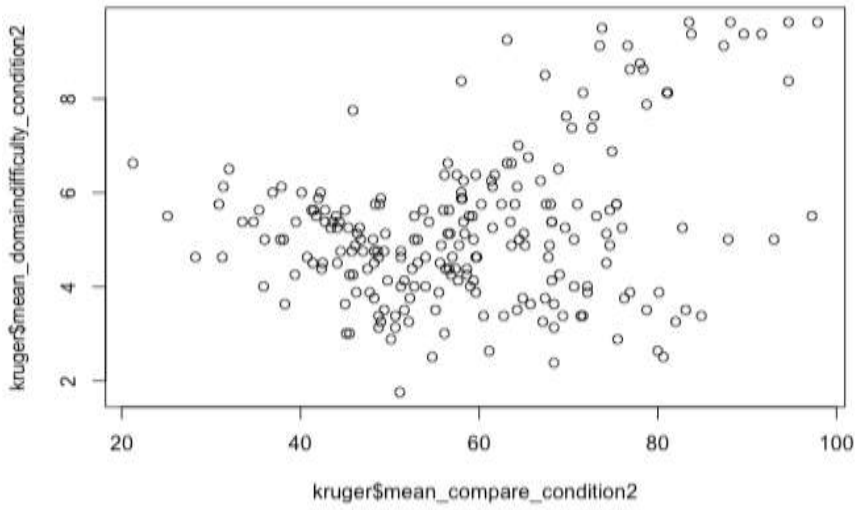


Figure 2.1. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the easy domain group.

Table 4

Easy domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	2.90	[-2.99, 8.80]						
kruger$mean_own_condition2	8.92**	[7.90, 9.93]	0.86	[0.76, 0.96]	.42	[.33, .51]	.83**	
kruger$mean_other_condition2	-0.50	[-1.59, 0.59]	-0.04	[-0.14, 0.05]	.00	[-.00, .01]	.52**	
								R^2 = .690** 95% CI[0.62,0.76]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
** indicates $p < .01$.

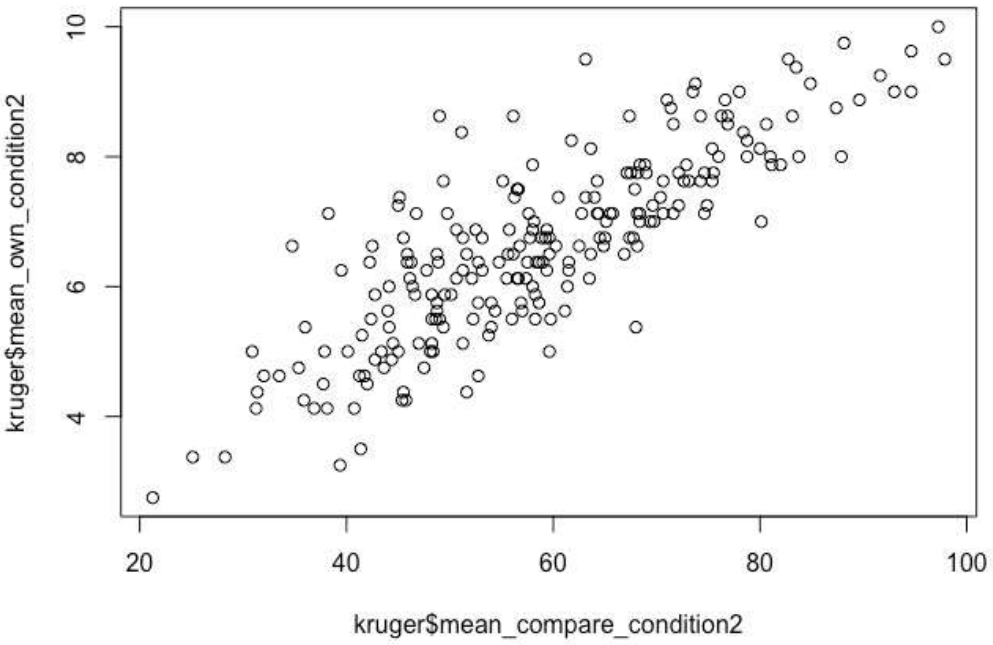


Figure 2.2. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the easy domain group.

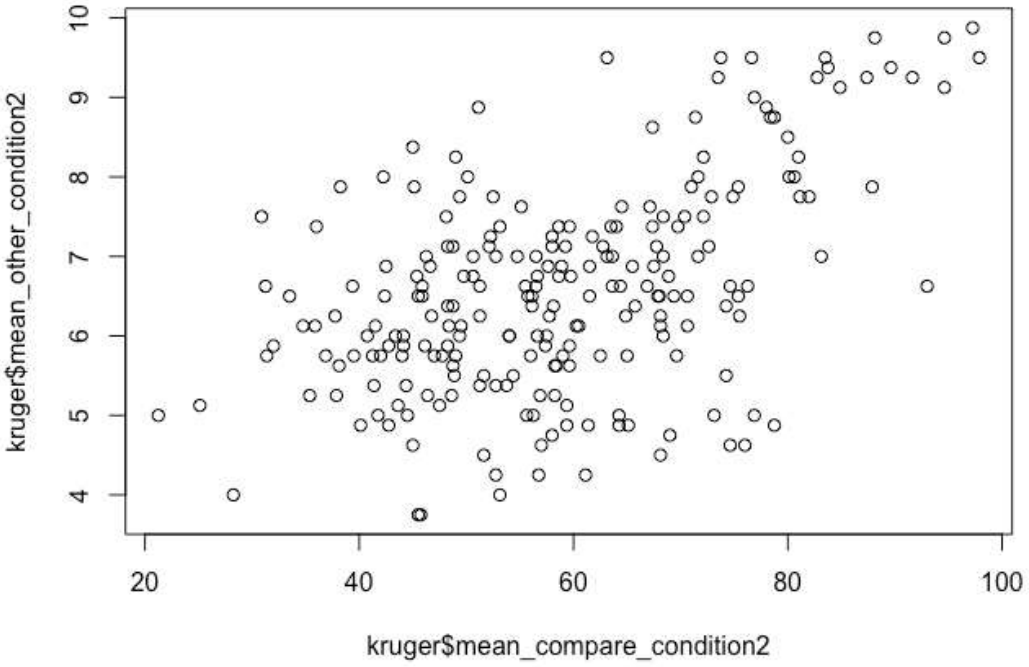


Figure 2.3. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the easy domain group.

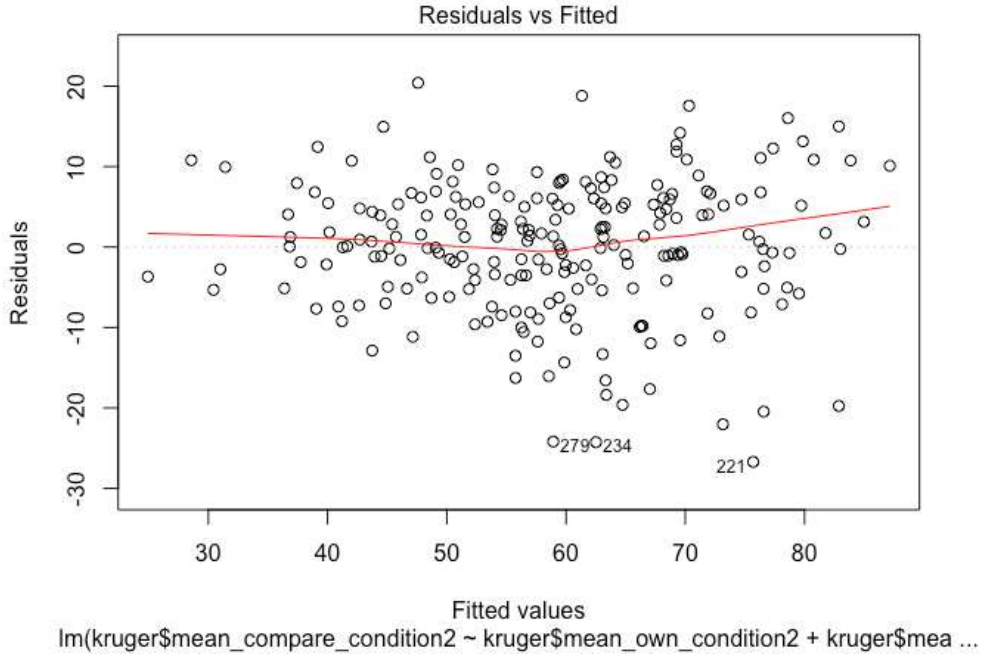


Figure 2.4. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

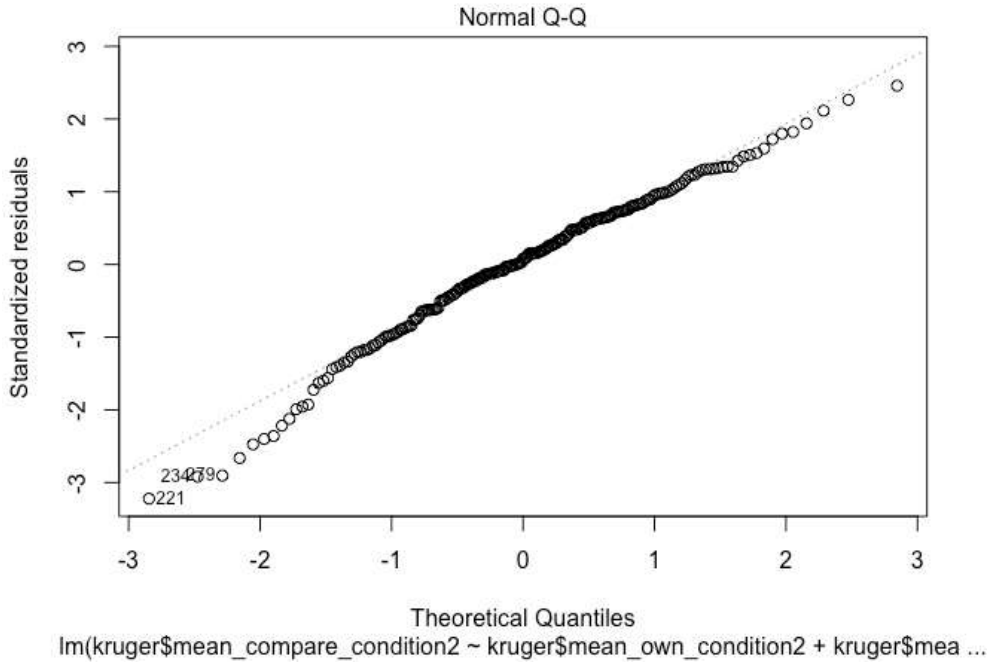


Figure 2.5. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

Correlation Matrix Easy Extension

Table 5.1

Person's r for mean values (across abilities) in the easy (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.83	1.00				
Own ability	0.32	0.28	1.00			
Peers' ability	0.52	0.66	0.37	1.00		
Desirability	0.29	0.40	0.15	0.41	1.00	
Ambiguity	0.08	0.18	-0.06	0.23	0.43	1.00

Table 5.2

P-values for correlations between mean values (across abilities) in the easy (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	<.001	0				
Own ability	<.001	<.001	0			
Peers' ability	<.001	<.001	<.001	0		
Desirability	<.001	<.001	0.0228	<.001	0	
Ambiguity	0.2630	0.0056	0.3420	0.0006	<.001	0

Difficult domain condition

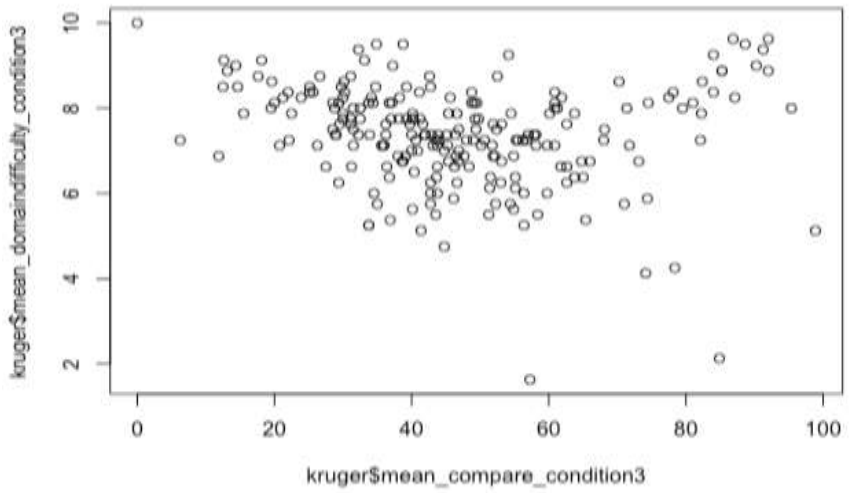


Figure 3.1. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the difficult domain group.

Table 6

Difficult domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	9.10**	[5.40, 12.80]						
kruger$mean_own_condition3	8.39**	[7.33, 9.45]	0.90	[0.79, 1.01]	.27	[.19, .34]	.87**	
kruger$mean_other_condition3	-0.42	[-1.62, 0.78]	-0.04	[-0.15, 0.07]	.00	[-.00, .00]	.70**	
								R^2 = .755**
								95% CI[0.70, 0.81]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

** indicates $p < .01$.

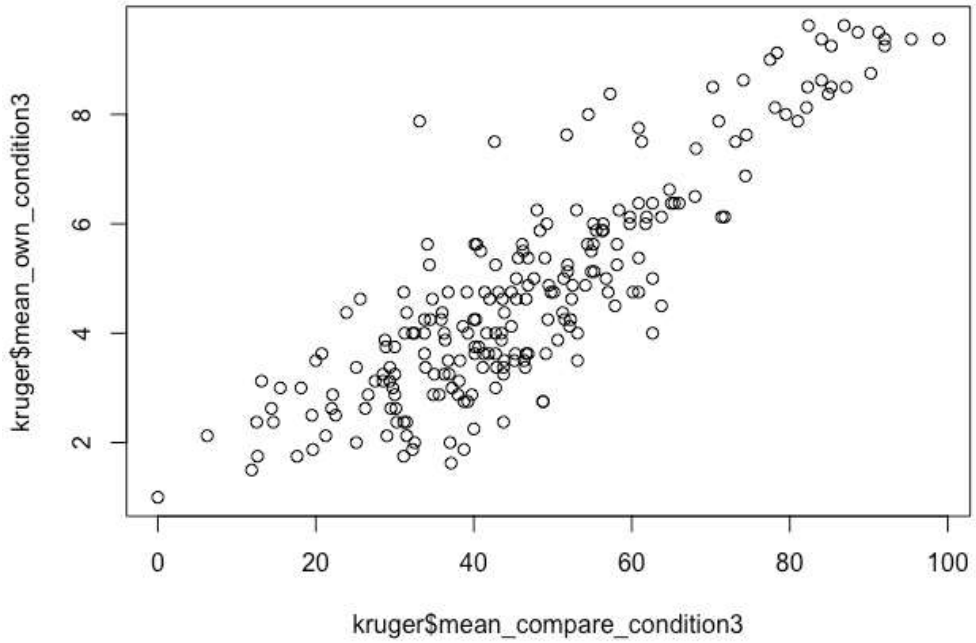


Figure 3.2. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the difficult domain group.

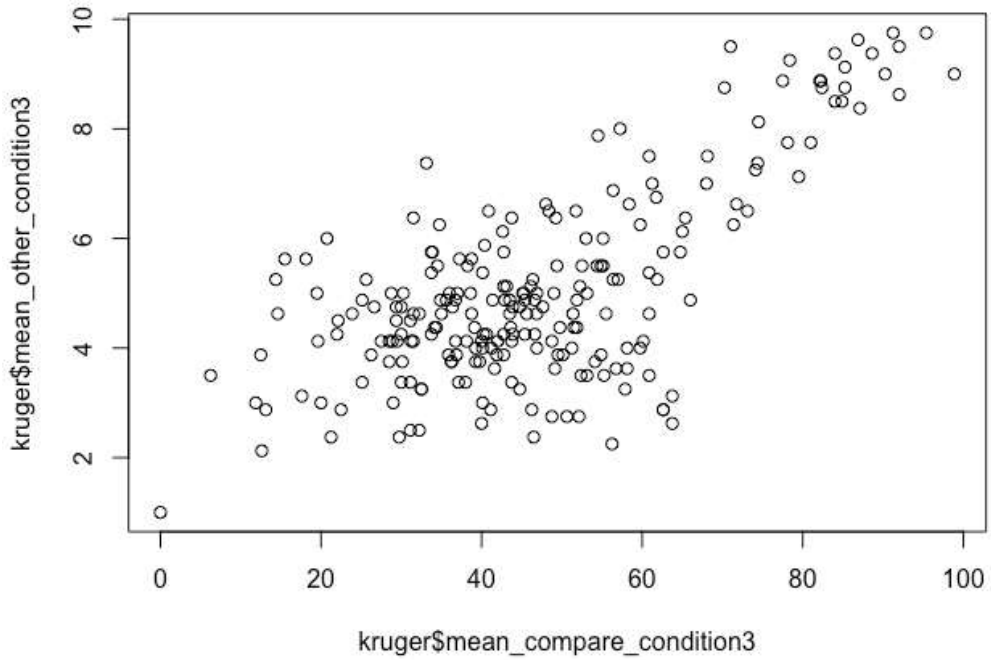


Figure 3.3. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the difficult domain group.

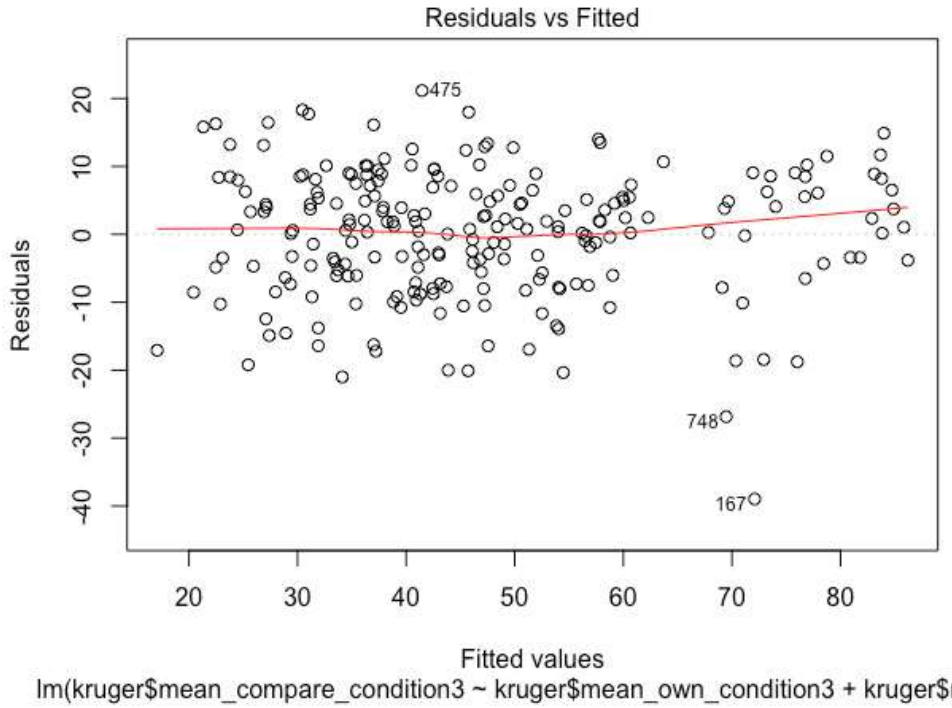


Figure 3.4. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

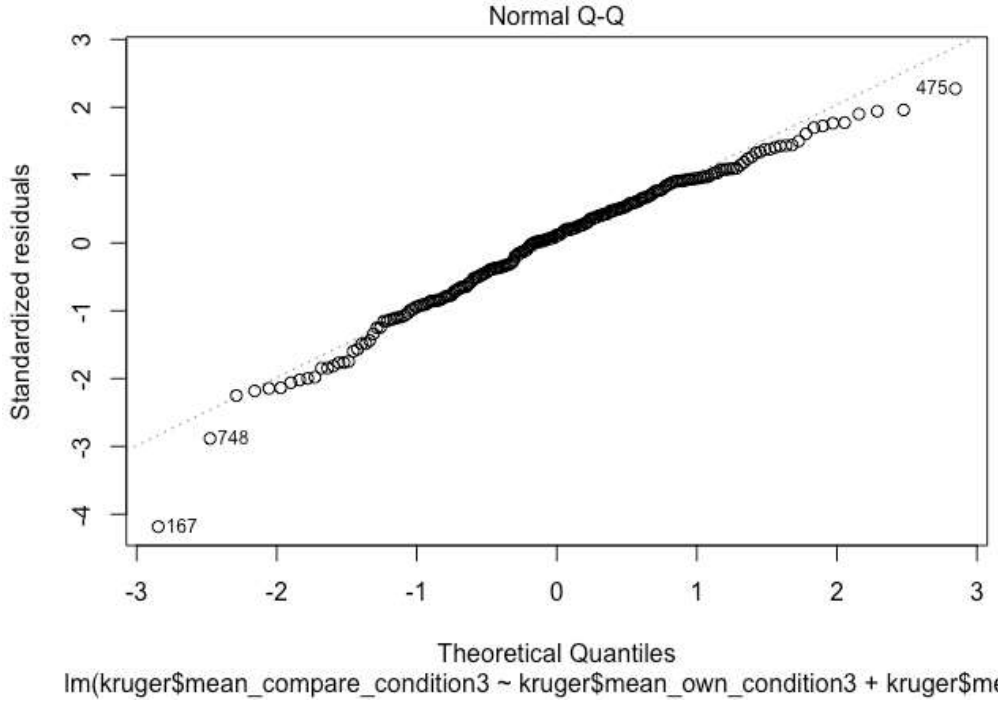


Figure 3.5. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

Correlation Matrices Difficult Extension

Table 7.1
Person's r for mean values (across abilities) in the difficult (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.86	1.00				
Own ability	-0.13	-0.13	1.00			
Peers' ability	0.70	0.82	-0.03	1.00		
Desirability	0.20	0.13	0.27	0.22	1.00	
Ambiguity	-0.03	-0.09	0.27	0.01	0.43	1.00

Table 7.2
P- values for correlations between mean values (across abilities) in the difficult (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	<.0001	0				
Own ability	0.0498	0.0595	0			
Peers' ability	<.0001	<.0001	0.6070	0		
Desirability	0.0025	0.0467	<.0001	0.0007	0	
Ambiguity	0.6860	0.1850	<.0001	0.8370	<.0001	0

One-sample Wilcoxon-tests

Table 8.1
One-sample Wilcoxon tests testing median comparative ability scores against the scale mid-point in the original (replication) condition

Item	Median	90% CI		P-value	Effect size *r*
		Lower	Upper		
Using mouse	75.50	73.00	78.00	<.001	0.81
Driving	70.50	67.70	73.00	<.001	0.61
Riding bicycle	64.50	61.00	67.50	<.001	0.52
Saving money	64.50	61.50	67.50	<.001	0.53
Telling jokes	53.00	50.00	56.50	0.0569	0.13
Playing chess	40.00	36.00	44.00	<.001	0.30
Juggling	28.00	24.00	33.50	<.001	0.55
Computer programming	40.00	36.00	44.00	<.001	0.30

Table 8.2
One-sample Wilcoxon tests testing median comparative ability scores against the scale mid-point in the easy domain (extension) condition

Item	Median	90% CI		P-value	Effect size *r*
		Lower	Upper		
Using mouse	76.50	74.00	79.50	<.001	0.76
Driving	71.00	67.50	74.00	<.001	0.62
Riding bicycle	69.00	65.60	70.50	<.001	0.63
Saving money	66.50	62.50	70.50	<.001	0.49
Telling jokes	62.00	58.50	65.50	<.001	0.43
Playing chess	47.50	43.50	52.00	0.26	0.06
Juggling	46.00	41.00	50.50	0.085	0.08
Computer programming	49.50	46.00	53.50	0.85	0.01

Table 8.3
One-sample Wilcoxon tests results testing against the scale mid-point for comparative ability in difficult domain (extension) condition

Item	Median	90% CI		P-value	Effect size *r*
		Lower	Upper		
Using mouse	59.00	55.50	62.50	<.001	0.33
Driving	39.50	34.50	44.00	<.001	0.30
Riding bicycle	49.50	45.00	53.50	0.69	0.02
Saving money	66.50	62.50	70.50	<.001	0.46
Telling jokes	40.00	35.50	44.00	<.001	0.30
Playing chess	40.50	36.00	44.50	<.001	0.25

Juggling	37.50	33.00	42.00	<.001	0.35
Computer programming	45.00	40.50	49.00	0.01	0.16

One-sample t-tests

Table 9.1
One-sample t-tests results testing against the scale mid-point for comparative ability in original (replication) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	71.20	18.34	1.6961E-47	239	68.92	73.47
Driving	65.17	10.64	6.7221E-22	239	62.36	67.97
Riding bicycle	61.01	8.33	6.3572E-15	239	58.41	63.62
Saving money	62.88	9.45	3.1702E-18	239	60.19	65.56
Telling jokes	52.42	1.66	0.0987806	239	49.54	55.30
Playing chess	40.98	-5.18	4.8043E-07	239	37.55	44.41
Juggling	31.98	-10.09	3.5936E-20	239	28.46	35.50
Computer programming	40.73	-4.92	1.6472E-06	239	37.01	44.44

Table 9.2
One-sample t-tests results testing against the scale mid-point for comparative ability in easy domain (extension) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	71.27	15.56	1.72E-37	224	68.58	73.97
Driving	66.32	10.77	4.63E-22	224	63.34	69.31
Riding bicycle	65.76	10.78	4.26E-22	224	62.88	68.64
Saving money	63.63	8.00	6.58E-14	224	60.27	66.98
Telling jokes	59.74	7.11	1.56E-11	224	57.04	62.44
Playing chess	47.81	-1.19	0.23413041	224	44.19	51.43
Juggling	46.76	-1.75	0.08125851	224	43.11	50.41
Computer programming	49.57	-0.25	0.80363761	224	46.20	52.95

Table 9.3
One-sample t-tests results testing against the scale mid-point for comparative ability in difficult domain (extension) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	55.79	4.12	5.2822E-05	225	53.02	58.56
Driving	40.63	-4.81	2.7508E-06	225	36.79	44.47
Riding bicycle	48.90	-0.60	0.54982103	225	45.29	52.51
Saving money	62.68	7.40	2.7397E-12	225	59.30	66.06
Telling jokes	40.82	-5.10	7.0415E-07	225	37.28	44.37

Playing chess	41.86	-4.49	1.1181E-05	225	38.29	45.43
Juggling	39.67	-5.61	5.7662E-08	225	36.04	43.29
Computer programming	45.36	-2.68	0.00792618	225	41.95	48.77

Table 9.4

One-sample t-tests results testing against the scale mid-point for desirability in original (replication) condition

Item	**estimate**	**statistic**	**p-value**	**df**	**d**	**Lower CI**	**Upper CI**
Using a computer mouse	8.654	34.058	1.11E-93	239	2.203	1.968	2.435
Driving	9.146	47.234	3.34E-123	239	3.055	2.750	3.354
Riding bicycle	7.996	27.992	2.14E-77	239	1.811	1.604	2.016
Saving money	9.129	45.568	7.60E-120	239	2.948	2.651	3.237
Telling jokes	7.500	20.858	9.87E-56	239	1.349	1.173	1.523
Playing chess	7.613	20.269	7.96E-54	239	1.311	1.138	1.483
Juggling	6.529	10.278	9.33E-21	239	0.665	0.524	0.804
Computer programming	8.663	33.998	1.58E-93	239	2.199	1.964	2.431

Power Simulation for Exploratory Analysis

Table 10
*Power Simulation for Main and Interaction effects for 3(Condition)*2(Difficulty) mixed design*

Effect	Power	Effect Size
Condition	100	0.12353699
Difficulty	100	0.28369743
Interaction	99.7	0.04595521

Power simulations in R using the "Superpower" package (Lakens & Caldwell, 2021) showed that using our sample of *n* = 691 (with sample size by cell/between factor: replication group n = 240, easy extension n = 225, difficult extension n = 226), near 100% power was reached to examine main & interaction effects. For some of the multiple comparisons we have however power close to 0%

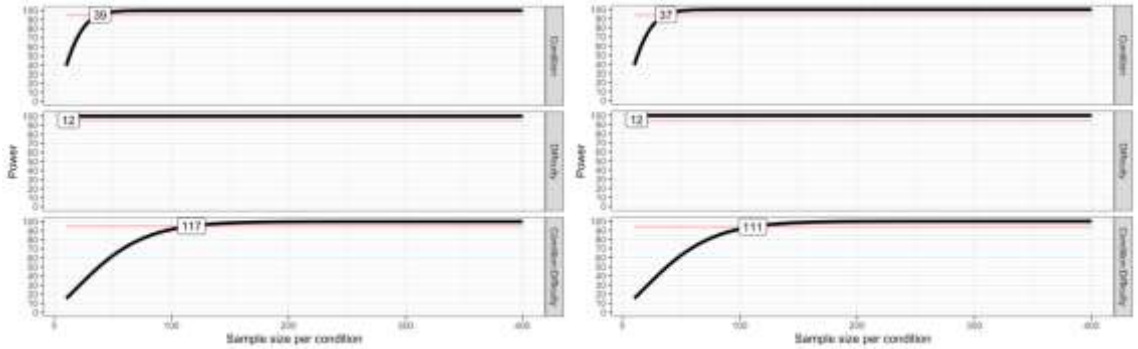
Comparison	Power	Effect Size
Condition_Replication_Difficulty_Easy VS Condition_Replication_Difficulty_Difficult	100	-1.3760623
Condition_Replication_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Easy	**3.8**	**0.10723954**
Condition_Replication_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Difficult	100	-0.8515678
Condition_Replication_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Easy	100	-0.7868727
Condition_Replication_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	100	-1.2543963
Condition_Replication_Difficulty_Difficult VS Condition_Easy Extension_Difficulty_Easy	100	1.39440291
Condition_Replication_Difficulty_Difficult VS Condition_Easy Extension_Difficulty_Difficult	98.5	0.48786589
Condition_Replication_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Easy	99.2	0.53517156
Condition_Replication_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Difficult	**0.5**	**0.01912579**
Condition_Easy Extension_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Difficult	100	-0.8971173
Condition_Easy Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Easy	100	-0.8359101
Condition_Easy Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	100	-1.2817852
Condition_Easy Extension_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Easy	**1**	**0.05121554**
Condition_Easy Extension_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Difficult	94.7	-0.4397807
Condition_Difficult Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	98.2	-0.4847737

Those are:

(1) Replication Condition Easy items compared to Easy Extension Condition Easy Items,
(2) Replication Condition Difficult Items vs Difficult Extension Difficult Items, and
(3) Easy Extension Difficulty Items vs Difficult Condition Easy Items.

Moreover, the sample size per cell was slightly too small to reach 95% power

Figure 2.13
*Power Curves for Main and Interaction effects for 3(Condition)*2(Difficulty) mixed design*



Left panel: 95% power. Right panel: 94% power.
Note: the sample size required to reach 95% power was n = 117 for each cell. Using this calculation, our collected sample was slightly too small for the extension conditions (easy extension: n = 225/2 = 112 each cell, and difficult extension: n = 226/2 = 113 each cell).

Equivalence Tests

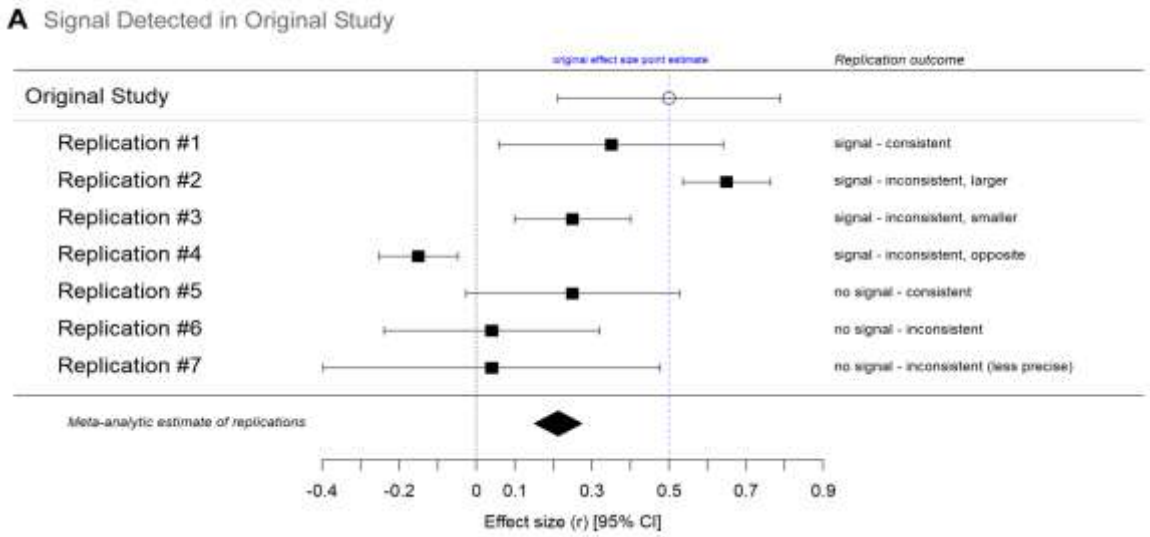
Table 11. Equivalence tests

EQ Test #	Correlation	Variable Controlled for	df	*r*diff	Upper 95% CI	Lower 95% CI	*p*	Condition
1	Comparative ability & domain difficulty (as in original)	desirability	5	0.08604	0.018	0.212	0.91	Replication
2		ambiguity	5	-0.0085	-0.012	-0.005	0.33	Replication
3	Comparative ability & domain difficulty (vector-wise)	desirability	1917	-0.0033	-0.016	0.007	0.55	Replication
4		ambiguity	1917	-0.0085	-0.016	-0.004	0.002	Replication
5	Comparative ability & domain difficulty (mean)	desirability	237	0.00527	-0.003	0.029	0.22	Replication
6		ambiguity	237	0.00054	-0.008	0.016	0.69	Replication
7	Comparative ability & domain difficulty (as in original)	desirability	5	0.09	0.02	0.159	0.005	Easy Extension
8		ambiguity	5	-0.084	-0.193	-0.011	0.03	Easy Extension
9	Comparative ability & domain difficulty (as in original)	desirability	5	0.04022	-0.238	0.126	0.67	Difficult Extension
10		ambiguity	5	0.03	-0.009	0.578	0.91	Difficult Extension
11	Comparative ability & domain difficulty (vector-wise)	desirability	1798	0.006	-0.007	0.018	0.32	Easy Extension

12		ambiguity	1798	-0.03	-0.041	-0.02	0.001	Easy Extension
13	Comparative ability & domain difficulty (vector-wise)	desirability	1806	0.023	0.015	0.034	0.002	Difficult Extension
14		ambiguity	1806	0.001	-0.0003	0.005	0.13	Difficult Extension
15	Comparative ability & domain difficulty (mean)	desirability	223	0.029	-0.0003	0.067	0.049	Easy Extension
16		ambiguity	223	-0.006	-0.028	0.002	0.15	Easy Extension
17	Comparative ability & domain difficulty (mean)	desirability	223	0.064	0.028	0.126	< .001	Difficult Extension
18		ambiguity	223	0.0025	-0.046	0.0349	0.93	Difficult Extension

Criteria for evaluation of replications

A simplified replication taxonomy for comparing replication effects to target article original findings by (LeBel et al., 2019)



Replication evaluation

We used the replication classification criteria by LeBel and colleagues' (2018) summarized in Table 6. We categorized the current replication as a "close replication" and provided details in Table 7. Variables and questions were the same as in the original, with the addition of extensions and adjustments to fit the MTurk sample, instead of Cornell university students.

Criteria for evaluation of replications by LeBel et al. (2018)

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication		Conceptual replication		
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
IV operationalization	Same/similar	Same/similar	Same/similar	Different	
DV operationalization	Same/similar	Same/similar	Same/similar	Different	
IV stimuli	Same/similar	Same/similar	Different		
DV stimuli	Same/similar	Same/similar	Different		
Procedural details	Same/similar	Different			
Physical setting	Same/similar	Different			
Contextual variables	Different				

A classification of relative methodological similarity of a replication study to an original study. "Same" ("different") indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. "Everything controllable" indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.).
"Similar" category was added to the LeBel et al. (2018) typology to refer to minor deviations, aimed to adjust the study to the target sample, that are not expected to have major implications on replication success.

Correlations per condition

Correlation matrix for the replication condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.35*** [-0.39, -0.31]				
Own Ability	0.81*** [0.79, 0.82]	-0.46*** [-0.50, -0.43]			
Others' Ability	0.50*** [0.46, 0.53]	-0.37*** [-0.41, -0.33]	0.64*** [0.62, 0.67]		
Desirability	0.30*** [0.25, 0.34]	-0.07** [-0.11, -0.02]	0.34*** [0.30, 0.38]	0.29*** [0.25, 0.33]	
Ambiguity	-0.10*** [-0.14, -0.06]	0.13*** [0.09, 0.17]	-0.15*** [-0.19, -0.10]	-0.20*** [-0.24, -0.15]	-0.17*** [-0.21, -0.12]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Correlation matrix for the easy extension condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.27** [-0.31, -0.22]				
Own Ability	0.78*** [0.76, 0.80]	-0.37*** [-0.41, -0.33]			
Others' Ability	0.47*** [0.44, 0.51]	-0.24*** [-0.28, -0.20]	0.61*** [0.58, 0.64]		
Desirability	0.28*** [0.24, 0.32]	-0.02 [-0.06, 0.03]	0.36*** [0.31, 0.39]	0.34*** [0.29, 0.38]	
Ambiguity	-0.19*** [-0.23, -0.14]	0.19*** [0.15, 0.24]	-0.26*** [-0.30, -0.22]	-0.28*** [-0.32, -0.23]	-0.25*** [-0.29, -0.21]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Correlation matrix for the difficult extension condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.31*** [-0.35, -0.27]				
Own Ability	0.78*** [0.76, 0.80]	-0.33*** [-0.37, -0.29]			
Others' Ability	0.45*** [0.41, 0.48]	-0.18*** [-0.23, -0.14]	0.56*** [0.52, 0.59]		
Desirability	0.13*** [0.08, 0.17]	0.14*** [0.09, 0.18]	0.14*** [0.09, 0.18]	0.15*** [0.10, 0.19]	
Ambiguity	-0.02 [-0.07, 0.02]	-0.03 [-0.08, 0.01]	-0.01 [-0.06, 0.04]	-0.01 [-0.06, 0.03]	-0.24*** [-0.28, -0.19]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Replication condition: Item-wise correlations between domain difficulty and comparative ability ratings for each ability domain

			95% CI	
Ability domain	***p***	***r***	**Lower**	**Upper**
Using a computer mouse	.768	-0.02	-0.15	0.11
Driving	.006	-0.18	-0.30	-0.05
Riding a bicycle	.790	0.02	-0.11	0.14
Saving money	.030	-0.14	-0.26	-0.01
Telling jokes	.161	-0.09	-0.22	0.04
Playing chess	.020	-0.15	-0.27	-0.02
Juggling	<.001	-0.25	-0.37	-0.13
Computer programming	<.001	-0.23	-0.34	-0.10

Reliability for domains across conditions

Variable	Original domains condition (n = 240)	Easy domains condition (n = 225)	Difficult domains condition (n = 226)
Domain difficulty	.64 (.68, .48)	.76 (.76, .48)	.68 (.47, .61)
Comparative ability	.76 (.61, .72)	.77 (.67, .72)	.86 (.71, .83)
Own absolute ability	.71 (.51, .72)	.68 (.48, .69)	.85 (.65, .82)
Others' absolute ability	.74 (.59, .77)	.77 (.60, .76)	.90 (.78, .85)
Desirability	.69 (.56, .68)	.74 (.70, .65)	.77 (.67, .64)
Ambiguity	.70 (.62, .50)	.73 (.64, .60)	.81 (.75, .59)

Note: Reliabilities are Cronbach's α. Reporting structure is the following: full inventory (easy items, difficult items). Reliability met requirements ($\alpha \geq .7$, see Tavakol & Dennick, 2011).

Extension conditions: correlations between comparative ability and domain difficulty ratings for each ability domain

Ability domain	Easy domain condition				Difficult domain condition			
	p	*r*	95% CI		*p*	*r*	95% CI	
			Lower	Upper			Lower	Upper
Using a computer mouse	.647	0.03	-0.10	0.16	<.001	-0.2	-0.37	-0.12
Driving	.082	-0.12	-0.24	0.01	<.001	-0.27	-0.39	-0.15
Riding a bicycle	.071	-0.12	-0.25	0.01	<.001	-0.24	-0.36	-0.12
Saving money	.144	-0.10	-0.23	0.03	<.001	-0.36	-0.47	-0.24
Telling jokes	.935	0.01	-0.13	0.14	<.001	-0.23	-0.35	-0.11
Playing chess	<.001	-0.36	-0.47	-0.24	<.001	-0.27	-0.39	-0.15
Juggling	<.001	-0.26	-0.38	-0.13	<.001	-0.25	-0.37	-0.12
Computer programming	<.002	-0.22	-0.34	-0.09	.002	-0.20	-0.33	-0.08

Mixed Models

Replication Condition

	Comparative Ability Model 1				Comparative Ability Model 2				Comparative Ability Model 3				Comparative Ability Model 4			
Predictors	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	53.29	5.03	43.44 – 63.15	**<0.001**	12.56	1.33	9.95 – 15.18	**<0.001**	12.38	1.97	8.51 – 16.25	**<0.001**	8.72	2.40	4.01 – 13.43	**<0.001**
Own					7.18	0.16	6.86 – 7.50	**<0.001**	7.18	0.17	6.85 – 7.52	**<0.001**	7.07	0.18	6.73 – 7.42	**<0.001**
Other					-0.42	0.21	-0.84 – -0.01	**0.045**	-0.42	0.21	-0.84 – -0.00	**0.047**	-0.48	0.21	-0.90 – -0.06	**0.025**
Difficulty									0.02	0.16	-0.30 – 0.34	0.902	-0.04	0.16	-0.36 – 0.28	0.817
Desirability													0.57	0.21	0.16 – 0.99	**0.007**
Ambiguity													0.12	0.17	-0.21 – 0.45	0.480

Random Effects				
σ^2	407.38	215.06	215.20	214.13
τ_{00}	159.27 $_{ID}$	37.78 $_{ID}$	37.73 $_{ID}$	39.04 $_{ID}$
	195.08 $_{ItemID}$	1.56 $_{ItemID}$	1.58 $_{ItemID}$	1.15 $_{ItemID}$
ICC	0.47	0.15	0.15	0.16
N	240 $_{ID}$	240 $_{ID}$	240 $_{ID}$	240 $_{ID}$
	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$
Observations	1920	1920	1920	1920
Marginal R^2/ Conditional R^2	0.000 / 0.465	0.638 / 0.694	0.638 / 0.694	0.638 / 0.696
AIC	17366.244	15990.513	15994.274	15993.984
log-Likelihood	-8679.122	-7989.257	-7990.137	

Easy Extension

	Comparative Ability Model 1				Comparative Ability Model 2				Comparative Ability Model 3				Comparative Ability Model 4			
Predictors	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	58.86	3.48	52.04 – 65.68	**<0.001**	15.48	1.30	12.94 – 18.02	**<0.001**	18.89	1.92	15.13 – 22.66	**<0.001**	18.60	2.35	13.99 – 23.21	**<0.001**
Own					6.56	0.16	6.25 – 6.88	**<0.001**	6.41	0.17	6.08 – 6.75	**<0.001**	6.37	0.18	6.02 – 6.71	**<0.001**
Other					-0.04	0.20	-0.43 – 0.36	0.861	-0.08	0.21	-0.48 – 0.32	0.685	-0.13	0.21	-0.54 – 0.28	0.546
Difficulty									-0.40	0.15	-0.70 – -0.10	**0.009**	-0.41	0.16	-0.72 – -0.11	**0.008**
Desirability													0.15	0.21	-0.26 – 0.56	0.468
Ambiguity													-0.10	0.19	-0.47 – 0.27	0.600

Random Effects

σ^2	414.81	204.40	201.63	201.51
τ_{00}	170.62 $_{ID}$	52.64 $_{ID}$	57.98 $_{ID}$	58.80 $_{ID}$
	88.95 $_{ItemID}$	0.36 $_{ItemID}$	1.05 $_{ItemID}$	1.01 $_{ItemID}$
ICC	0.38	0.21	0.23	0.23
N	225 $_{ID}$	225 $_{ID}$	225 $_{ID}$	225 $_{ID}$
	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$
Observations	1800	1800	1800	1800
Marginal R^2/ Conditional R^2	0.000 / 0.385	0.590 / 0.674	0.586 / 0.680	0.585 / 0.680
AIC	16318.406	14950.165	14948.412	14954.319
log-Likelihood	-8155.203	-7469.082	-7467.206	-7468.159

Difficult Extension

					Compare				Compare				Compare			
Predictors	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	46.97	3.17	40.74 – 53.19	**<0.001**	16.53	1.49	13.61 – 19.44	**<0.001**	25.92	2.32	21.38 – 30.46	**<0.001**	25.57	2.66	20.36 – 30.79	**<0.001**
Own					6.40	0.16	6.09 – 6.72	**<0.001**	6.14	0.17	5.81 – 6.47	**<0.001**	6.11	0.17	5.78 – 6.45	**<0.001**
Other					-0.02	0.22	-0.44 – 0.41	0.940	-0.15	0.22	-0.58 – 0.27	0.478	-0.16	0.22	-0.59 – 0.26	0.451
Difficulty									-1.01	0.20	-1.40 – -0.62	**<0.001**	-1.04	0.20	-1.43 – -0.64	**<0.001**
Desirability													0.16	0.20	-0.22 – 0.55	0.405
Ambiguity													-0.17	0.19	-0.55 – 0.21	0.370

Random Effects				
σ^2	399.17	227.50	222.88	223.05
τ_{00}	305.86 $_{ID}$	71.00 $_{ID}$	76.39 $_{ID}$	75.65 $_{ID}$
	68.00 $_{ItemID}$	5.13 $_{ItemID}$	3.51 $_{ItemID}$	3.88 $_{ItemID}$
ICC	0.48	0.25	0.26	0.26
N	226 $_{ID}$	226 $_{ID}$	226 $_{ID}$	226 $_{ID}$
	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$	8 $_{ItemID}$
Observations	1808	1808	1808	1808
Marginal R^2/ Conditional R^2	0.000 / 0.484	0.558 / 0.669	0.556 / 0.673	0.556 / 0.673
AIC	16434.072	15249.889	15228.477	15233.599
log-Likelihood	-8213.036	-7618.945	-7607.238	-7607.800

References

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge. doi: https://doi.org/10.4324/9780203771587

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. Download PDF

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77*(2), 221–232. https://doi.org/10.1037/0022-3514.77.2.221

Lakens, D., Caldwell, A. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, **4**(1), 251524592095150. doi: 10.1177/2515245920951503

Lenhard, W., & Lenhard, A. (2016). *Calculation of Effect Sizes*. Retrieved from: https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): Psychometrica. DOI: 10.13140/RG.2.2.17823.92329

The Pennsylvania State University. (2020). *6.3 - Testing for Partial Correlation: STAT 505.* Retrieved from https://online.stat.psu.edu/stat505/lesson/6/6.3

Kruger (1999): Replication and extensions Supplementary

Contents

Open Science disclosures

Main manuscript, data and code

The manuscript, data and code are shared using the Open Science Framework. Review link for data and code of the study: osf.io/7yfkc

Final pre-registration is on: https://osf.io/byx4z
(there was one previous pre-registration, 10 minutes before the final one. Both were conducted before data collection, and the second one simply fixed a minor glitch regarding a default set in one of the comparative questions. Previous pre-registration is on: https://osf.io/nm568/

Procedure and data disclosures

Data collection

Data collection was completed before analysing the data.

Conditions reporting

All collected conditions are reported.

Data exclusions

Details are reported in the materials section of this document

Variables reporting

All variables collected for this study are reported and included in the provided data.

Analysis of the original article

Original article methods

Type of study

The original article was a comparative study.

Experimental design

The study was a within-subject design with 1 independent and 6 dependent variables.

Independent variables (IV)

Independent variable 1: ability domains - easy VS difficult domains

- Participants were assigned to a total of 8 ability domains. The 4 easy domains were using mouse, driving, riding bicycle, and saving money; the 4 difficult domains were telling jokes, playing chess, juggling, and computer programming.

Dependent variables

The original study had 6 dependent variables, all of which were participant ratings in response to each of the 8 ability domains. For each ability domain, the order in which the dependent variables were presented were counterbalanced across participants.

Dependent variable 1: comparative ability rating

- For each ability, participants compared themselves to other students from their course by writing down a percentile ranging from 0 (I'm at the very bottom) to 50 (I'm exactly average) to 100 (I'm at the very top). The exact wording of the question was not provided.

Dependent variable 2: own absolute ability rating

- Participants' estimates of their own absolute ability on a scale from 1 (very unskilled) to 10 (very skilled) for each ability domain.

Dependent variable 3: peers' absolute ability rating

- Participants' estimates of their peers' absolute ability on a scale from 1 (very unskilled) to 10 (very skilled) for each ability domain.

For dependent variables 2 and 3, the judgmental weight of own absolute ability and peers' absolute ability ratings in predicting comparative ability ratings (dependent variable) were measured. A series of multiple regressions were conducted to test how well the two estimates predict comparative ability ratings.

Dependent variable 4: desirability rating

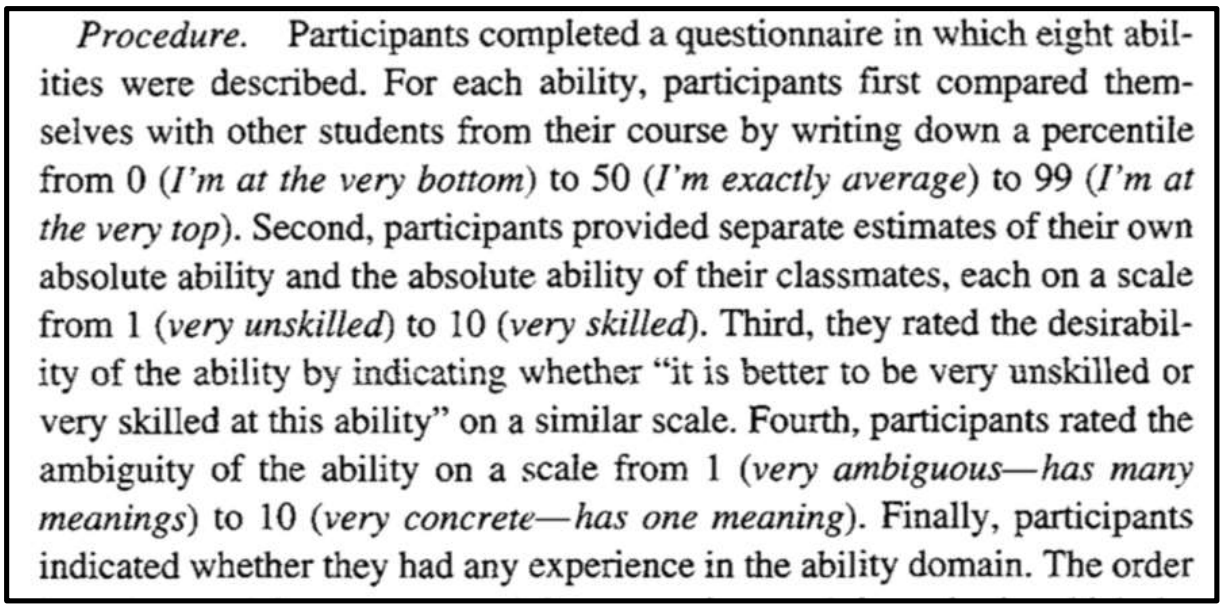
- For each ability, participants indicated whether "it is better to be very unskilled or very skilled at this ability." The rating ranged from a scale of 1 (very unskilled) to 10 (very skilled).

Dependent variable 5: ambiguity rating

- For each ability, participants indicated the ambiguity of the ability. The rating ranged from a scale of 1 (very ambiguous - has many meanings) to 10 (very concrete - has one meaning). The exact wording of the question was not provided.

Dependent variable 6: experience in the ability domain

- For each ability, participants indicated whether they had any experience in the domain. The exact wording of the question and the scale used to measure this indication were not provided.



Procedure. Participants completed a questionnaire in which eight abilities were described. For each ability, participants first compared themselves with other students from their course by writing down a percentile from 0 (*I'm at the very bottom*) to 50 (*I'm exactly average*) to 99 (*I'm at the very top*). Second, participants provided separate estimates of their own absolute ability and the absolute ability of their classmates, each on a scale from 1 (*very unskilled*) to 10 (*very skilled*). Third, they rated the desirability of the ability by indicating whether "it is better to be very unskilled or very skilled at this ability" on a similar scale. Fourth, participants rated the ambiguity of the ability on a scale from 1 (*very ambiguous—has many meanings*) to 10 (*very concrete—has one meaning*). Finally, participants indicated whether they had any experience in the ability domain. The order

Screenshot 1. Screenshot of original study on independent variables 2, 3, and all dependent variables

Original article results

Sample size before and after exclusions

The original study sample size was 37. In addition to the original study's sample of participants, Kruger also conducted two separate tests. Prior to conducting the original study, a pretest of 39 participants was conducted. Pretest participants provided domain difficulty ratings to determine the level of difficulty (easy or difficult) of the 8 ability domains used in the original study. To confirm it was domain difficulty, instead of the rarity or importance of the abilities, that led to the original study's results, another test of 48 Cornell undergraduates was conducted. For the three samples used, exclusions were not specified.

The age of the original sample was not specified in the original study. For gender, the sample consisted of 8 males and 29 females. Participants were university students from Cornell University enrolled in an introductory psychology course, and the study took place in New York, the United States.

Relationship between absolute and comparative ability

Table 1 provides a summary of the original study correlations and partial correlations. The correlation between participants' comparative ability ratings and their own absolute ability ratings was highly significant at $r = .95$, $p < .001$, 95% CI [0.90, 0.97]. For each of the 8 abilities, a series of multiple regressions predicting comparative ability ratings from participant ratings of their own and their peers' absolute ability were conducted. Participant ratings of their own absolute ability, rather than ratings of their peers' absolute ability, better predicted comparative ability ratings ($p < .01$ for all abilities).

Table 1
Correlations and Partial correlations between comparative ability judgments and ratings across all abilities

Control variable	Variable	Comparative ability
	Domain difficulty	-.96***
	Own absolute ability	.95***
	Desirability	.77*
Desirability	Domain difficulty	-.93**
Ambiguity	Domain difficulty	-.97***

*$p < .05$. **$p < .01$. *** $p<.001$

Relationship between domain difficulty and comparative ability

In the original study's pretest, 39 participants provided domain difficulty ratings for each of the 8 abilities. In study one, 37 participants then provided a series of ratings on comparative ability, own and others' absolute abilities, desirability, and ambiguity for each of the 8 abilities. As shown in Table 1, the correlation between participants' comparative ability ratings and pretest domain difficulty was highly significant at $r = -0.96$, $p < .001$, 95% CI [-0.98, -0.92]. Higher domain difficulty scores were significantly correlated with lower percentiles of participants' comparative ability judgments.

To test whether traditional motivational accounts hold true, correlational and partial correlational studies were conducted to determine if the observed results can be explained by factors of desirability or ambiguity. For desirability, the correlation between desirability and comparative ability estimates was significant, $r = .77$, $p < .025$, 95% CI [0.60, 0.87]. The positive correlation shows that participants consider themselves to be below average even in difficult domains that were high in desirability. The partial correlation between comparative ability and domain difficulty while holding desirability constant was also highly significant, $r = -.93$, $p < .01$, 95% CI [-0.96, -0.870]. Similarly, the partial correlation between comparative ability and domain difficulty while holding ambiguity constant was highly significant, $r = -.97$, $p < .001$, 95% CI [-0.98, -0.94]. Both partial correlations demonstrate that the results cannot be explained by desirability or ambiguity, rejecting the motivational account.

One sample experiment

For each ability, one sample t-tests were conducted to compare comparative ability estimates with the midpoint on the scale (50th percentile) ranging from 0 to 99. Participant comparative judgments were compared to the midpoint of the scale indicating "same amount of ability." An above average effect would be above the 50th percentile, whereas a below average effect would be below the 50th percentile. One sample T-tests were also used for comparing the

mean desirability ratings for easy and difficult abilities respectively, measuring whether it was better to be skilled or unskilled at the abilities. The mean desirability ratings were compared to the midpoint of the scale indicating "same amount of desirability."

For mean comparative ability ratings comparing own ability to other students' ability, percentile ratings were given on a scale from 0 to 99. The sample size was 37. The mean, degrees of freedom, and p value for each ability are listed below:

1. Mean:

Mean of easy abilities:

- Using a mouse: 58.8
- Driving: 65.4
- Riding bicycle: 64.0
- Saving money: 61.5

Mean of difficult abilities:

- Telling jokes: 46.4
- Playing chess: 27.8
- Juggling: 26.5
- Computer programming: 24.8

2. The reported degrees of freedom of the mean comparative ratings for each ability was 36.

3. The reported p-values:

- Across easy abilities: $p < .01$
- Across difficult abilities: 3 less than $p < .0001$, 1 not significant (specific p=value unreported)

4. The reported t-statistic:

- Easy and difficult abilities: unreported

For mean desirability ratings comparing own ability to other students' ability, ratings were given on a scale from 1 to 10. The sample size was 37. The mean, degrees of freedom, and p value for the easy and difficult abilities are listed below:

1. Mean

- Desirability rating for easy abilities: 7.6
- Desirability rating for difficult abilities: unspecified, exceeded midpoint by more than a full scale point

2. The reported degrees of freedom of the mean desirability ratings for the easy and difficult abilities was 36.

3. The reported p-values

- Desirability rating for easy abilities: $p < .0001$
- Desirability rating for difficult abilities: $p < .0001$

4. The reported t-statistic

- Desirability rating for easy abilities: $t = 13.51$
- Desirability rating for difficult abilities: $t = 5.06$

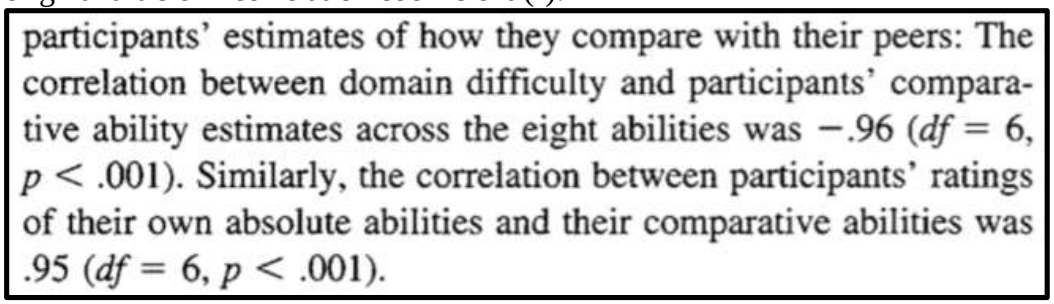
For all four easy abilities, participants gave significantly higher comparative ability estimates ($p < .01$) than the midpoint. For three of the four difficult abilities, participants gave significantly lower comparative ability estimates ($p < .0001$) than the midpoint. The results thus

support the above average effect for easy ability domains and the below average effect for difficult ability domains.

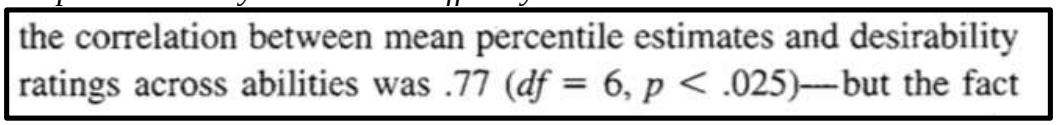
Effect size calculations of the original study effects

Effect size and confidence intervals for correlations and partial correlations

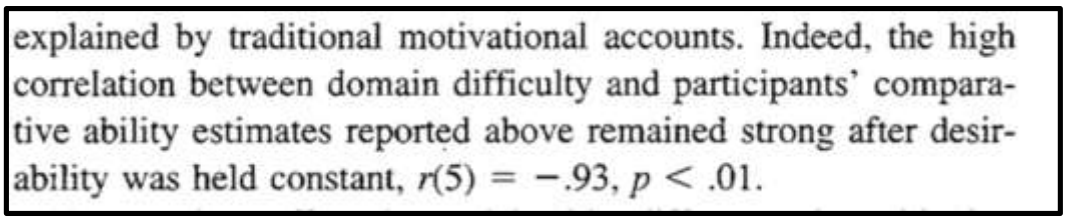
All correlation effect sizes in correlation coefficient (r) are summarised in Table 1 under the section ""relationship between absolute and comparative ability." The effects sizes of the correlations and partial correlations were also reported in p.224, under the "results" section of the original article in correlation coefficient (r):


participants' estimates of how they compare with their peers: The correlation between domain difficulty and participants' comparative ability estimates across the eight abilities was −.96 ($df = 6$, $p < .001$). Similarly, the correlation between participants' ratings of their own absolute abilities and their comparative abilities was .95 ($df = 6$, $p < .001$).

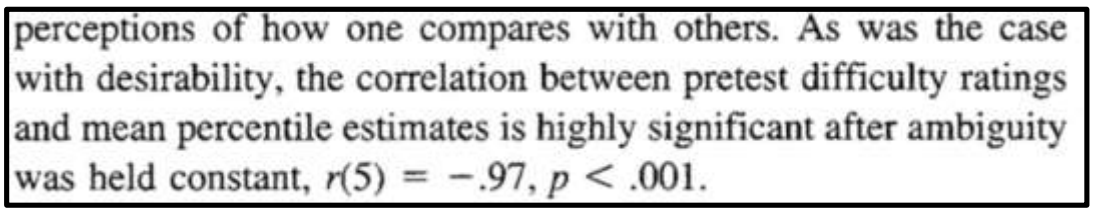
Screenshot 2.1 Screenshot of original study on the effect sizes of the correlations between comparative ability and domain difficulty or own absolute abilities.


the correlation between mean percentile estimates and desirability ratings across abilities was .77 ($df = 6$, $p < .025$)—but the fact

Screenshot 2.2 Screenshot of original study on the effect sizes of correlation between comparative ability and desirability ratings.


explained by traditional motivational accounts. Indeed, the high correlation between domain difficulty and participants' comparative ability estimates reported above remained strong after desirability was held constant, $r(5) = -.93$, $p < .01$.

Screenshot 2.3 Screenshot of original study on the effect sizes of correlation between domain difficulty and comparative ability, holding desirability constant.


perceptions of how one compares with others. As was the case with desirability, the correlation between pretest difficulty ratings and mean percentile estimates is highly significant after ambiguity was held constant, $r(5) = -.97$, $p < .001$.

Screenshot 2.4 Screenshot of original study on the effect sizes of correlation between domain difficulty and comparative ability, holding ambiguity constant.

The R code used to calculate the confidence intervals for correlations:

1. CIr(r, n, level = 0.95)
   Where r = correlation coefficient, n = sample size, level = significance level for constructing the CI

For correlations:

Formula: CIr(r, n, level = 0.95)

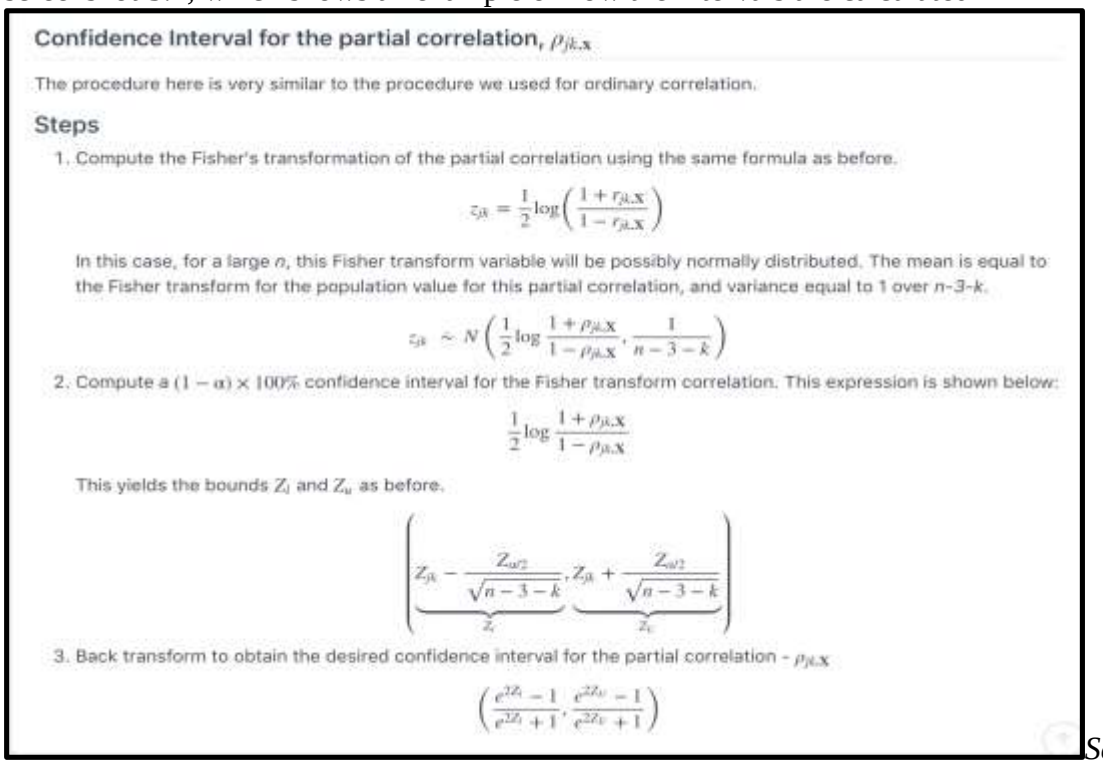
1. Domain difficulty and comparative ability: r = -.96, n = 39
   r = -.96, 95% CI [-0.98, -0.93]
2. Own absolute ability and comparative ability: r = .95, n = 37
   r = .95, 95% CI [0.90, 0.97]
3. Others' absolute ability and comparative ability: r = .77, n = 37
   r = .77, 95% CI [0.59, 0.88]

The input and calculated confidence intervals for correlations and partial correlations:

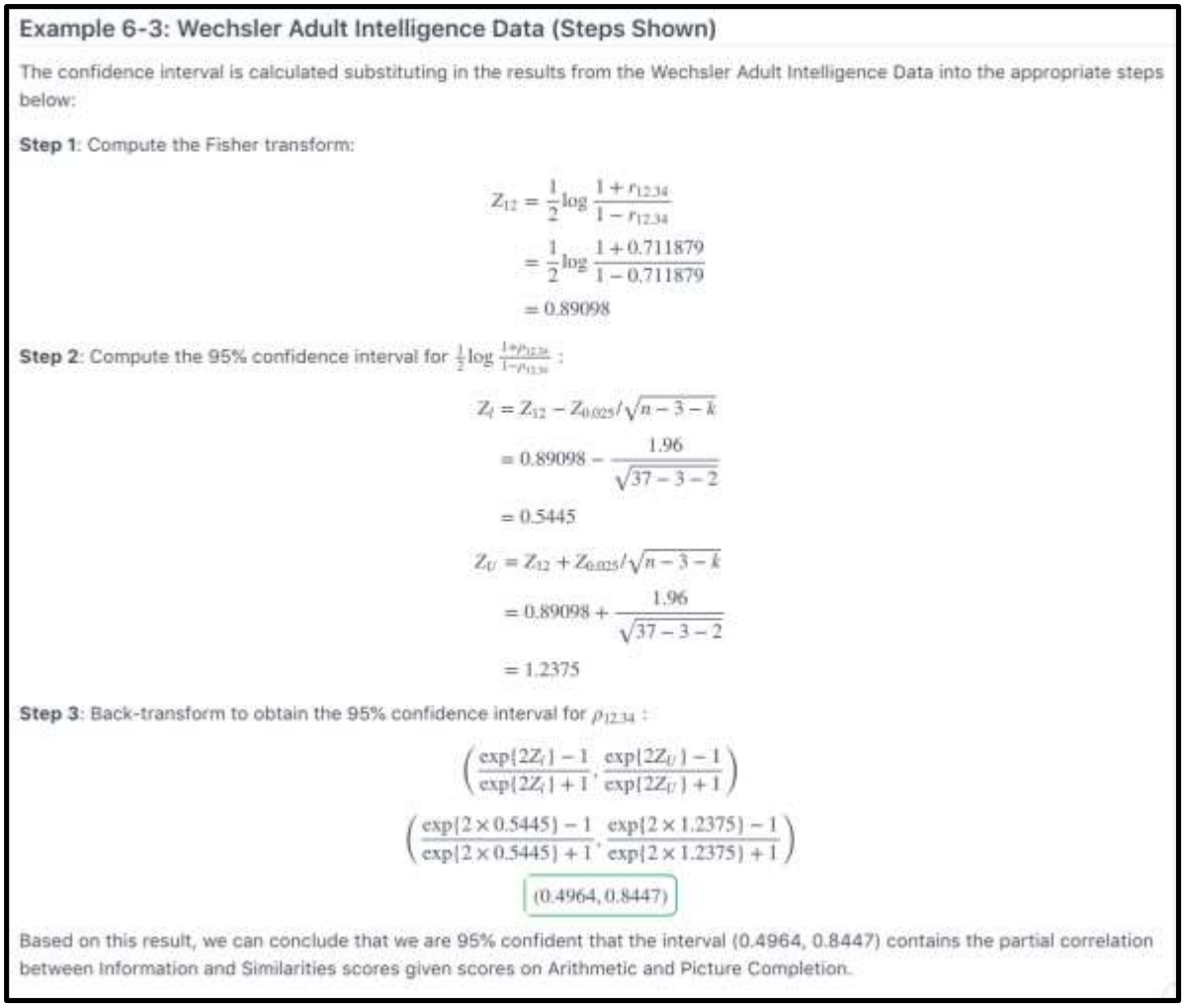
```
> #correlation between domain difficulty and comparative ability across abilities
> CIr(r=-.96, n = 39, level = .95)
[1] -0.9789858 -0.9245153
> #correlation between own difficulty and comparative ability across abilities
> CIr(r=.95, n = 37, level = .95)
[1] 0.9043590 0.9741562
> #correlation between desirability and comparative ability ratings across abilities
> CIr(r=.77, n = 37, level = .95)
[1] 0.5942409 0.8755692
```

The formula used to calculate the confidence intervals for partial correlations:

The formula used to calculate the confidence intervals for partial correlations was referenced from https://online.stat.psu.edu/stat505/lesson/6/6.3 (The Pennsylvania State University, 2020). Screenshots 3.6 and 3.7 show the steps of first computing the Fisher's transformation and obtaining the intervals from the Fisher's transformation correlation. The current partial correlation confidence interval calculations will follow the steps listed in screenshot 3.7, which shows an example of how the intervals are calculated



Screenshot 2.5. Screenshot showing the formula for partial correlation confidence intervals.



Example 6-3: Wechsler Adult Intelligence Data (Steps Shown)

The confidence interval is calculated substituting in the results from the Wechsler Adult Intelligence Data into the appropriate steps below:

Step 1: Compute the Fisher transform:

$$\begin{aligned} Z_{12} &= \frac{1}{2}\log\frac{1+r_{12.34}}{1-r_{12.34}} \\ &= \frac{1}{2}\log\frac{1+0.711879}{1-0.711879} \\ &= 0.89098 \end{aligned}$$

Step 2: Compute the 95% confidence interval for $\frac{1}{2}\log\frac{1+\rho_{12.34}}{1-\rho_{12.34}}$:

$$\begin{aligned} Z_l &= Z_{12} - Z_{0.025}/\sqrt{n-3-k} \\ &= 0.89098 - \frac{1.96}{\sqrt{37-3-2}} \\ &= 0.5445 \end{aligned}$$

$$\begin{aligned} Z_U &= Z_{12} + Z_{0.025}/\sqrt{n-3-k} \\ &= 0.89098 + \frac{1.96}{\sqrt{37-3-2}} \\ &= 1.2375 \end{aligned}$$

Step 3: Back-transform to obtain the 95% confidence interval for $\rho_{12.34}$:

$$\left(\frac{\exp\{2Z_l\}-1}{\exp\{2Z_l\}+1}, \frac{\exp\{2Z_U\}-1}{\exp\{2Z_U\}+1}\right)$$

$$\left(\frac{\exp\{2\times 0.5445\}-1}{\exp\{2\times 0.5445\}+1}, \frac{\exp\{2\times 1.2375\}-1}{\exp\{2\times 1.2375\}+1}\right)$$

$$(0.4964, 0.8447)$$

Based on this result, we can conclude that we are 95% confident that the interval (0.4964, 0.8447) contains the partial correlation between Information and Similarities scores given scores on Arithmetic and Picture Completion.

Screenshot 2.6. Screenshot showing an example of computing the partial correlation confidence intervals.

Following the three steps in screenshot 3.7, the fisher z-transformation is conducted in step one. For step one, the fisher z-transformation is first calculated in r. The calculated z-scores are then applied into steps two and three to calculate the confidence intervals.

Step one (*the fisher z-transformation from r to z-score)*:

```
> #correlation between domain difficulty and comparative ability, holding desirability
constant
> fisherz(-.93)
[1] -1.65839
> #correlation between domain difficulty and comparative ability, holding ambiguity
constant
> fisherz(-.97)
[1] -2.092296
```

For the partial correlations between domain difficulty and comparative ability, holding desirability constant:

Step two:

$$\left(\underbrace{Z_{jk} - \frac{Z_{\alpha/2}}{\sqrt{n-3-k}}}_{Z_l}, \underbrace{Z_{jk} + \frac{Z_{\alpha/2}}{\sqrt{n-3-k}}}_{Z_U} \right)$$

Z_I= -1.65839 - (1.96 / (√ 39 – 3 – 2))
=-1.994527
Z_U = -1.65839 + (1.96 / (√ 39 – 3 – 2))
= -1.322253

Step three:

$$\left(\frac{e^{2Z_l} - 1}{e^{2Z_l} + 1}, \frac{e^{2Z_U} - 1}{e^{2Z_U} + 1} \right)$$

Z_I = Exp (2 x -1.994527) -1 / Exp (2 x -1.994527) +1 = -0.9636389
Z_U = Exp (2 x -1.322253) -1 / Exp (2 x -1.322253) +1 = -0.8673431
r = -.93, 95% CI [-0.964, -0.867]

For the partial correlations between domain difficulty and comparative ability, holding ambiguity constant:

Step two:

$$\left(\underbrace{Z_{jk} - \frac{Z_{\alpha/2}}{\sqrt{n-3-k}}}_{Z_l}, \underbrace{Z_{jk} + \frac{Z_{\alpha/2}}{\sqrt{n-3-k}}}_{Z_U} \right)$$

Z_I= -2.092296 - (1.96 / (√ 39 – 3 – 2))
=-2.428433
Z_U = -2.092296 + (1.96 / (√ 39 – 3 – 2))
= -1.756159

Step three:

$$\left(\frac{e^{2Z_l} - 1}{e^{2Z_l} + 1}, \frac{e^{2Z_U} - 1}{e^{2Z_U} + 1} \right)$$

Z_I = Exp (2 x -2.428433) -1 / Exp (2 x -2.428433) +1 = =0.9845703
Z_U = Exp (2 x -1.756159) -1 / Exp (2 x -1.756159) +1 = -0.9420725
r = -.97, 95% CI [-0.985, -0.942]

Effect size and confidence intervals for one sample experiment

For the one sample experiments regarding participant comparative ability judgments, the t-statistic, standard deviation, and standard error were not reported in the original study. The esc_t package in R was used to compute the effect sizes.

The R code used to calculate the effects:

1. esc_t(p = X, totaln = X)
   Where p = the p-value of the t-test, total n = total sample size.
2. ES.t.one(t = X,df = X)

Where t = t-statistic, df = degree of freedom

The R code used to calculate the confidence intervals:

1. cohen.d.ci(d=X, n1 = X, alpha = X)
   Where d = Cohen's d statistic, n1 = sample size, alpha = 1-alpha is the width of the confidence interval

For comparative ability judgments for each of the 4 easy ability domains:

Formula: esc_t(p = X, totaln = X)

p < .01, total n = 37

Cohen's d = 0.896, 95% CI [0.22, 1.57]

The calculated effect size and confidence intervals for comparative ability judgments across the 4 easy ability domains:

```
> #effect size of comparative ability ratings for the easy abilities
> esc_t(p = 0.01, totaln = 37)

Effect Size Calculation for Meta Analysis
     Conversion: t-value to effect size d
    Effect Size:   0.8956
  Standard Error:   0.3449
       Variance:   0.1189
       Lower CI:   0.2196
       Upper CI:   1.5715
         Weight:   8.4071
```

For comparative ability judgments for each of the 4 difficult ability domains (excluding telling jokes):

Formula: esc_t(p = X, totaln = X)

p < .0001, total n = 37

Cohen's d = 1.443, 95% CI [0.72, 2.17]

The calculated effect size and confidence intervals for comparative ability judgments across the 4 difficult ability domains:

```
> #effect size of comparative ability ratings for the difficult abilities
> esc_t(p = 0.0001, totaln = 37)

Effect Size Calculation for Meta Analysis


     Conversion: t-value to effect size d
    Effect Size:   1.4430
  Standard Error:   0.3691
       Variance:   0.1362
       Lower CI:   0.7196
       Upper CI:   2.1665
         Weight:   7.3396
```

The original study notes that for 3 of the 4 difficult abilities, the mean percentile estimates were significantly less than 50%. The t-statistic, SE, and SD are unreported for the not

significant ability domain (telling jokes). The effect size and confidence interval calculations thus does not include calculations for the ability domain telling jokes.

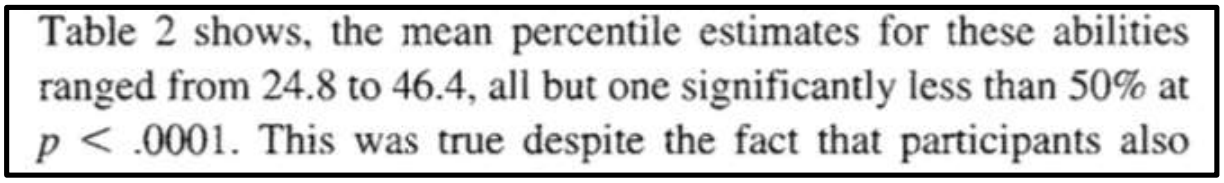


Table 2 shows, the mean percentile estimates for these abilities ranged from 24.8 to 46.4, all but one significantly less than 50% at $p < .0001$. This was true despite the fact that participants also

Screenshot 3. Screenshot of original study on comparative ability judgments across the 4 difficult ability domains.

For mean desirability rating across the 4 easy ability domains:

Formulas:

1. ES.t.one(t = X,df = X)

t = 13.51, df = 36

2. cohen.d.ci(d=X, n1 = X, alpha = X)

d = 2.25, n1 = 37, alpha = .05

Cohen's d = 2.252, 95% CI [1.64, 2.86]

The calculated effect size for mean desirability rating across the 4 easy ability domains:

```
> #effect size of mean desirability rating across the easy abilities
> ES.t.one(t = 13.51, df=36)

    effect size (Cohen's d) of one-sample t test

          d = 2.251667
  alternative = two.sided
```

The calculated confidence intervals for mean desirability rating across the 4 easy ability domains:

```
> #confidence intervals of mean desirability rating across the easy abilities
> cohen.d.ci(d=2.251667, n1 = 37, alpha = .05)
      lower   effect    upper
[1,] 1.636927 2.251667 2.856963
```

For mean desirability rating across the 4 difficult ability domains:

Formula:

1. ES.t.one(t = X,df = X)

t = 5.06, df = 36

2. cohen.d.ci(d=X, n1 = X, alpha = X)

d = 0.84, n1 = 37, alpha = .05

Cohen's d = 0.843, 95% CI [0.46, 1.22]

The calculated effect size for mean desirability rating across the 4 easy ability domains:

```
> #mean desirability rating across the difficult abilities
> ES.t.one(t = 5.06, df=36)

    effect size (Cohen's d) of one-sample t test
```

```
          d = 0.8433333
   alternative = two.sided
```

The calculated confidence intervals for mean desirability rating across the 4 difficult ability domains:

```
> #confidence intervals of mean desirability rating across the difficult abilities
> cohen.d.ci(d=0.8433333, n1 = 37, alpha = .05)
      lower    effect    upper
[1,] 0.4628362 0.8433333 1.215157
```

Effect size and confidence intervals for paired sample t test

The R code used to calculate the effects and confidence intervals:

1. ES.t.paired (t=X ,df=X)
   Where t = t statistic, df = degree of freedom
2. cohen.d.ci(d=0.5666667, n1 = 37, alpha = .05)
   Where d = Cohen's d statistic, n1 = sample size, alpha = 1-alpha is the width of the confidence interval

For ambiguity ratings between easy and difficult abilities:

Formulas:

1. ES.t.paired (t=X ,df=X)
t = 3.40, df = 36
2. cohen.d.ci(d=X, n1 = X, alpha = X)
d = 0.57, n1 = 37, alpha = .05

Cohen's d = 0.57, 95% CI [0.22, 0.91]

The calculated effect size for ambiguity ratings between easy and difficult abilities:

```
> #paired sample t-test
> #ambiguity ratings between easy and difficult abilities
> ES.t.paired(t=3.40,df=36)

    effect size (Cohen's d) of paired two-sample t test

          d = 0.5666667
   alternative = two.sided

NOTE: The alternative hypothesis is md != 0
small effect size:  d = 0.2
medium effect size: d = 0.5
large effect size:  d = 0.8
```

The calculated confidence intervals for ambiguity ratings between easy and difficult abilities:

```
> #confidence intervals of ambiguity ratings between easy and difficult abilities
> cohen.d.ci(d=0.5666667, n1 = 37, alpha = .05)
      lower    effect     upper
[1,] 0.2156665 0.5666667 0.9108903
```

Effect size for multiple regressions

In the original study, table 2 in p. 224 shows the standardised betas for the multiple regressions conducted to predict comparative ability estimates from participant (N = 37) estimates of their own and their peers' absolute abilities. Table 2 below shows the standardised betas for the eight regressions in the columns "own ability" and "Peers' ability."

Of the 8 standardised betas reported, the smallest significant beta was -.25 for the abilities driving and saving money. The effect size for driving / saving money was therefore calculated. Computation of the effect size was first done by converting the standardised β coefficient, -.25, to r. The standardised β coefficient was entered into the table under the "imputation of r from standardised β weights from multiple regression analysis" section in the psychometrica website: *https://www.psychometrica.de/effect_size.html*, adapted from Lenhard & Lenhard (2016). The resulting r value was -.3.

Table 2

Original study results showing the standardised betas from multiple regressions predicting comparative ability judgments from own and peers' absolute abilities

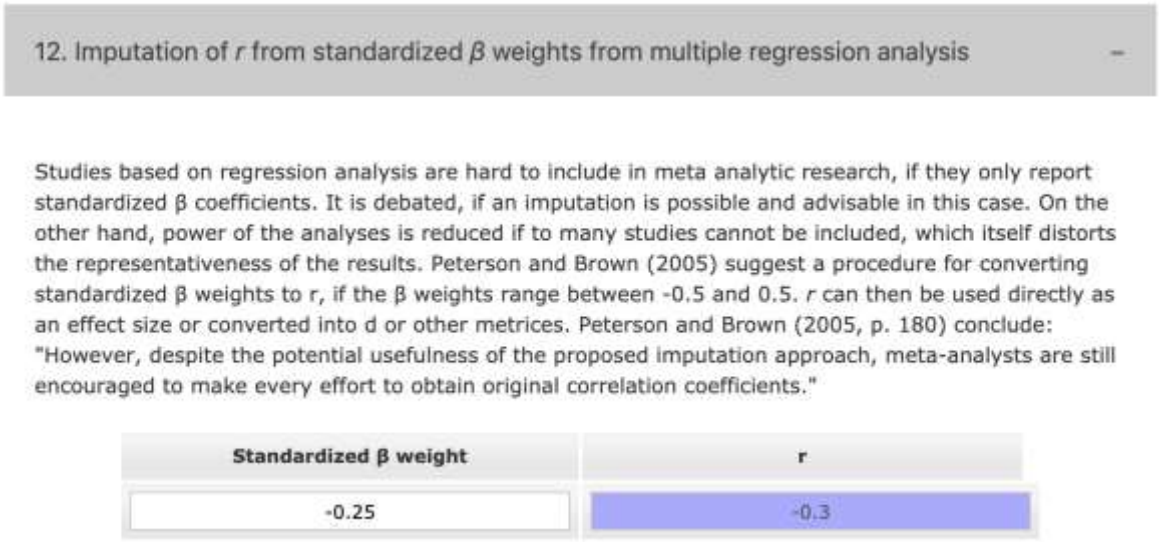
Table 2

Mean Comparative Ability Estimates and Judgmental Weight of Own Versus Peers' Abilities by Domain Difficulty, Study 1 (N = 37)

Ability	Domain difficulty[a]	Percentile estimate[b]	Judgmental weight of	
			Own ability[c]	Peers' ability[c]
Easy				
Using mouse	3.1	58.8**	.21	.06
Driving	3.6	65.4****	.89****	−.25*
Riding bicycle	3.9	64.0****	.61****	−.02
Saving money	4.3	61.5**	.90****	−.25***
Difficult				
Telling jokes	6.1	46.4	.91****	−.03
Playing chess	7.1	27.8****	.96****	−.22**
Juggling	8.3	26.5****	.89****	−.16
Computer programming	8.7	24.8****	.85****	−.10

[a]Higher numbers reflect greater difficulty. [b]Mean percentile estimates above 50 reflect an above-average effect, estimates below 50 reflect a below-average effect. [c]Standardized betas from multiple regressions predicting participants' comparative ability (percentile) estimates from their estimates of their own absolute ability and the absolute ability of their peers, respectively.

*$p < .05$. **$p < .01$. ***$p < .001$. ****$p < .0001$.



Screenshot 4. Screenshot showing input to calculate r from standardised β coefficient in the Psychometrica website: https://www.psychometrica.de/effect_size.html.

The resulting value is r = -.3, and $\mathbf{r^2 = 0.9}$. $r^2 = 0.9$ was then applied into the equation $f^2 = r^2 / (1 - r^2)$ to find the effect size f^2 (Cohen, 2013).

$$f^2 = \frac{r^2}{1 - r^2}$$

Figure 1. f^2 formula. Adapted from 'Statistical Power Analysis for the Behavioral Sciences' by J. Cohen, 2013, *Academic Press*, p. 422.

$$f^2 = r^2 / (1 - r^2)$$
$$f^2 = (0.09) / (1 - (0.09))$$
$$\mathbf{f^2 = 0.099}$$

The resulting value is $f^2 =$ **0.099.**

Confidence intervals for multiple regressions

The R code used to calculate the confidence intervals:

CI.Rsq(rsq, n, k, level = 0.95)

Where rsq = squared multiple correlation, n = sample size, k = number of predictors in model, level = significance level for constructing the CI

For comparative ability estimates from participant estimates of their own and their peers' absolute abilities:

Formula: CI.Rsq(rsq, n, k, level = 0.95)

rsq = 0.09, SErsq = 0.079, n = 37, k =2, level = 0.95

95% CI [-0.07, 0.25]

The input and calculated confidence intervals for multiple regressions:

```
> #multiple regression
> #own ability and others ability predicting comparative ability
> CI.Rsq(0.09, 37, 2, level = 0.95)
  Rsq     SErsq       LCL       UCL
```

1 0.09 0.07935943 -0.06554162 0.2455416

Power analysis of original study effect to assess required sample for replication

Minimum required sample size for replication

All power analysis calculations were done using the G* Power 3.1 program (Buchner et al, 2007) downloaded from the website: http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html. The current replication aimed for a power of .95, and the alpha error probability was set at .05.

Since a total of 11 effect sizes were calculated, 11 power analyses based on the 11 effect sizes were conducted. Of the 11 power analyses, 5 of them tested for hypotheses 1, 2 and 3 of the original study, whereas the other 6 did not test for the main hypotheses. Therefore, the 5 power analyses testing for the original study's hypotheses were used to calculate the minimum required sample size needed for replication.

Table 3 summarises the 11 power analyses and the 5 primary power analyses are labelled with [a]. From table 3, power analyses for hypotheses 1, 2 and 3 showed that the minimum required sample size in this study would be 160 participants. For the current replication, we aimed to collect more than three times the minimum sample size of 500 participants. The following section is the protocol of inputs and outputs from G* Power.

Table 3
Summary table of power analysis

Variables	Required sample size
Correlational study	
Correlation between domain difficulty and comparative ability (2) (3)[a]	6
Correlation between own absolute ability and comparative ability (1) [a]	7
Correlation between others' absolute ability and comparative ability	15
Correlation between domain difficulty and comparative ability, holding desirability constant	8
Correlation between domain difficulty and comparative ability, holding ambiguity constant	7
One sample experiment	
Comparative ability judgments across the 4 easy ability domains (2)[a]	19
Comparative ability judgments across the 4 difficult ability domains (3)[a]	9
Mean desirability ratings across the 4 easy ability domains	5
Mean desirability ratings across the 4 difficult ability domains	21
Paired sample t-test	
Ambiguity ratings between easy and difficult abilities	43
Multiple regression	
Comparative ability estimates from participant estimates of their own and their peers' absolute abilities (driving / saving money) (1) [a]	160

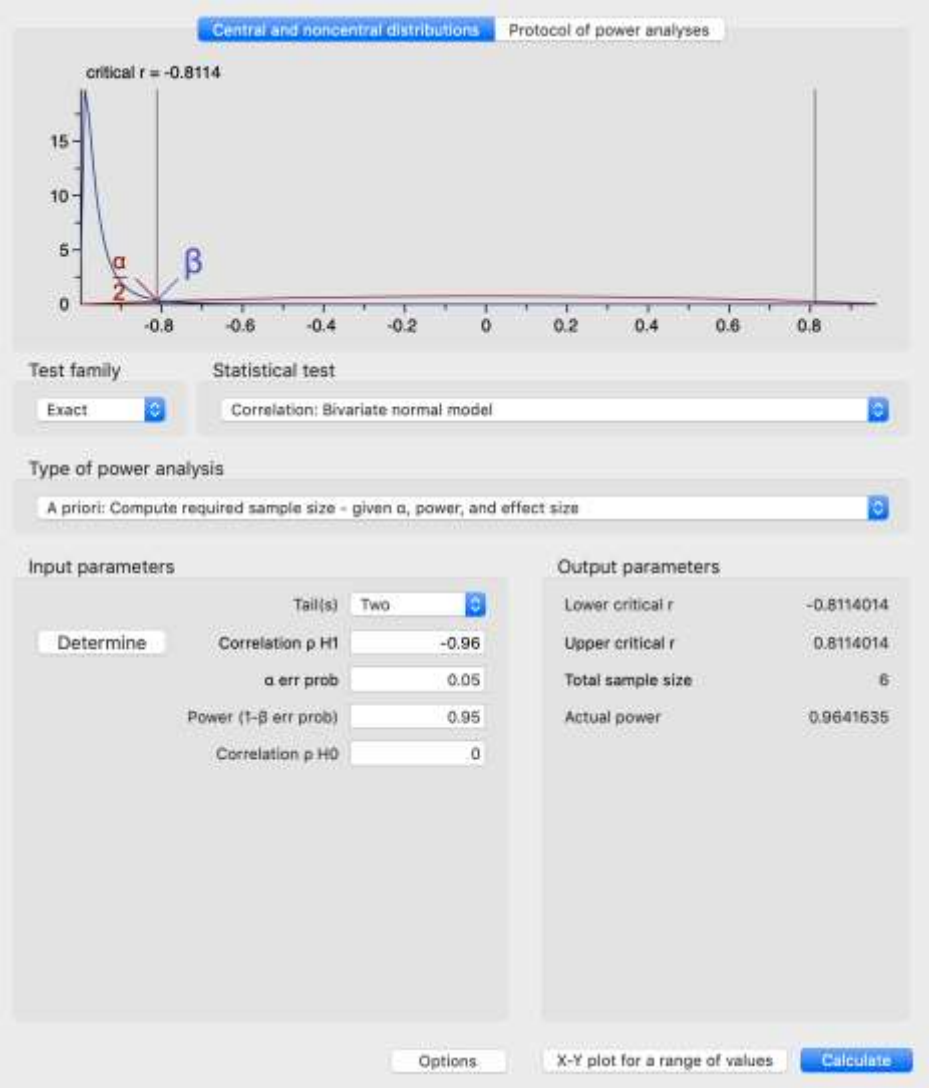
Note. Variables with (1), (2), or (3) are power analysis for hypothesis 1, 2, or 3 respectively. Variables not marked with (1), (2), or (3) do not test for the main effects of interest in this study.

[a]Calculations related to the main effects of interest in the original study.

G* power protocol

Power analysis for correlations

Domain difficulty and comparative ability



Exact - Correlation: Bivariate normal model

Options: exact distribution

Analysis: A priori: Compute required sample size

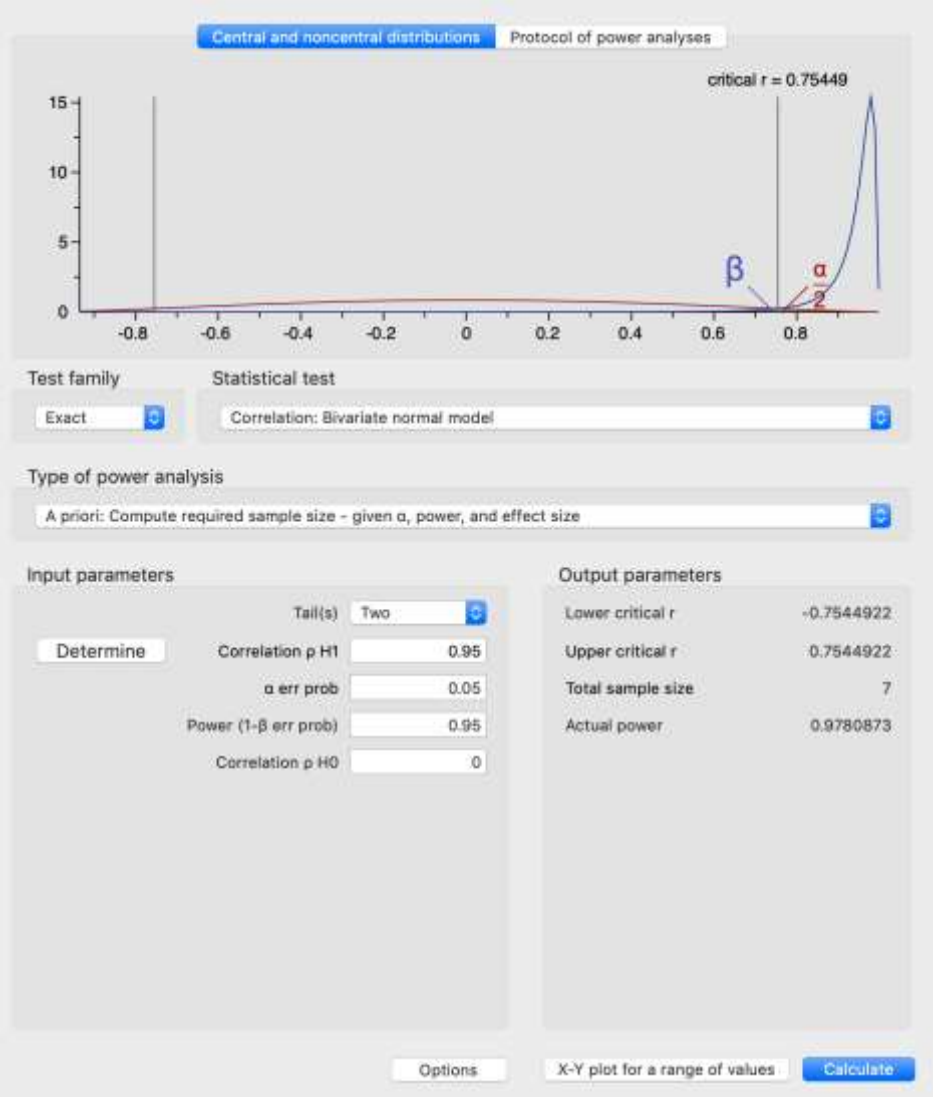
Input: Tail(s) = Two

Correlation ρ H1 = -0.96

α err prob = 0.05

	Power (1-β err prob)	=	0.95
	Correlation ρ H0	=	0
Output:	Lower critical r	=	-0.8114014
	Upper critical r	=	-0.8114014
	Total sample size	=	6
	Actual power	=	0.9641635

Own absolute ability and comparative ability



Exact - Correlation: Bivariate normal model

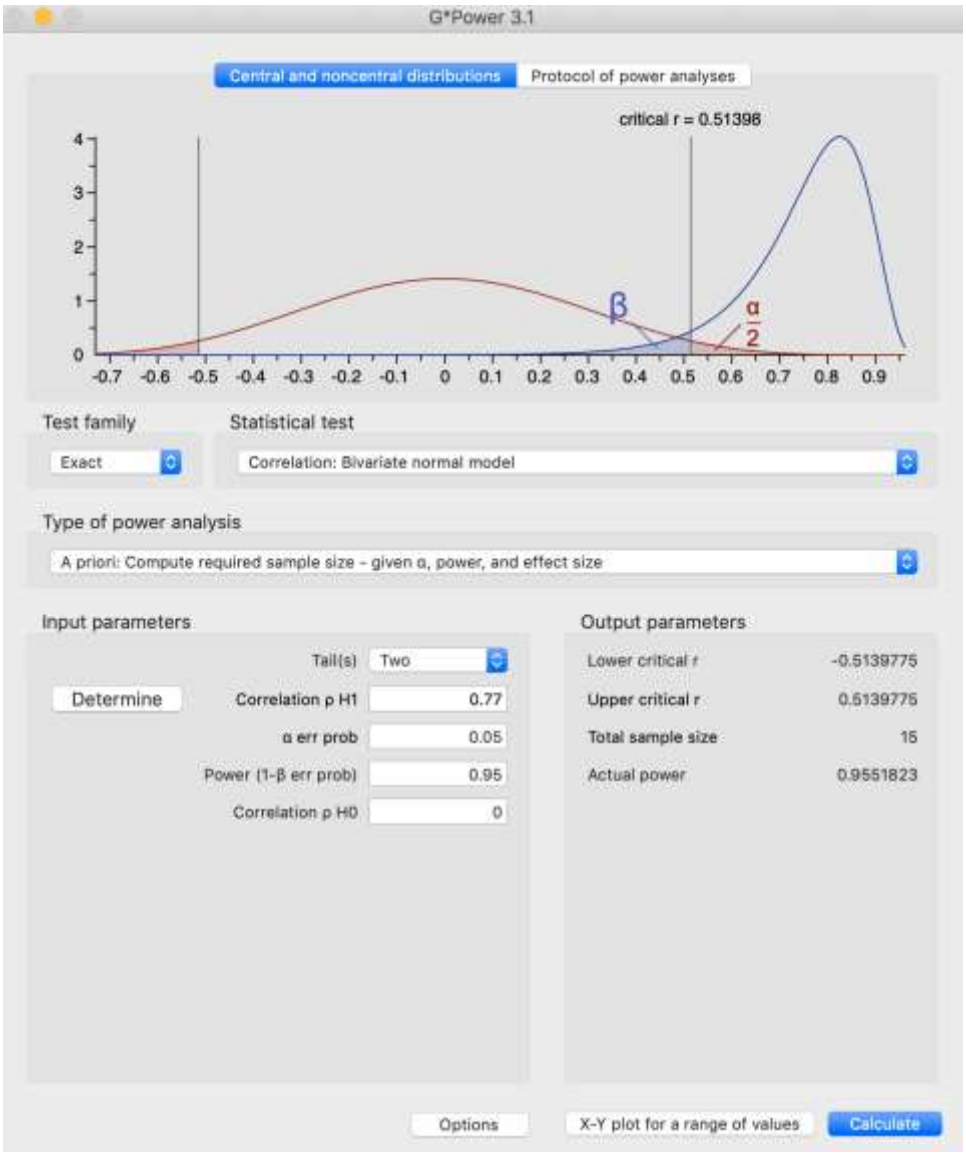
Options: exact distribution

Analysis: A priori: Compute required sample size

Input:	Tail(s)	=	Two

	Correlation ρ H1	=	0.95
	α err prob	=	0.05
	Power (1-β err prob)	=	0.95
	Correlation ρ H0	=	0
Output:	Lower critical r	=	-0.7544922
	Upper critical r	=	0.7544922
	Total sample size	=	7
	Actual power	=	0.9780873

Others' absolute ability and comparative ability



Exact - Correlation: Bivariate normal model

Options: exact distribution

Analysis:	A priori: Compute required sample size		
Input:	Tail(s)	=	Two
	Correlation ρ H1	=	0.77
	α err prob	=	0.05
	Power (1-β err prob)	=	0.95
	Correlation ρ H0	=	0
Output:	Lower critical r	=	-0.5139775
	Upper critical r	=	0.5139775
	Total sample size	=	15
	Actual power	=	0.9551823

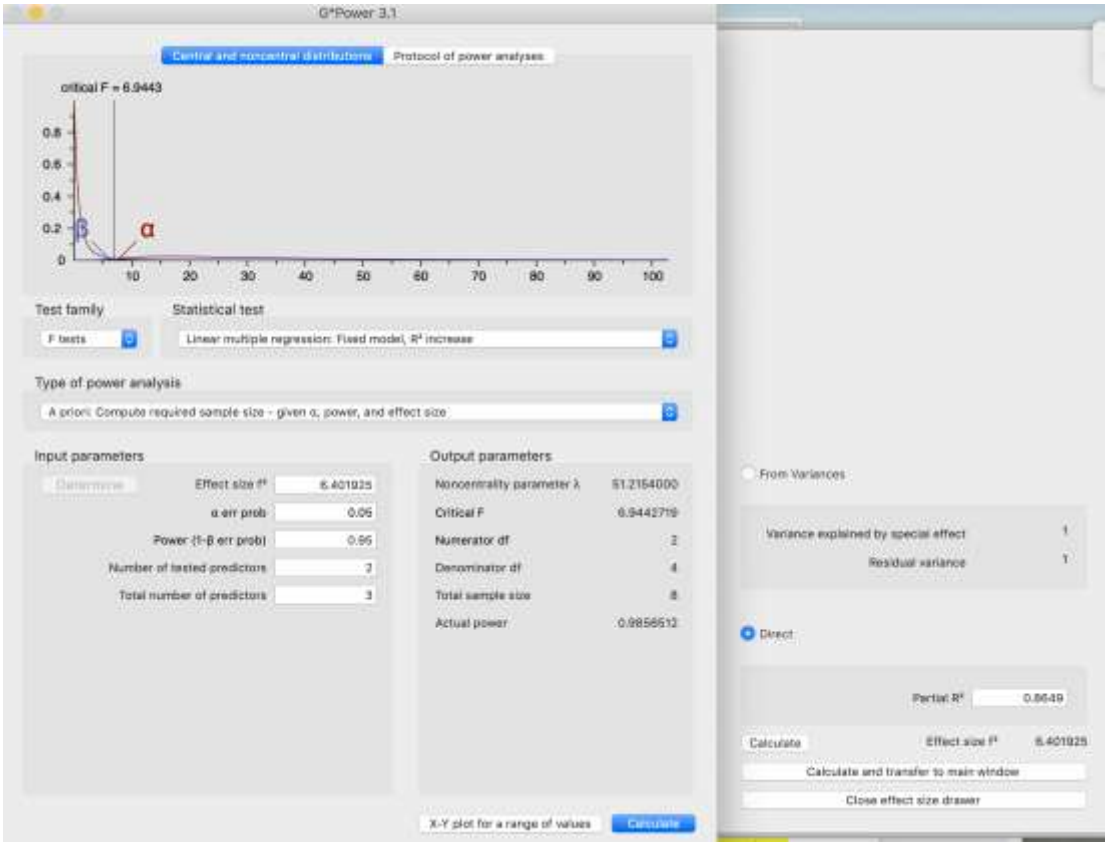
Power analysis for partial correlations

The partial correlation between domain difficulty and comparative ability, holding desirability constant

Partial correlation = -0.93

$R2 = (-0.93)^2$

=0.8649



F tests - Linear multiple regression: Fixed model, R^2 increase

Analysis:	A priori: Compute required sample size		
Input:	Effect size f^2	=	6.401925
	α err prob	=	0.05

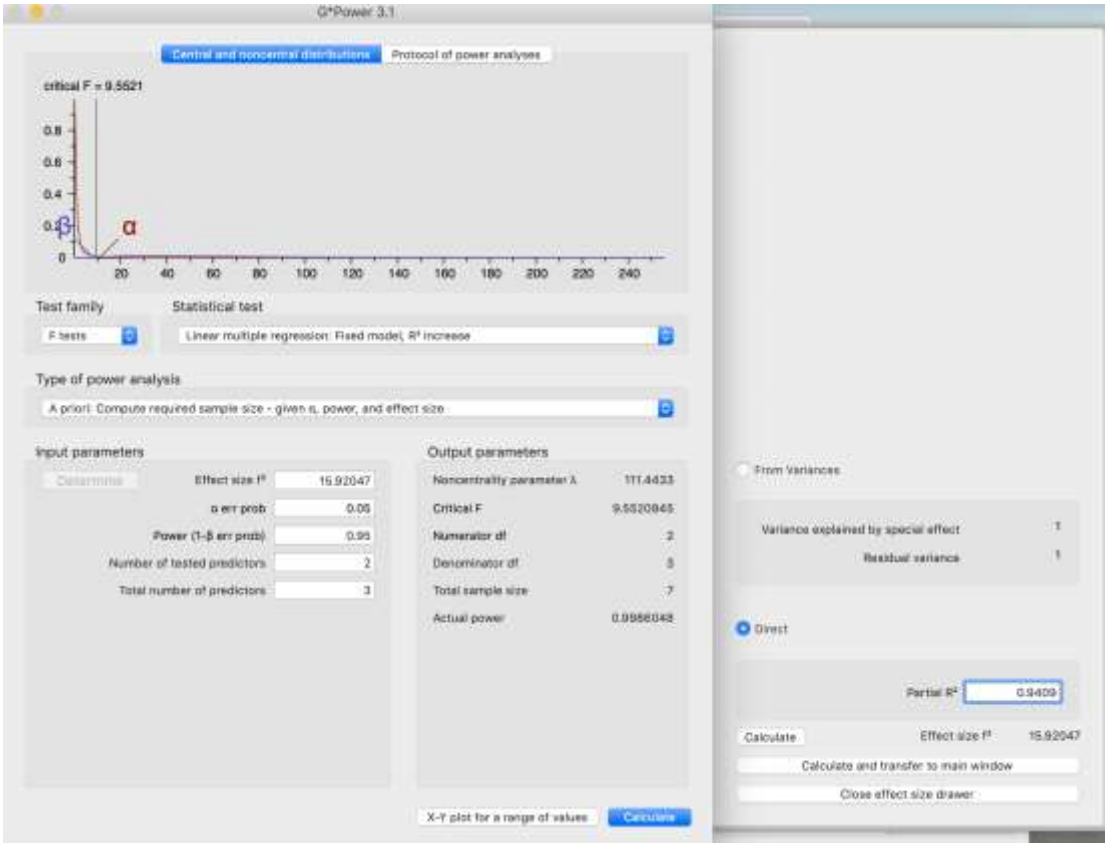
	Power (1-β err prob)	=	0.95	
	Number of tested predictors		=	2
	Total number of predictors		=	3
Output:	Noncentrality parameter λ		=	51.2154000
	Critical F	=	6.9442719	
	Numerator df	=	2	
	Denominator df	=	4	
	Total sample size	=	8	
	Actual power	=	0.9856512	

The partial correlation between domain difficulty and comparative ability, holding ambiguity constant

Partial correlation = -0.97

$R2 = (-0.97)^2$

=0.9409



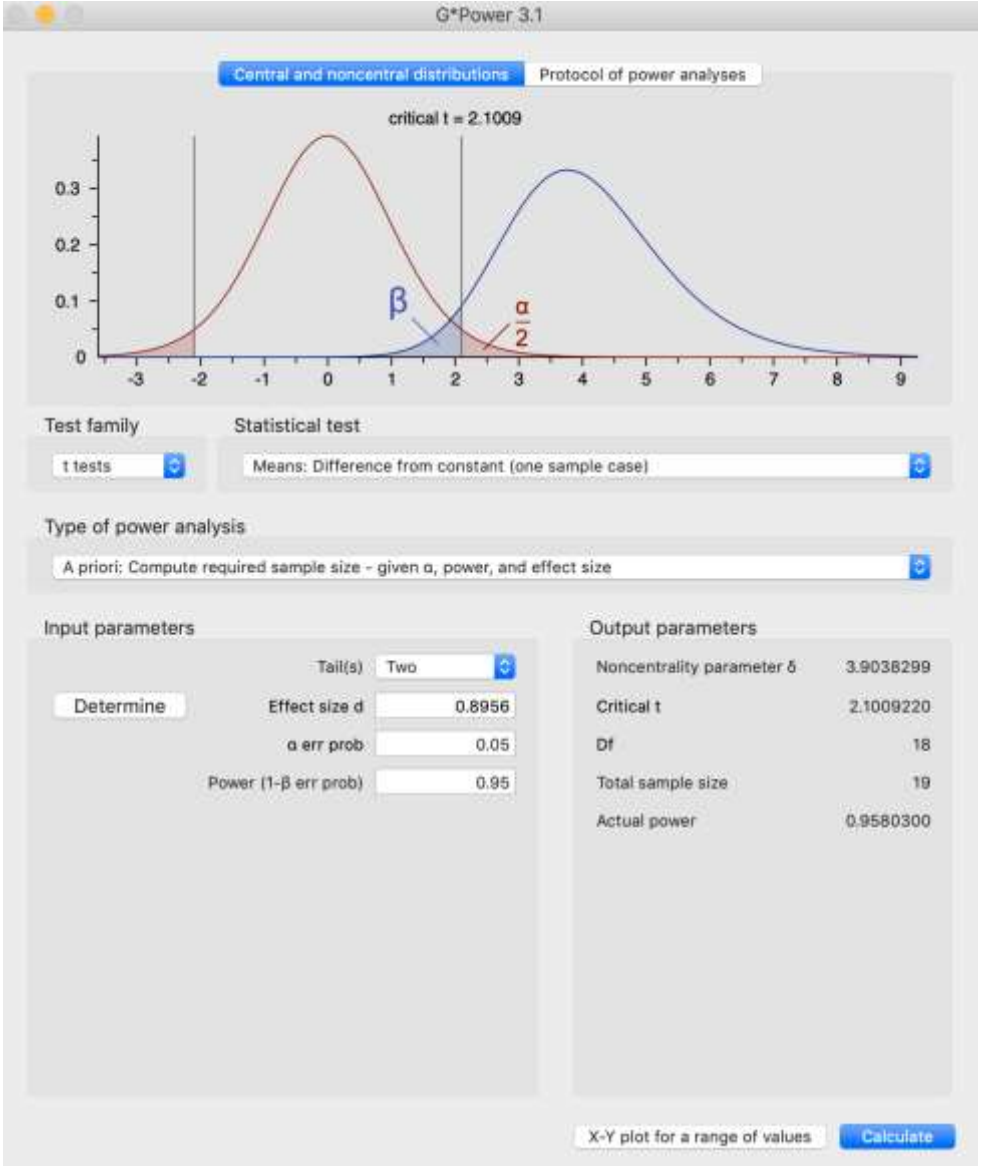
F tests - Linear multiple regression: Fixed model, R^2 increase

Analysis:	A priori: Compute required sample size			
Input:	Effect size f^2	=	15.92047	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.95	

	Number of tested predictors		=	2
	Total number of predictors		=	3
Output:	Noncentrality parameter λ		=	111.4433
	Critical F	=	9.5520945	
	Numerator df		=	2
	Denominator df		=	3
	Total sample size		=	7
	Actual power		=	0.9986048

Power analysis for one sample experiments

For comparative ability judgments across the 4 easy ability domains

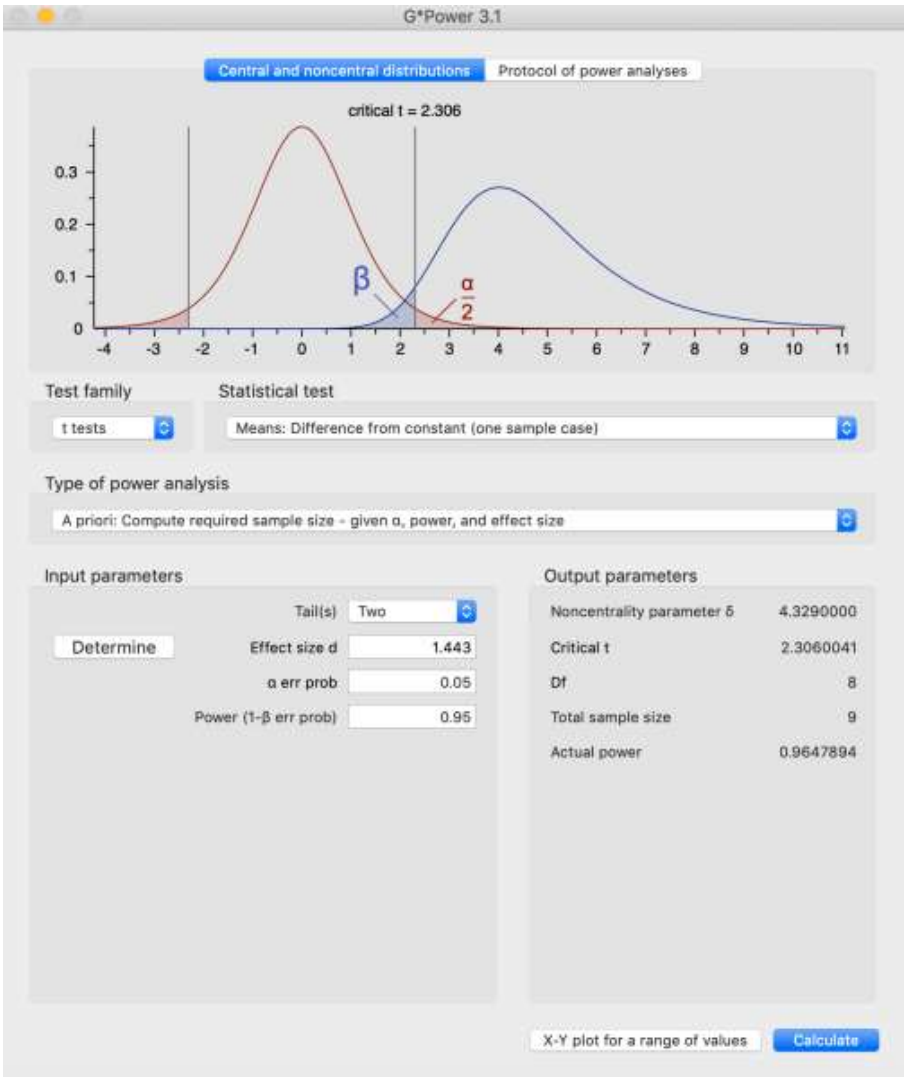


t tests - Means: Difference from constant (one sample case)

Analysis: A priori: Compute required sample size

Input:	Tail(s)	=	Two	
	Effect size d	=	0.8956	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.95	
Output:	Noncentrality parameter δ	=	3.9038299	
	Critical t	=	2.1009220	
	Df	=	18	
	Total sample size	=	19	

For comparative ability judgments across the 4 difficult ability domains

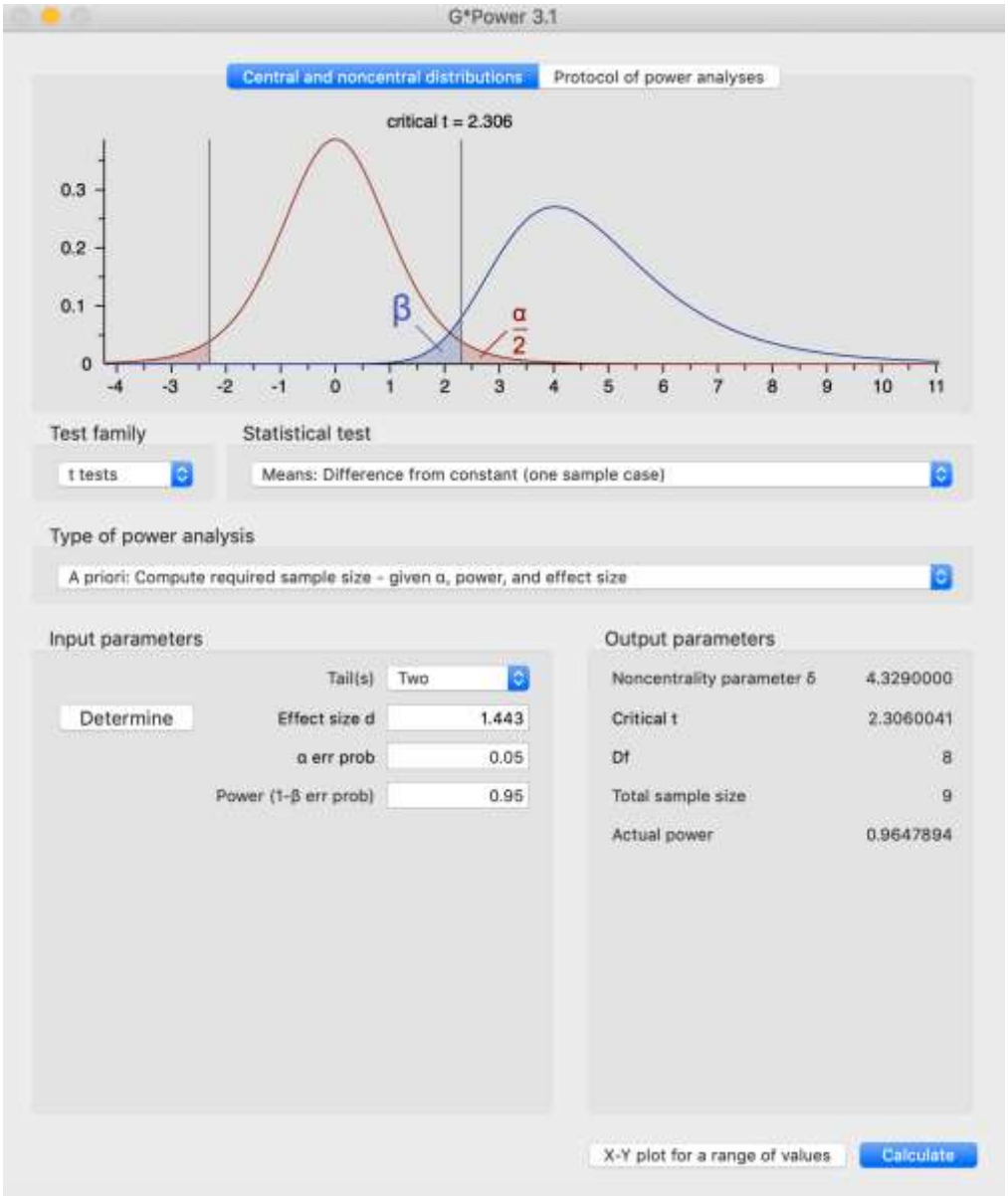


t tests - Means: Difference from constant (one sample case)

Analysis:	A priori: Compute required sample size		
Input:	Tail(s)	=	Two

	Effect size d	=	1.4430
	α err prob	=	0.05
	Power (1-β err prob)	=	0.95
Output:	Noncentrality parameter δ	=	4.3290000
	Critical t	=	2.3060041
	Df	=	8
	Total sample size	=	9
	Actual power	=	0.9647894

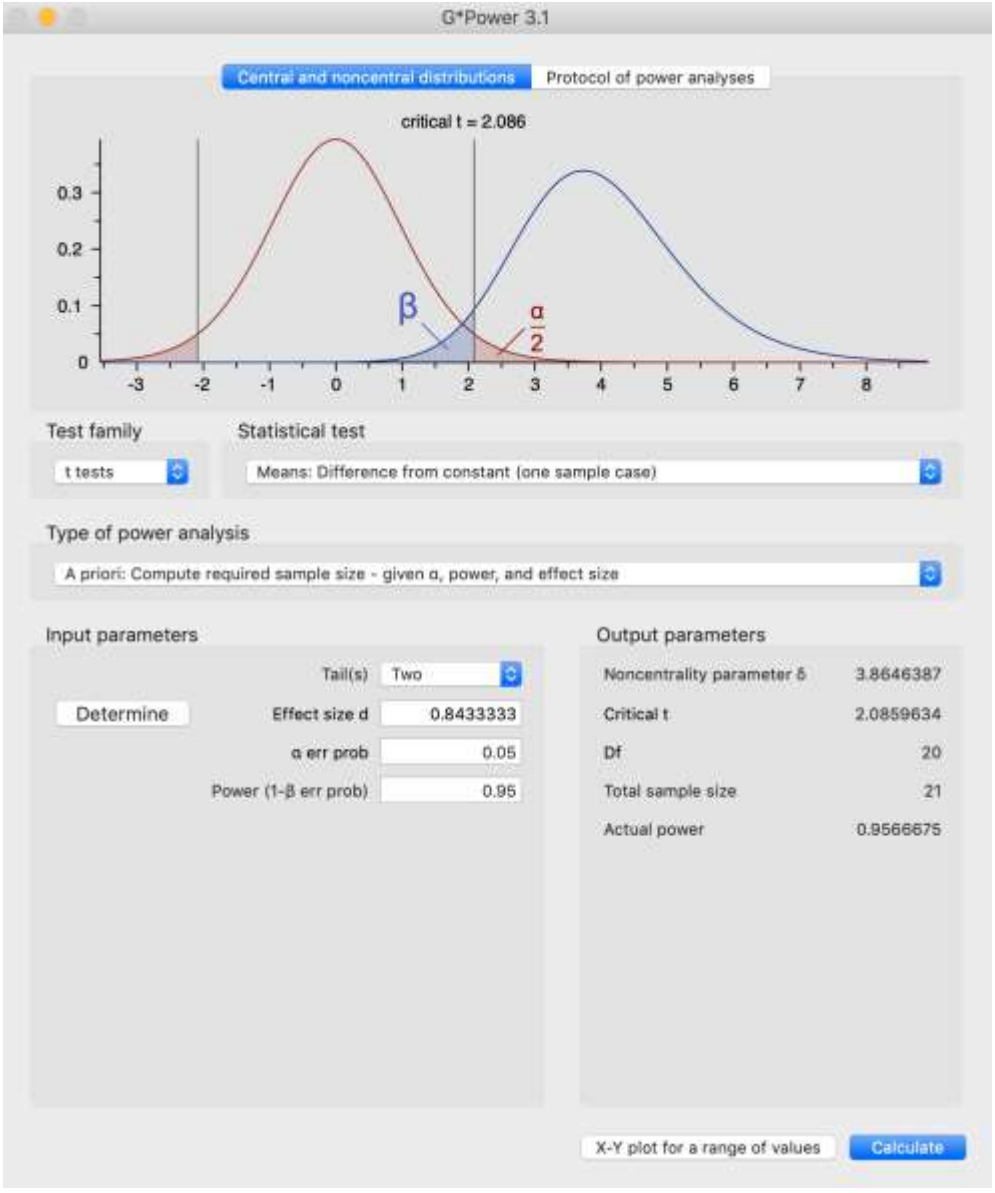
For mean desirability rating across the 4 easy ability domains



t tests - Means: Difference from constant (one sample case)

Analysis:	A priori: Compute required sample size			
Input:	Tail(s)	=	Two	
	Effect size d	=	2.251667	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.95	
Output:	Noncentrality parameter δ	=	5.0348805	
	Critical t	=	2.7764451	
	Df	=	4	
	Total sample size	=	5	
	Actual power	=	0.9570378	

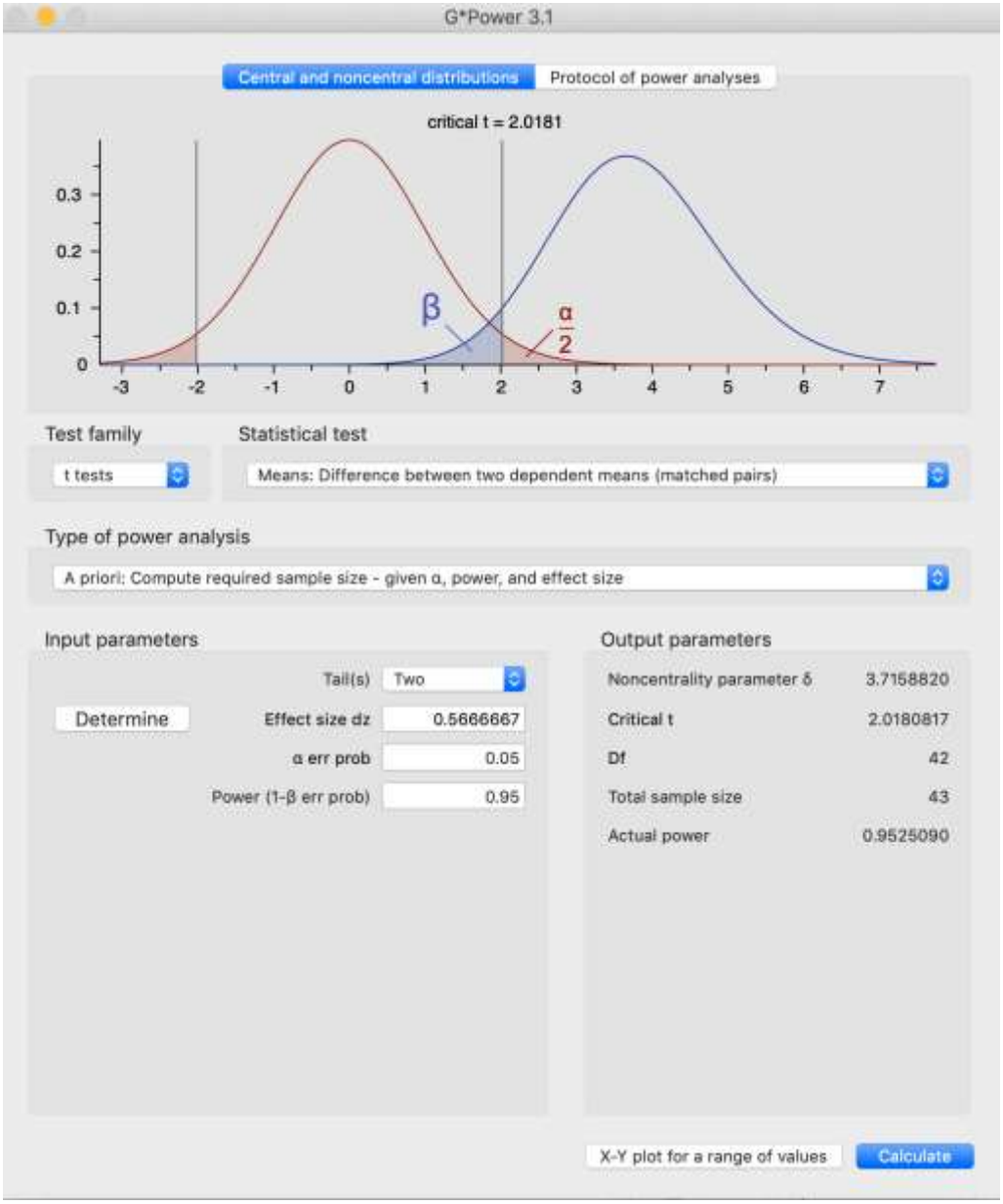
For mean desirability rating across the 4 difficult ability domains



Analysis:	A priori: Compute required sample size			
Input:	Tail(s)	=	Two	
	Effect size d	=	0.8433333	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.95	
Output:	Noncentrality parameter δ	=	3.8646387	
	Critical t	=	2.0859634	
	Df	=	20	
	Total sample size	=	21	
	Actual power	=	0.9566675	

Power analysis for paired sample t-test

For ambiguity ratings between easy and difficult abilities

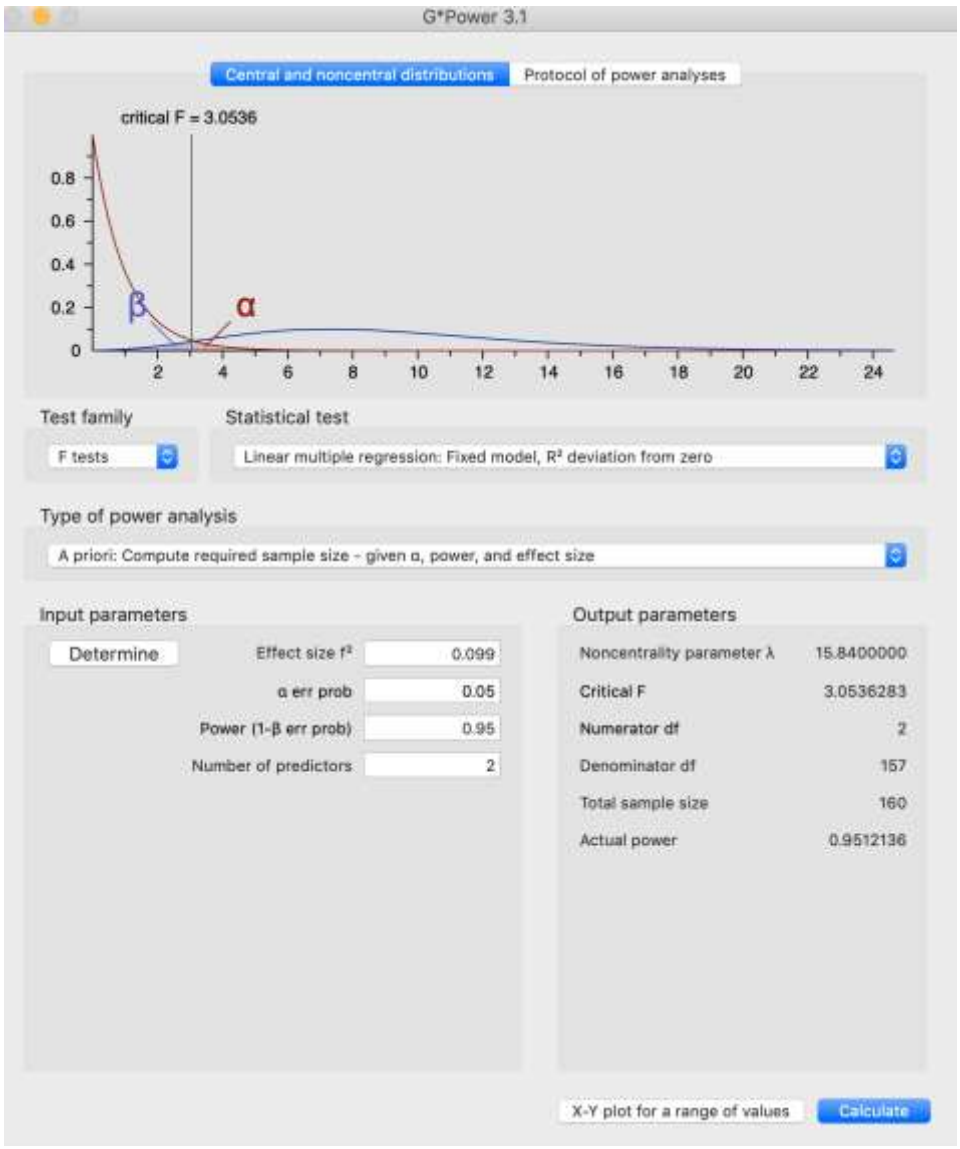


t tests - Means: Difference between two dependent means (matched pairs)

Analysis:	A priori: Compute required sample size			
Input:	Tail(s)	=	Two	
	Effect size dz	=	0.5666667	
	α err prob	=	0.05	
	Power (1-β err prob)		=	0.95
Output:	Noncentrality parameter δ		=	3.7158820
	Critical t	=	2.0180817	
	Df	=	42	
	Total sample size		=	43
	Actual power		=	0.9525090

Power analysis for multiple regression

For comparative ability estimates from participant estimates of their own and their peers' absolute abilities



F tests - Linear multiple regression: Fixed model, R^2 deviation from zero

Analysis:	A priori: Compute required sample size			
Input:	Effect size f^2	=	0.099	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.95	
	Number of predictors	=	2	
Output:	Noncentrality parameter λ	=	15.8400000	
	Critical F	=	3.0536283	
	Numerator df	=	2	
	Denominator df	=	157	
	Total sample size	=	160	
	Actual power	=	0.9512136	

Materials and scales used in the replication + extension experiment

Extension introduction and explanation

For extension one, two additional conditions were added: the easy domain and the difficult domain conditions. For extension two, an additional dependent variable of domain difficulty ratings is added. Domain difficulty would be scored on a scale from 1 (very easy) to 10 (very difficult).

Table of design

The table of design is found under Table 7 of the methods section in the main manuscript.

Instructions and experimental material

Consent

All participants first had to give consent before being randomly assigned to one of the three conditions.

Consent Form

This study is conducted by Gilad Feldman of the psychology department at University of Hong Kong and colleagues. If you have questions or concerns regarding this project, please do not hesitate to contact Gilad Feldman gfeldman@hku.hk at any time.

Purpose of the study
To understand how people think, feel, make decisions, and act in various types of situations. Preferences and individual differences between people, as well as both internal and external factors, may affect these types of responses and this research intends to uncover and/or understand these processes.

Procedures.
This study will ask you to complete a set of questionnaires requiring decision making in various scenarios. The duration of this study has been indicated on the Amazon Mechanical Turk HIT that you accepted.

Potential risks.
This procedure has no known risks greater than those of ordinary daily life.

Potential benefits.
This study aims to add to existing research line in the field of social-cognitive-personality psychology. We also hope that this study can provide you with a learning experience of participating in psychological research and possibly learning more about yourself and your beliefs, evaluations, preferences, personality, etc..

Compensation.
Compensation is offered through the Amazon Mechanical Turk platform. The level of compensation has been indicated on the Amazon Mechanical Turk HIT that you accepted.

Confidentiality.
Your questionnaire responses are anonymous and strictly confidential. No personal identifiers are kept. Information obtained will only be used as aggregates for research purposes.

Participation and withdrawal.
Your participation is voluntary. This means that you can choose to stop at any time without negative consequences. If at any time you wish to withdraw, please simply indicate eight zeros as your completion code on MTurk, and you will receive compensation regardless.

Questions and concerns
If you have any questions about the research, please feel free to contact Gilad Feldman at the University of Hong Kong (gfeldman@hku.hk). If you have questions about your rights as a research participant, contact the Human Research Ethics Committee, HKU (+852 2241-5267).

Please print a copy of this consent form for your records, if you so desire.

Consent agreement
Please select the message box below to indicate that you are 18 years old or older and have read and agree to the above.

I fully understand the contents of this consent form and agree to participate in this study. I also agree not to disclose the details of the study to other parties.
(Participants would provide a tick to this statement to give consent)

Instruction
In this study, you will be presented with a series of questions on different abilities. Please read and answer all questions carefully. When you are ready to begin, please select >> (Next) to proceed to the study.

Ability domains

Participants were then randomly assigned into one of the three conditions: original domain group, easy domain group, or difficult domain group. Depending on the condition, participants were shown one of the three versions of eight ability domain descriptions.

Ratings
For each of the eight ability domains, participants provided seven ratings.
1.Domain difficulty
Question: Please rate the difficulty of this ability from 1 (very easy) to 10 (very difficult):
- Scale: 1 (very easy) to 10 (very difficult)

2.Comparative ability
Question: Please compare yourself with other MTurk workers of the same age, gender, and socioeconomic background as you on this ability from 0 to 100. Please click on the bar to indicate your rank compared to others on this ability.
- Scale: 0 (I'm at the very bottom) to 50 (I'm exactly average) to 100 (I'm at the very top).

3.Own absolute ability
Question: For this ability, please rate your own ability from 1 (very unskilled) to 10 (very skilled):
- Scale: 1 (very unskilled) to 10 (very skilled)

4.Others' absolute ability
Question: For this ability, please rate the ability of other MTurk workers of the same age, gender, and socioeconomic background as you from 1 (very unskilled) to 10 (very skilled):

- Scale: 1 (very unskilled) to 10 (very skilled)

5.Desirability

Question: Please rate ability desirability: is it better to be very unskilled or very skilled at this ability from 1 (better to be very unskilled) to 10 (better to be very skilled).

- Scale: 1 (better to be very unskilled) to 10 (better to be very skilled)

6.Ambiguity

Question: Please rate the ambiguity of the phrase used for this ability "(ability domain)" from 1 (very ambiguous, has many meanings) to 10 (very clear, has one meaning).

- Scale: 1 (very ambiguous, has many meanings) to 10 (very clear, has one meaning)

7.Experience in the ability domain

Question: Please rate your experience in using this ability from 1 (no experience at all) to 10 (very experienced).

- Scale: 1 (no experience at all) to 10 (very experienced)

Funnelling section

Five funnelling questions were presented:

1. How serious were you in filling out this questionnaire?
- 1 (not at all) to 5 (very much)
2. Have you ever seen the materials used in this study or similar before? If yes - please indicate where.
3. What do you think the purpose of the last part was? (one sentence)
4. Help us improve for the next studies - Did you spot any errors? Anything missing or wrong? Something we should pay attention to in next runs? (briefly)
5. Please rate your satisfaction with the pay/compensation offered for this MTurk HIT (note - this will not impact your pay in any way)
- 0 (extremely unsatisfied) to 3 (neutral) to 6 (very satisfied)

Demographics

Four questions on demographics were presented:

1. Please indicate your gender
    - Male
    - Female
    - Other/Would rather not disclose
2. Which country are you originally from? (Country of birth)
    - (type country of birth)
3. Please estimate your family's social class
    - Lower class
    - Working class
    - Lower Middle class
    - Middle class
    - Upper Middle class
    - Upper class
4. How would you generally rate your understanding of English used in this study?
    - Very bad
    - Bad
    - Poor
    - Neither good or bad
    - Fair

- Good
- Very good

Exclusion criteria

The current replication's exclusion criteria are summarised in table 4.

Table 4.

Summary of exclusion criteria

Exclusion Criteria	Reason
Exclusion Criteria 1a Participants indicating low proficiency of English. Item: "On a scale from 1-7, what do you think is your proficiency of English?" (1 = being not proficient at all, 7 = being very proficient.) *Exclusion: if self-report less than 5 on a 1-7 scale*	Without a fair proficiency of English participants may not fully understand the questions and may affect results.
Exclusion Criteria 1b Participants who self-report not being serious about filling in the survey. Item: "On a scale from 1-5, what do you think is your seriousness in filling in the survey?" (1 = not serious at all, 5 = very serious.) *Exclusion: if self-report less than 4 on a 1-5 scale*	Nonserious answering behavior increases noise and reduces experimental power. Excluding their response can increase data validity.
Exclusion Criteria 1c Participants who correctly guessed any one of the hypotheses of this study in the funnelling section. Item: "What do you think the purpose of the last part was?" (If you are not sure please write "not sure") *Exclusion: if guessed correctly for both replication and extension*	Participants who could guess any of the hypotheses of the study may commit experimental bias and do not reflect the true nature of the investigated phenomenon

Exclusion Criteria 1d Participants who have already seen or done the survey before. Item: "Have you ever seen the materials used in this study or similar before? If yes - please indicate where." *Exclusion: answered 'yes'*	Experimental bias in responses
Exclusion Criteria 1e Participants who failed to complete the survey. (duration = 0, leave question blank)	Incomplete data
Exclusion Criteria 1f Participants not from the United States. Item: "Which country are you originally from? (Country of birth)" *Exclusion: if response is not part of the United States*	Does not fit our targeted population criteria

Comparisons and deviations

Similarities and differences between the original study and replication study

	Original	Replication	Reason for change
Study design	Within-subjects design	Mixed design	From one condition to three conditions (add two extension conditions to test hypotheses 3 &4)
Procedure	Participants were all assigned to the same condition and were presented with the same ability domains.	Participants were randomly assigned to one of the three conditions, and in each condition participants were presented with different versions (original, easy interpretations, difficult interpretations) of the 8 ability domains.	Extension: changed independent variable of ability domain definitions
Conditions	One condition: original ability domains.	Three conditions: 1. original domain group, 2. easy domain group, and 3. difficult domain group.	Extension: changed independent variable of ability domain definitions
Domain difficulty ratings	Ratings obtained in pretest by a separate group of participants (N = 39).	Ratings obtained for each of the eight abilities across all three conditions.	Extension: additional dependent variable of domain difficulty ratings
Measures / stimulus	Written questionnaire	Online Qualtrics survey	
	4 easy abilities, 4 difficult abilities	8 abilities, abilities not categorised as easy or difficult.	Issues with categorising continuous variables (see adjustments to original study)
	Comparison group: other students from their psychology course	Comparison group: others of the same age, gender, and socioeconomic background as you	Used Mturk instead of college setting to recruit participants
	DV2: Comparative ability rating (scale from 0 to 99)	DV2: Comparative ability rating (scale from 0 to 100)	Use 100 instead of 99 for easier understanding and comparison

Location	Introductory psychology students at Cornell University, New York	Online	Recruit participants online on MTurk
Remuneration	Extra course credit	Money	Used Mturk instead of college setting to recruit participants
Calculations	One sample t -tests, correlations, multiple regressions	Correlations, Hotelling-Williams-test, multiple regressions, ANOVA	Issues with categorising continuous variables (section: “adjustments to original study” in main manuscript)

Pre-exclusions versus post-exclusions

Overview of pre-exclusions and post-exclusions

	Before exclusion	Fulfilled any of the exclusion criteria	After exclusion
Total number of cases	756	65	691

Summary of exclusions

Exclusion Criteria	Cases fulfilling the exclusion criteria	Cases remaining after exclusion
Exclusion Criteria 1a Participants indicating low proficiency of English.	10	746
Exclusion Criteria 1b Participants who self-report not being serious about filling in the survey.	19	727
Exclusion Criteria 1c Participants who correctly guessed any one of the hypotheses of this study in the funnelling section.	0	727
Exclusion Criteria 1d Participants who have already seen or done the survey before.	7	720
Exclusion Criteria 1e Participants who failed to complete the survey.	0	720
Exclusion Criteria 1f Participants not from the United States.	29	691

Pre-registration plan versus final report

See Preregistration Planning and Deviation Documentation (PPDD) document for latest updates.

Components in your preregistration	Location of 1) preregistered decision/plan and 2) rational for decision/plan	Were there deviations? What type? minor	If yes - describe details of deviation(s)	Rationale for deviation	How might the results be different if you had/had not deviated	Date/time of decision for deviation + stage
Study design		No	/	/	/	/
Measured variables		No	/	/	/	/
Exclusion criteria		No	/	/	/	/
IV		No	/	/	/	/
DV		No	/	/	/	/
Data analysis		Minor	1. Added Welch's T-tests under the section "comparisons between the three groups" in the main manuscript. 2. Added one-sample T-tests and correlations for ratings across all abilities under "Replication: original domain condition" in the main manuscript. 3. Condensed code files into 1 R file. Redundant calculations on both JAMOVI and R during pre-registration.	1.Shapiro-wilk and Levene's tests show that assumptions of normality and homogeneity of variance are not met 2.Pre-registered that if the paired samples t-test for the replication condition is significant, original study analysis will be carried out. 3. Easier viewing, avoid redundancy	1. No difference, added tests to supplement data analysis 2. No difference 3. No difference	1, 2, & 3: 03/07/2020, after data collection

Table notes: Locations should include page number (section) and paragraph or line number (as specific as possible). Where possible, please embed in-document hyperlinks to make browsing easier. *Categories for deviations: Minor - Change probably did not affect results or interpretations; Major - Change likely affected results or interpretations.

Variable computation

Non-excluded variables
Columns from file: Kruger (1999) non-excluded dataset
Variable: easy abilities difficulty (non-excluded, replication)
mouse_difficulty + drive_difficulty + ride_difficulty + saving_difficulty
(4 columns combined into one total column)
Variable: difficult abilities difficulty (non-excluded, replication)
joke_difficulty + chess_difficulty + juggle_difficulty + program_difficulty
(4 columns combined into one total column)
Variable: difficulty across all abilities (replication, non-excluded)
mouse_difficulty + drive_difficulty + ride_difficulty + saving_difficulty + joke_difficulty + chess_difficulty + juggle_difficulty + program_difficulty
(8 columns combined into one total column)
Variable: compare across all abilities (replication, non-excluded)
mouse_compare_8 + drive_compare_8 + ride_compare_8 + saving_compare_8 + joke_compare_8 + chess_compare_8 + juggle_comapre_8 + program_compare_8
(8 columns combined into one total column)
Variable: Own ability across all abilities(replication, non-excluded)
mouse_own + drive_own + ride_own + saving_own + joke_own + chess_own + juggle_own + program_own
(8 columns combined into one total column)

Excluded variables
Columns from file: Kruger (1999) excluded dataset
Variable: easy abilities difficulty (replication group)
mouse_difficulty + drive_difficulty + ride_difficulty + saving_difficulty
(4 columns combined into one total column)
Variable: difficult abilities difficulty (replication group)
joke_difficulty + chess_difficulty + juggle_difficulty + program_difficulty
(4 columns combined into one total column)
Variable: easy abilities difficulty (easy group)
mouse_difficulty_e + drive_difficulty_e + ride_difficulty_e + saving_difficulty_e
(4 columns combined into one total column)
Variable: difficult abilities difficulty (easy group)
joke_difficulty_e + chess_difficulty_e + juggle_difficulty_e + program_difficulty_e
(4 columns combined into one total column)
Variable: easy abilities difficulty (difficult group)
mouse_difficulty_d + drive_difficulty_d + ride_difficulty_d + saving_difficulty_d
(4 columns combined into one total column)
Variable: difficult abilities difficulty (difficult group)
joke_difficulty_d + chess_difficulty_d + juggle_difficulty_d + program_difficulty_d
(4 columns combined into one total column)
Variable:difficulty across all abilities (replication group)

mouse_difficulty + drive_difficulty + ride_difficulty + saving_difficulty + joke_difficulty + chess_difficulty + juggle_difficulty + program_difficulty
(8 columns combined into one total column)
Variable: compare across all abilities (replication group)
mouse_compare_8 + drive_compare_8 + ride_compare_8 + saving_compare_8 + joke_compare_8 + chess_compare_8 + juggle_comapre_8 + program_compare_8
(8 columns combined into one total column)
Variable: own absolute ability across all abilities (replication group)
mouse_own + drive_own + ride_own + saving_own + joke_own + chess_own + juggle_own + program_own
(8 columns combined into one total column)

Pre-exclusion versus post-exclusion results

Summary of pre-exclusion versus post-exclusion main results for the replication condition

Test	**Pre-exclusion (*N* = 756)**	**Post-exclusion (*N* = 691)**
Paired samples t-test Domain difficulty ratings between easy and difficult abilities	$t(1031) = -29.3$, $p < .001$	$t(959) = -29.1$, $p < .001$
Correlational studies Domain difficulty and comparative ability across all abilities	$r = -0.31$, $p < .001$, 95% CI [-0.35, -0.27]	$r = -0.35$, $p < .001$, 95% CI [-0.39, -0.31]
Own absolute ability and comparative ability across all abilities	$r = 0.81$, $p < .001$, 95 % CI [0.80, 0.83]	$r = 0.81$, $p < .001$, 95 % CI [0.79, 0.83]
Mean domain difficulty and mean comparative ability across all abilities	$r = 0.27$, $p < .001$, 95 % CI [0.15, 0.38]	$r = 0.16$, $p = 0.012$, 95% CI [0.04, 0.28]
Mean own absolute ability and mean comparative ability across all abilities	$r = 0.87$, $p < .001$, 95 % CI [0.83, 0.89]	$r = 0.85$, $p < .001$, 95% CI [0.82, 0.89]
Multiple regression Mean own and others' absolute ability ratings predicting mean comparative ability judgments	$F(2, 255) = 384.4$, $p < .001$, with an R^2 of 0.75.	$F(2, 237) = 323.9$, $p < .001$, with an R^2 of 0.73.

Note. No significant differences in main results were found for pre-exclusion versus post-exclusion analysis for the replication condition.

Statistical assumptions and normality Tests

File: Kruger (1999) excluded final.R

Effect size and confidence intervals for independent samples t-tests between the replication and easy or difficult domain conditions:

Domain difficulty ratings between the replication condition and easy domain condition

```
> #cohen's d domaindifficulty1and2
> ES.t.two(m1=6.05,m2=5.22,sd1=1.15,sd2=1.63,n1=240,n2=225)

    effect size (Cohen's d) of independent two-sample t test

           d = 0.5916382
   alternative = two.sided

NOTE: The alternative hypothesis is m1 != m2
small effect size:  d = 0.2
medium effect size: d = 0.5
large effect size:  d = 0.8

> d.ci(0.59,n=465,n1=240,n2=225,alpha=.05)
        lower effect     upper
[1,] 0.4038918   0.59 0.7754965
```

Domain difficulty ratings between the replication condition and difficult domain condition

```
> #cohen's d domaindifficulty1and3
> ES.t.two(m1=6.05,m2=7.39,sd1=1.15,sd2=1.19,n1=240,n2=226)

    effect size (Cohen's d) of independent two-sample t test

           d = 1.145723
   alternative = two.sided

NOTE: The alternative hypothesis is m1 != m2
small effect size:  d = 0.2
medium effect size: d = 0.5
large effect size:  d = 0.8

> d.ci(1.15,n=466,n1=240,n2=226,alpha=.05)
       lower effect    upper
[1,] 0.953311    1.15 1.345615
```

Ambiguity ratings between the replication condition and easy domain condition

```
> #cohen's d ambiguity1and2
> ES.t.two(m1=8.00,m2=8.32,sd1=1.24,sd2=1.23,n1=240,n2=225)

    effect size (Cohen's d) of independent two-sample t test

           d = 0.2590732
   alternative = two.sided

NOTE: The alternative hypothesis is m1 != m2
small effect size:  d = 0.2
medium effect size: d = 0.5
large effect size:  d = 0.8

> d.ci(0.26,n=465,n1=240,n2=225,alpha=.05)
        lower effect    upper
[1,] 0.07721565   0.26 0.442508
```

Ambiguity ratings between the replication condition and difficult domain condition

```
> #cohen's d ambiguity1and3
> ES.t.two(m1=8.00,m2=8.24,sd1=1.24,sd2=1.43,n1=240,n2=226)

    effect size (Cohen's d) of independent two-sample t test

           d = 0.1797061
   alternative = two.sided

NOTE: The alternative hypothesis is m1 != m2
small effect size:  d = 0.2
medium effect size: d = 0.5
large effect size:  d = 0.8

> d.ci(0.18,n=466,n1=240,n2=226,alpha=.05)
         lower effect     upper
[1,] -0.00213244   0.18 0.3619412
```

Confidence intervals for paired sample t-test results comparing the domain difficulty ratings between easy and difficult abilities:

```
> #confidence intervals for paired sample t-test comparing the domain difficulty ratings
between easy and difficult abilities
> #replication condition
> cohen.d.ci(d=-0.939, n1 = 960, alpha = .05)
```

```
       lower effect      upper
[1,] -1.014767 -0.939 -0.8628849
> #easy domain condition
> cohen.d.ci(d=-0.624, n1 = 900, alpha = .05)
       lower effect      upper
[1,] -0.6952683 -0.624 -0.5524394
> #difficult domain condition
> cohen.d.ci(d=-0.429, n1 = 904, alpha = .05)
       lower effect      upper
[1,] -0.4970143 -0.429 -0.3607673
```

R^2 confidence intervals for multiple regressions on mean own and others' absolute abilities predicting mean comparative ability:

```
> #Multiple regression CI for the replication condition
> CI.Rsq(0.7322,239, 2, level = 0.95)
    Rsq      SErsq       LCL       UCL
1 0.7322 0.02909147 0.6751818 0.7892182

> #Multiple regression CI for the easy domain condition
> CI.Rsq(0.6896, 224, 2, level = 0.95)
    Rsq      SErsq       LCL       UCL
1 0.6896 0.03375872 0.6234341 0.7557659

> #Multiple regression CI for the difficult domain condition
> CI.Rsq(0.7546, 225, 2, level = 0.95)
    Rsq     SErsq       LCL       UCL
1 0.7546 0.0278593 0.6999968 0.8092032
```

Independent sample t-tests for comparisons between the three conditions:
Shapiro-wilk test for normality

R codes for Shapiro-wilk tests for mean domain difficulty and ambiguity ratings for all three conditions. The hypotheses of normality is rejected for mean domain difficulty ratings for the conditions, $p < .001$. Welch's t-tests are therefore also included in the main manuscript to supplement the Student's t-tests.

Domain difficulty (replication condition)

```
> shapiro.test(kruger$mean_domaindifficulty_condition1)

        Shapiro-Wilk normality test

data:  kruger$mean_domaindifficulty_condition1
W = 0.96995, p-value = 0.00005811
```

Domain difficulty (easy domain condition)

```
> shapiro.test(kruger$mean_domaindifficulty_condition2)

        Shapiro-Wilk normality test

data:  kruger$mean_domaindifficulty_condition2
W = 0.93229, p-value = 0.0000000113
```

Domain difficulty (difficult domain condition)

```
> shapiro.test(kruger$mean_domaindifficulty_condition3)

        Shapiro-Wilk normality test

data:  kruger$mean_domaindifficulty_condition3
W = 0.94822, p-value = 0.0000003177
```

Ambiguity (replication condition)

```
> shapiro.test(kruger$mean_ambiguity_condition1)

        Shapiro-Wilk normality test

data:  kruger$mean_ambiguity_condition1
W = 0.96538, p-value = 0.00001442
```

Ambiguity (easy domain condition)

```
> shapiro.test(kruger$mean_ambiguity_condition2)

        Shapiro-Wilk normality test

data:  kruger$mean_ambiguity_condition2
W = 0.94575, p-value = 0.0000001912
```

Ambiguity (difficult domain condition)

```
> shapiro.test(kruger$mean_ambiguity_condition3)

        Shapiro-Wilk normality test

data:  kruger$mean_ambiguity_condition3
W = 0.89103, p-value = 0.00000000001025
```

F-test for equality in variances

R codes for F-tests for mean domain difficulty and ambiguity ratings between the replication and easy domain groups, and the replication and difficult domain groups. Apart from mean ambiguity ratings in the replication and difficult domain condition (p =.029),the p values for all other conditions were all larger than 0.05, meaning that variances were equal across groups.

Domain difficulty ratings between the replication condition and easy domain condition

```
> var.test(kruger$mean_domaindifficulty_condition1,
kruger$mean_domaindifficulty_condition2,
+         alternative = c("two.sided"),
+         conf.level = 0.95)

        F test to compare two variances

data:  kruger$mean_domaindifficulty_condition1 and
kruger$mean_domaindifficulty_condition2
F = 0.49985, num df = 239, denom df = 224, p-value = 0.0000001639
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3858126 0.6468819
sample estimates:
ratio of variances
         0.4998488
```

Domain difficulty ratings between the replication condition and difficult domain condition

```
> var.test(kruger$mean_domaindifficulty_condition1,
```

```
kruger$mean_domaindifficulty_condition3,
+         alternative = c("two.sided"),
+         conf.level = 0.95)

          F test to compare two variances

data:  kruger$mean_domaindifficulty_condition1 and
kruger$mean_domaindifficulty_condition3
F = 0.93829, num df = 239, denom df = 225, p-value = 0.6271
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7244734 1.2139816
sample estimates:
ratio of variances
        0.9382927
```

Ambiguity ratings between the replication condition and easy domain condition

```
> var.test(kruger$mean_ambiguity_condition1, kruger$mean_ambiguity_condition2,
+         alternative = c("two.sided"),
+         conf.level = 0.95)

          F test to compare two variances

data:  kruger$mean_ambiguity_condition1 and kruger$mean_ambiguity_condition2
F = 1.0099, num df = 239, denom df = 224, p-value = 0.9415
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7795024 1.3069713
sample estimates:
ratio of variances
         1.009903
```

Ambiguity ratings between the replication condition and difficult domain condition

```
> var.test(kruger$mean_ambiguity_condition1, kruger$mean_ambiguity_condition3,
+         alternative = c("two.sided"),
+         conf.level = 0.95)

        F test to compare two variances

data:  kruger$mean_ambiguity_condition1 and kruger$mean_ambiguity_condition3
F = 0.74993, num df = 239, denom df = 225, p-value = 0.0286
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5790329 0.9702707
sample estimates:
ratio of variances
       0.7499273
```

Additional Tables and Figures

Replication condition

Non-excluded

Table 5.1.

Partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	0.27	0.26	0.093	0.01 [-0.00, 0.04]
Comparative ability, domain difficulty	Ambiguity	0.27	0.27	0.346	0.00 [-0.00, 0.02]

Note. The replication group did not find a significant difference between the zero order correlation ***(p.xy)*** and partial correlation ***(p.xy.z)*** between comparative ability and domain difficulty ratings while holding desirability or ambiguity constant. Both tests were not able to reject the null hypothesis: ρ.xy - ρ.xy.z = 0.

Excluded

Table 5.2.

Partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	0.16	0.16	0.212	0.01 [-0.00, 0.03]
Comparative ability, domain difficulty	Ambiguity	0.16	0.16	0.659	0.001 [-0.01, 0.02]

Note. The replication group did not find a significant difference between the zero order correlation ***(p.xy)*** and partial correlation ***(p.xy.z)*** between comparative ability and domain difficulty ratings while holding desirability or ambiguity constant. Both tests were not able to reject the null hypothesis: ρ.xy - ρ.xy.z = 0.

Non-excluded

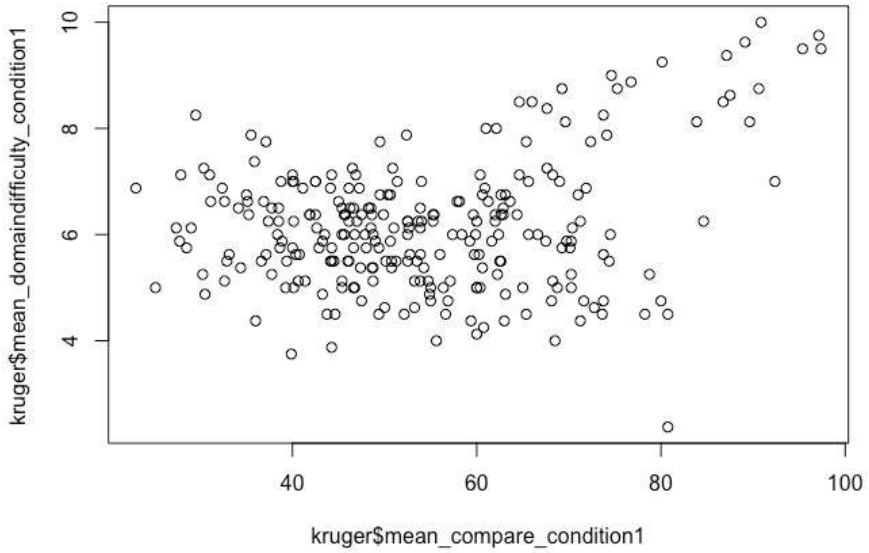


Figure 2.1 Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the replication group.

Excluded

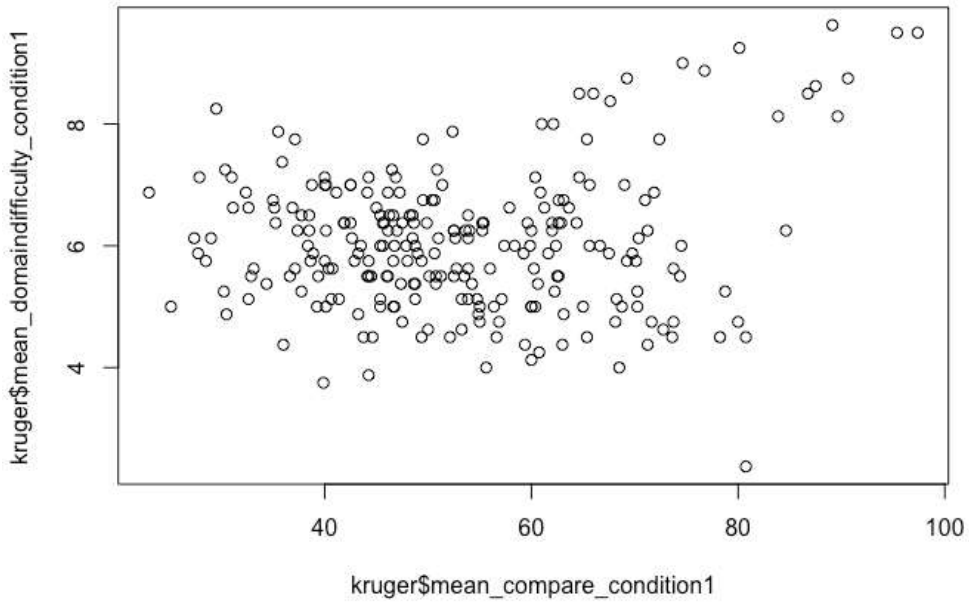


Figure 2.2 Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the replication group.

Non-excluded

Table 5.3

Replication condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	-1.80	[-6.80, 3.20]						
kruger$mean_own_condition1	9.46**	[8.53, 10.38]	0.88	[0.80, 0.97]	.39	[.31, .48]	.87**	
kruger$mean_other_condition1	-0.27	[-1.34, 0.80]	-0.02	[-0.11, 0.06]	.00	[-.00, .01]	.60**	
								R^2 = .751**
								95% CI[0.70, 0.80]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Excluded

Table 5.4

Replication condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	1.27	[-4.13, 6.67]						
kruger$mean_own_condition1	9.51**	[8.57, 10.44]	0.90	[0.81, 0.99]	.45	[.36, .54]	.85**	
kruger$mean_other_condition1	-0.87	[-2.00, 0.26]	-0.07	[-0.16, 0.02]	.00	[-.00, .01]	.53**	
								R^2 = .732**
								95% CI[0.68, 0.79]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Non-excluded

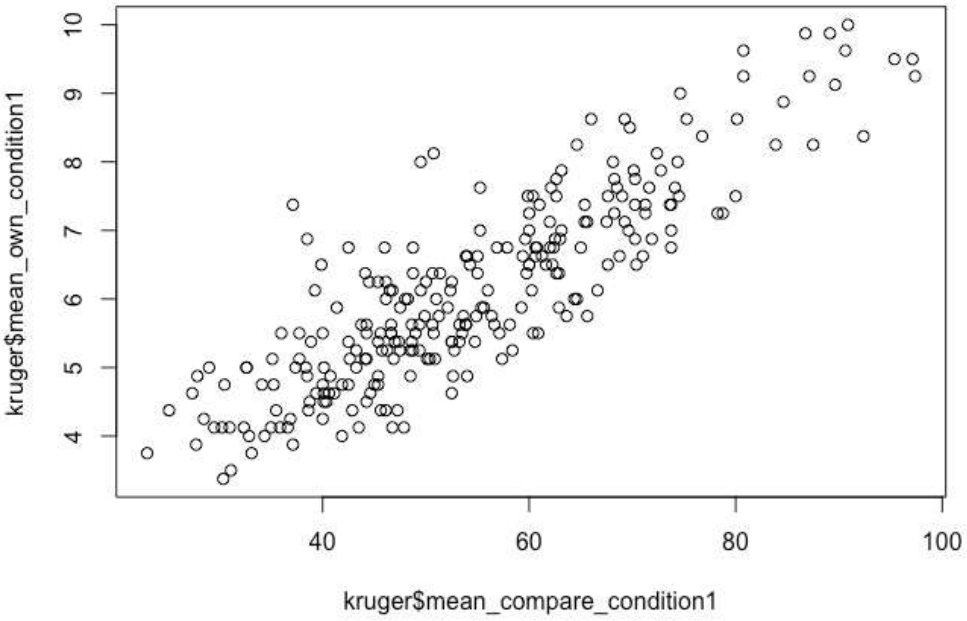


Figure 2.3. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the replication group.

Excluded

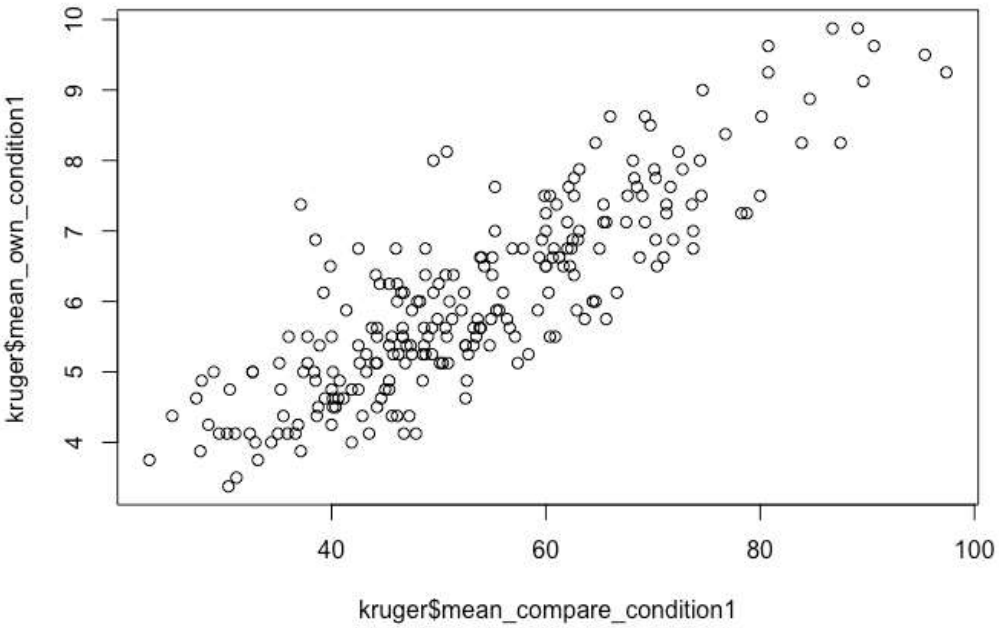


Figure 2.4. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the replication group.

Non-excluded

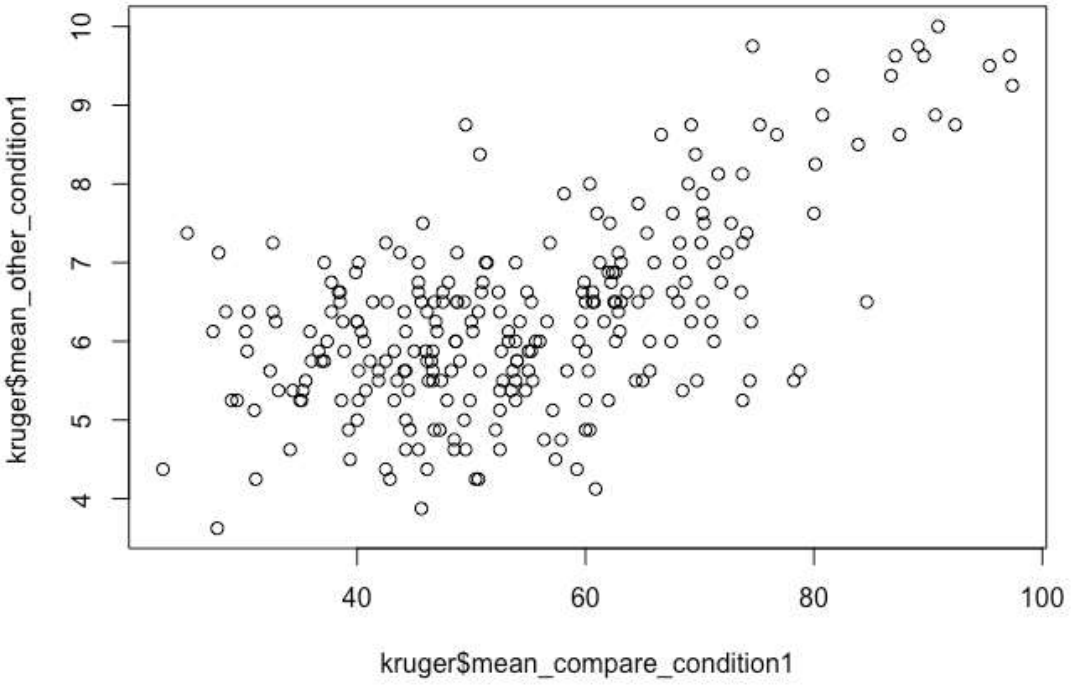


Figure 2.5. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the replication group.

Excluded

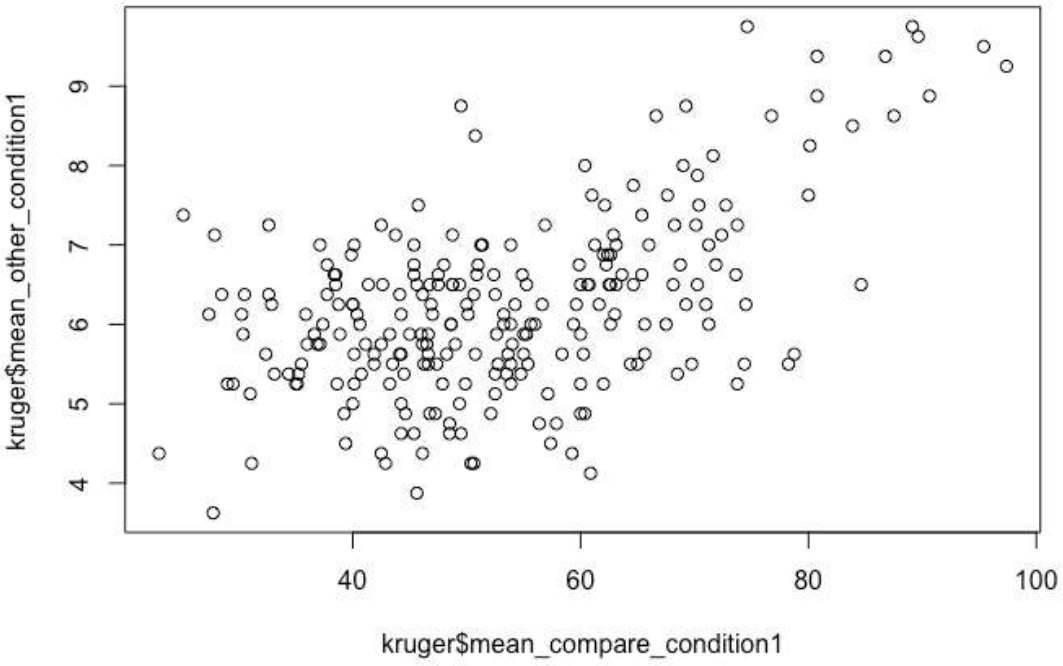


Figure 2.6. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the replication group.

Non-excluded

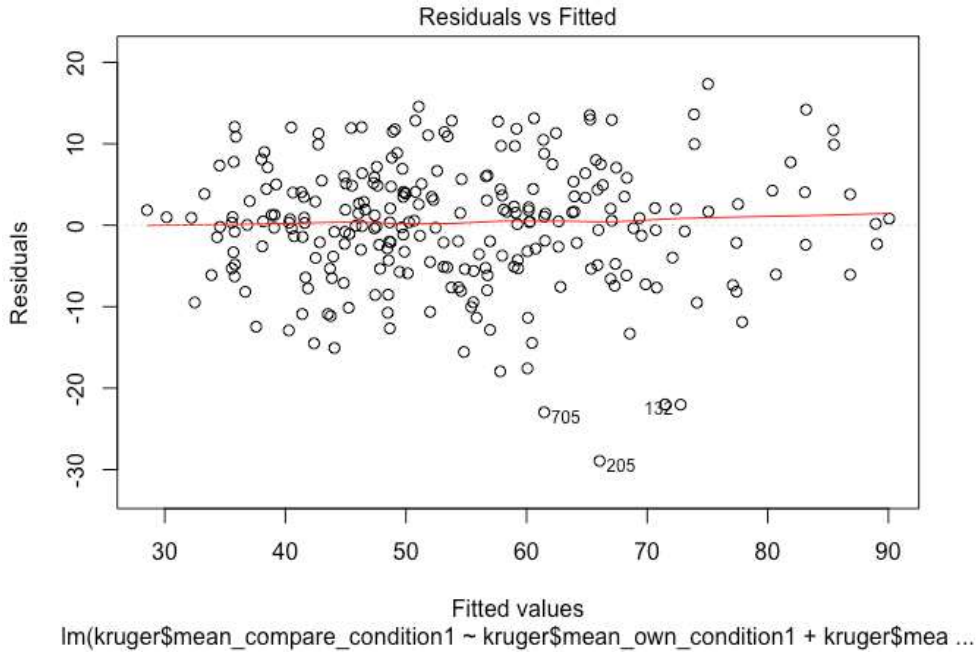


Figure 2.7. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the replication group.

Excluded

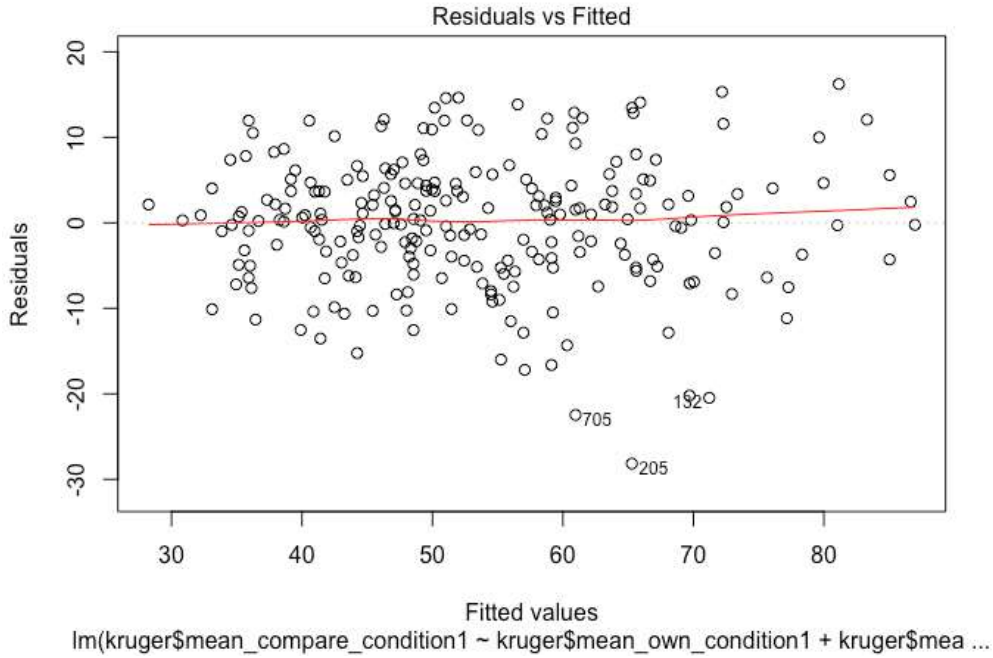


Figure 2.8. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the replication group.

Non-excluded

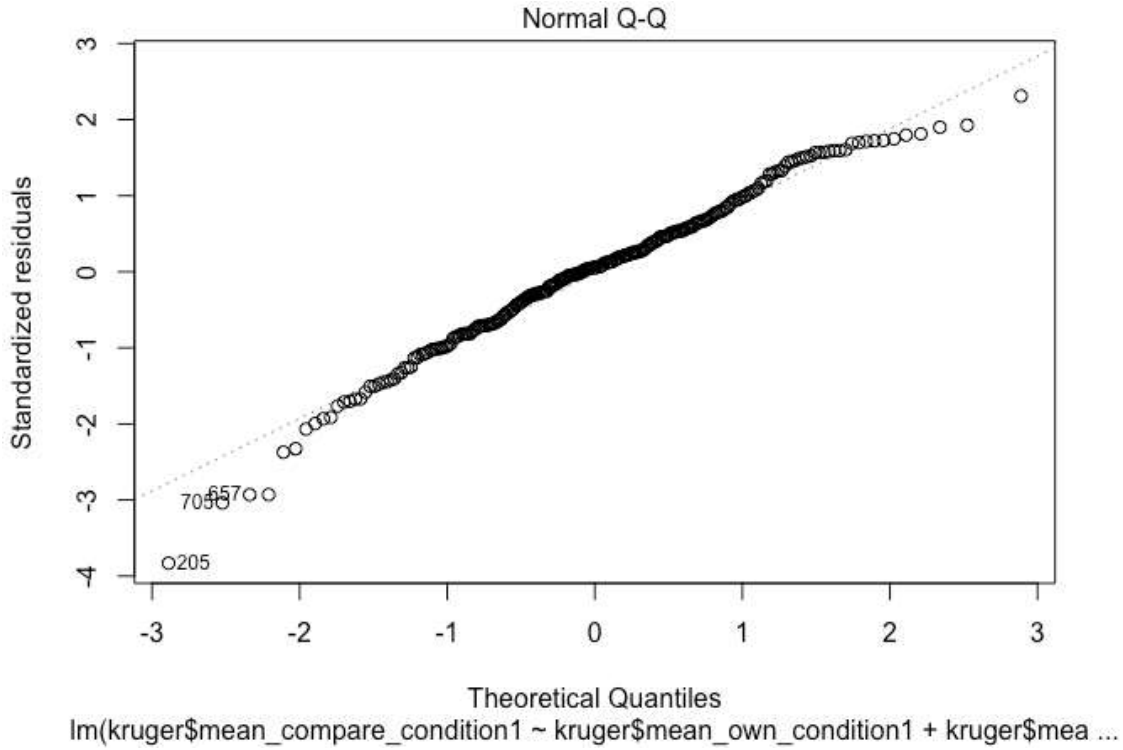


*Figure 2.*9. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the replication group.

Excluded

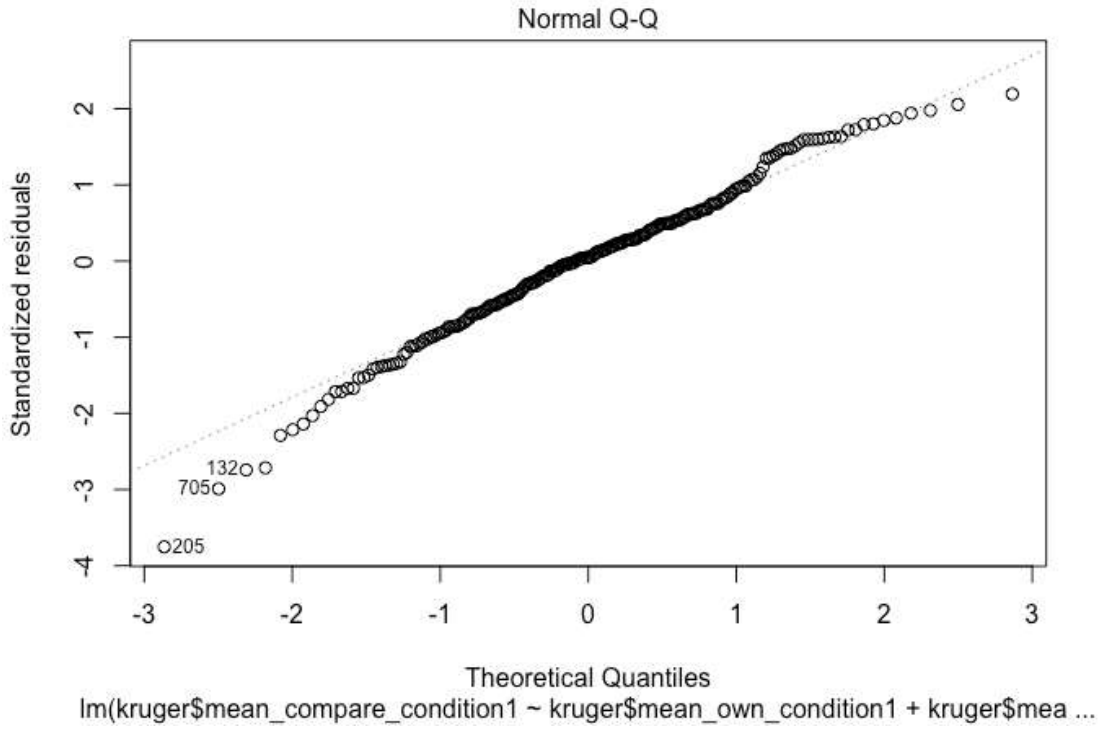


*Figure 2.*10. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the replication group.

Correlation Matrix Condition 1

Table 5.5
Person's r for mean values (across abilities) in the replication (original) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.16	1.00				
Own ability	0.85	0.21	1.00			
Peers' ability	0.53	0.34	0.67	1.00		
Desirability	0.11	0.06	0.14	0.20	1.00	
Ambiguity	-0.06	-0.01	-0.11	-0.13	-0.35	1.00

Table 5.6
P-values for correlations between mean values (across abilities) in the replication (original) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	.012	0				
Own ability	<.001	.001	0			
Peers' ability	<.001	<.001	<.001	0		
Desirability	0.090	0.362	.025	.002	0	
Ambiguity	.340	.829	.093	.039	<.001	0

Easy domain condition

Non-excluded

Table 6.1

Easy domain group: partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	0.35	0.33	0.139	0.02 [-0.01, 0.05]
Comparative ability, domain difficulty	Ambiguity	0.35	0.35	0.097	-0.01 [-0.03, 0.00]

Note. The replication group did not find a significant difference between the zero order correlation ***(p.xy)*** and partial correlation ***(p.xy.z)*** between comparative ability and domain difficulty ratings while holding desirability or ambiguity constant. Both tests were not able to reject the null hypothesis: ρ.xy - ρ.xy.z = 0.

Excluded

Table 6.1

Easy domain group: partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	0.32	0.29	0.052	0.03 [-0.001, 0.07]
Comparative ability, domain difficulty	Ambiguity	0.32	0.32	0.188	-0.01 [-0.03, 0.00]

Note. The replication group did not find a significant difference between the zero order correlation ***(p.xy)*** and partial correlation ***(p.xy.z)*** between comparative ability and domain difficulty ratings while holding desirability or ambiguity constant. Both tests were not able to reject the null hypothesis: ρ.xy - ρ.xy.z = 0.

Non-excluded

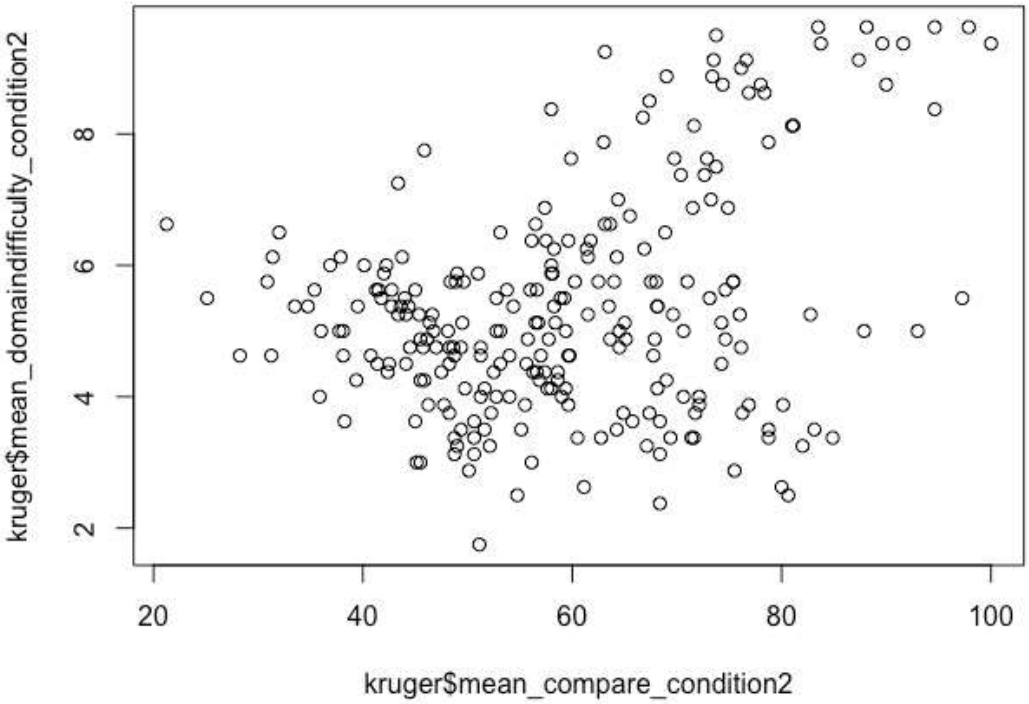


Figure 3.1. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the easy domain group.

Excluded

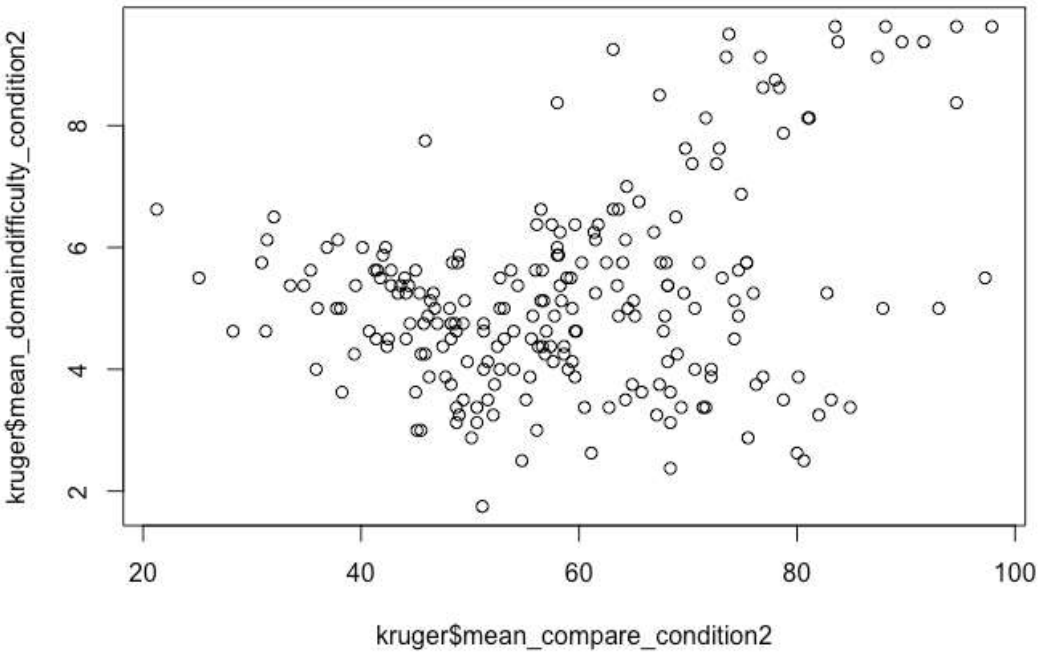


Figure 3.2. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the easy domain group.

Non-excluded

Table 6.3

Easy domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	2.20	[-3.53, 7.93]						
kruger$mean_own_condition2	8.69**	[7.69, 9.69]	0.83	[0.74, 0.93]	.38	[.30, .47]	.82**	
kruger$mean_other_condition2	-0.14	[-1.21, 0.94]	-0.01	[-0.11, 0.08]	.00	[-.00, .00]	.54**	
								R^2 = .679**
								95% CI[0.61,0.74]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Excluded

Table 6.4

Easy domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	2.90	[-2.99, 8.80]						
kruger$mean_own_condition2	8.92**	[7.90, 9.93]	0.86	[0.76, 0.96]	.42	[.33, .51]	.83**	
kruger$mean_other_condition2	-0.50	[-1.59, 0.59]	-0.04	[-0.14, 0.05]	.00	[-.00, .01]	.52**	
								R^2 = .690**
								95% CI[0.62,0.76]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Non-excluded

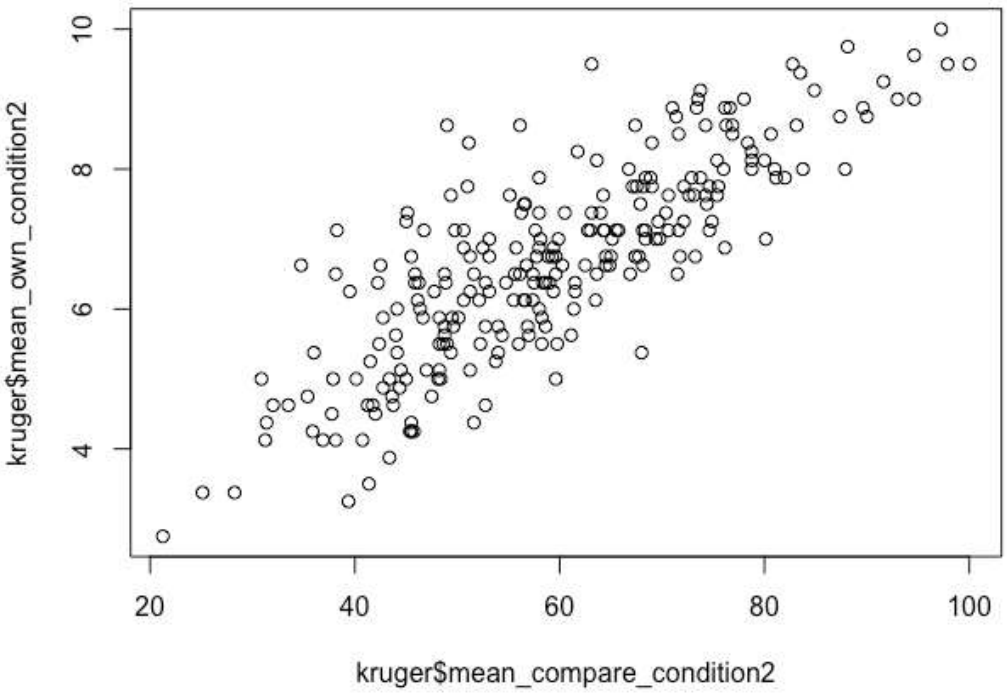


Figure 3.3. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the easy domain group.

Excluded

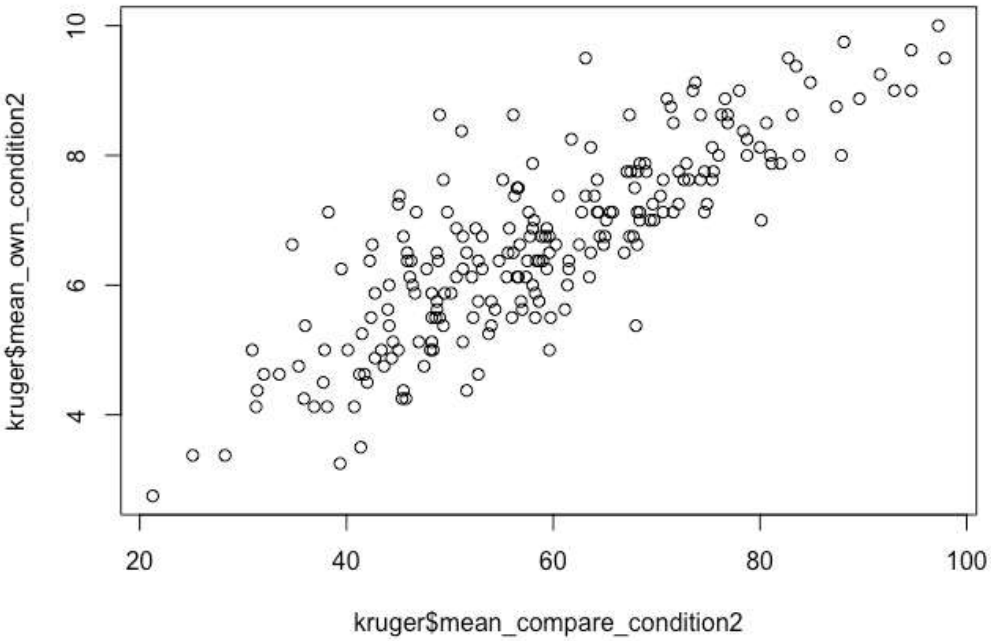


Figure 3.4. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the easy domain group.

Non-excluded

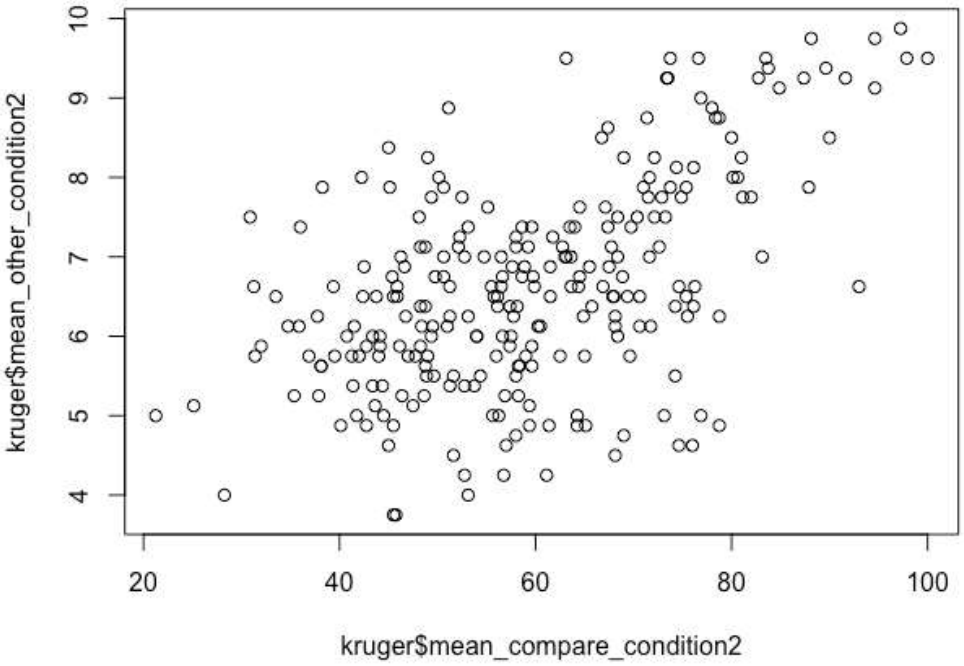


Figure 3.5. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the easy domain group.

Excluded

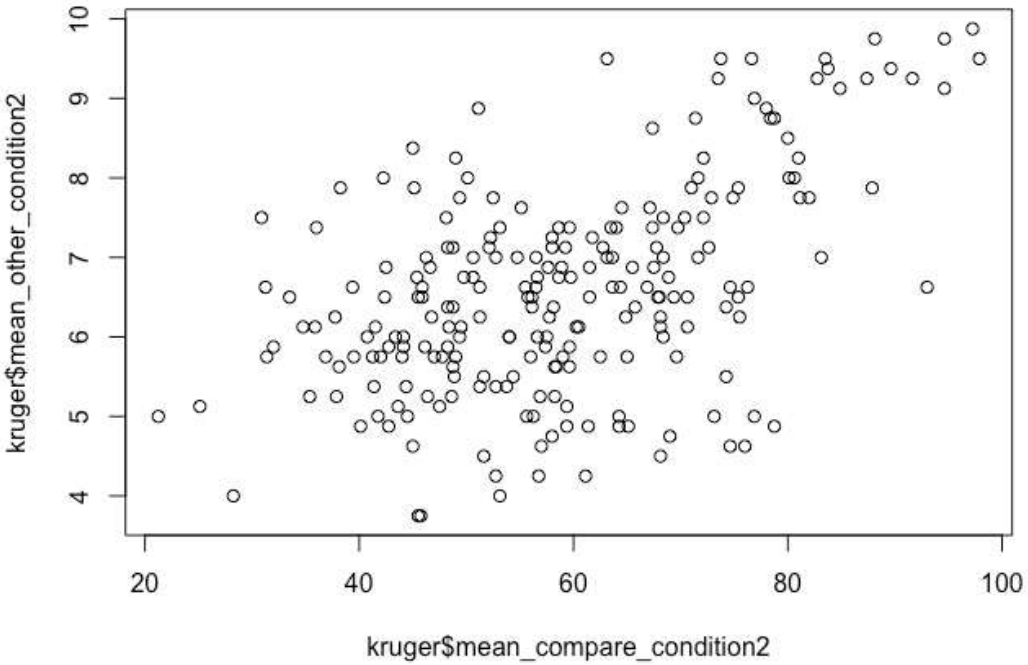


Figure 3.6. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the easy domain group.

Non-excluded

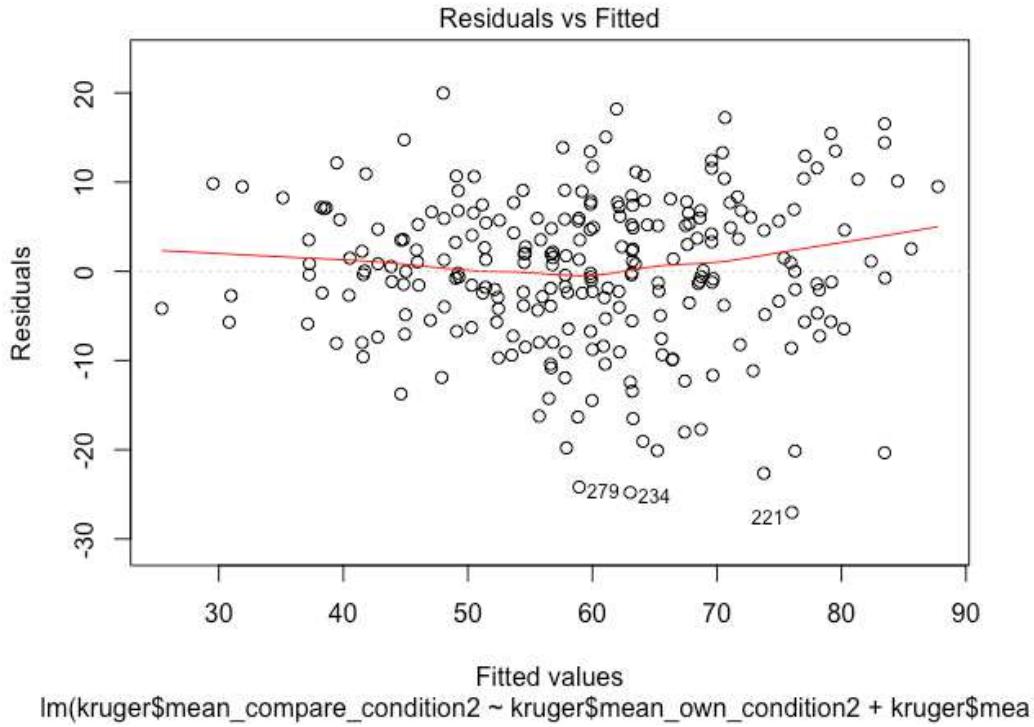


Figure 3.7. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

Excluded

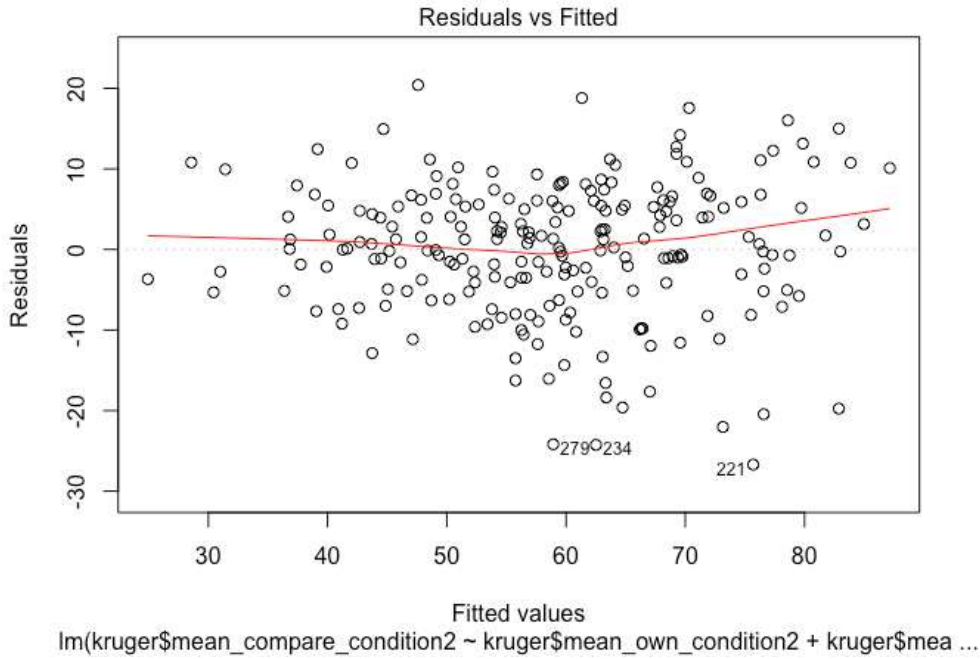


Figure 3.8. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

Non-excluded

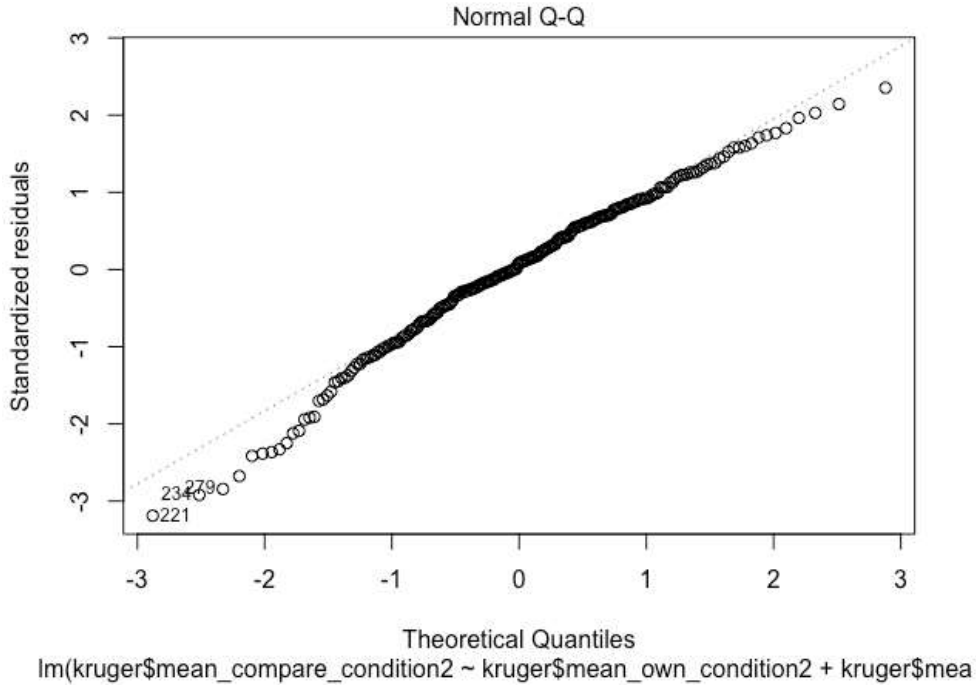


*Figure 3.*9. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

Excluded

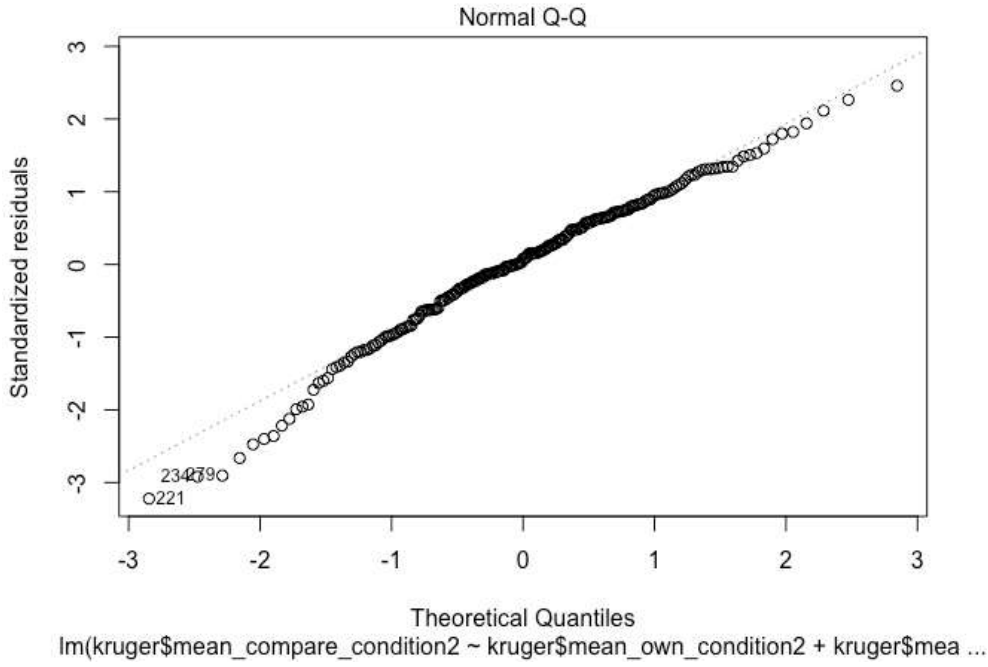


*Figure 3.*10. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the easy domain group.

Correlation Matrix Condition 2

Table 6.5

Person's r for mean values (across abilities) in the easy (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.83	1.00				
Own ability	0.32	0.28	1.00			
Peers' ability	0.52	0.66	0.37	1.00		
Desirability	0.29	0.40	0.15	0.41	1.00	
Ambiguity	0.08	0.18	-0.06	0.23	0.43	1.00

Table 6.6

P-values for correlations between mean values (across abilities) in the easy (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	<.001	0				
Own ability	<.001	<.001	0			
Peers' ability	<.001	<.001	<.001	0		
Desirability	<.001	<.001	0.0228	<.001	0	
Ambiguity	0.2630	0.0056	0.3420	0.0006	<.001	0

Difficult domain condition

Non-excluded

Table 7.1.

Difficult domain group: partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	-0.07	-0.15	<.001	0.08 [0.04, 0.14]
Comparative ability, domain difficulty	Ambiguity	-0.07	-0.07	.99	-0.001 [-0.05, 0.05]

Note. The replication group found a significant difference between the zero order correlation *(p.xy)* and partial correlation *(p.xy.z)* between comparative ability and domain difficulty ratings while holding desirability constant. No significant difference was found between the zero order correlation *(p.xy)* and partial correlation *(p.xy.z)* between comparative ability and domain difficulty ratings while holding ambiguity constant.

Excluded

Table 7.2.

Difficult domain group: partial correlations between mean comparative ability estimates and domain difficulty, holding desirability or ambiguity ratings constant

Variables	**Control variable**	**Correlation *(p.xy)***	**Partial correlation *(p.xy.z)***	***p***	**Difference *(p.xy - p.xy.z)***
Comparative ability, domain difficulty	Desirability	-0.13	-0.20	<.001	0.06 [0.03, 0.12]
Comparative ability, domain difficulty	Ambiguity	-0.13	-0.13	.954	-0.003 [-0.04, 0.04]

Note. The replication group found a significant difference between the zero order correlation *(p.xy)* and partial correlation *(p.xy.z)* between comparative ability and domain difficulty ratings while holding desirability constant. No significant difference was found between the zero order correlation *(p.xy)* and partial correlation *(p.xy.z)* between comparative ability and domain difficulty ratings while holding ambiguity constant.

Non-excluded

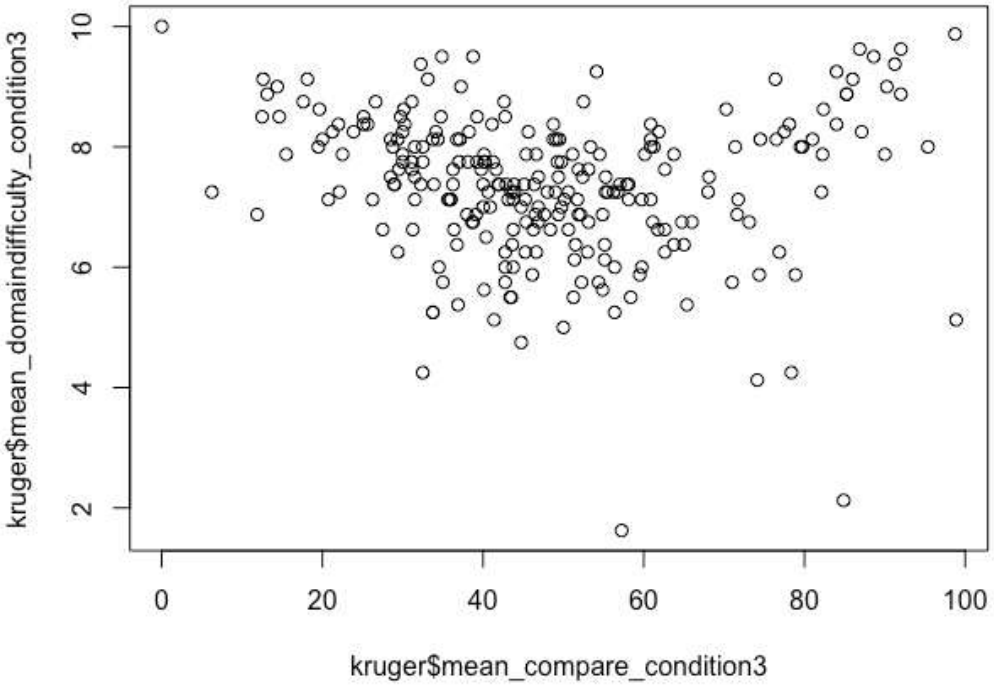


Figure 4.1. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the difficult domain group.

Excluded

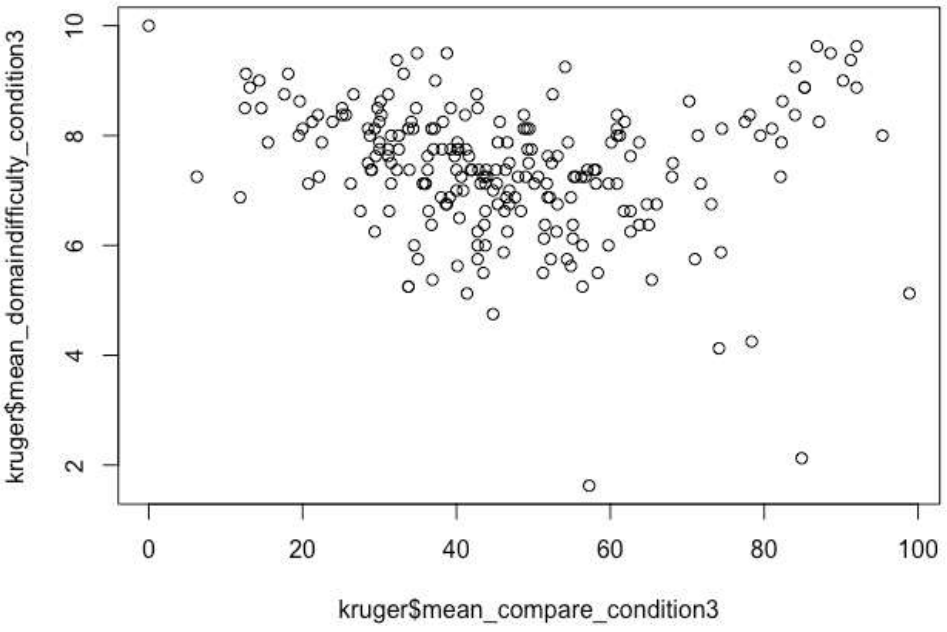


Figure 4.2. Scatterplot for the correlation between mean comparative ability and mean domain difficulty ratings in the difficult domain group.

Non-excluded

Table 7.3.

Difficult domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	8.765	[5.21, 12.28]						
kruger$mean_own_condition2	8.32**	[7.28, 9.36]	0.90	[0.78, 1.01]	.24	[.17, .31]	.87**	
kruger$mean_other_condition2	-0.28	[-1.46, 0.90]	-0.03	[-0.14, 0.09]	.00	[-.00, .00]	.73**	
								R^2 = .764**
								95% CI[0.71, 0.81]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Excluded
Table 7.4.

Difficult domain condition: regression results using mean comparative ability as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	9.10**	[5.40, 12.80]						
kruger$mean_own_condition3	8.39**	[7.33, 9.45]	0.90	[0.79, 1.01]	.27	[.19, .34]	.87**	
kruger$mean_other_condition3	-0.42	[-1.62, 0.78]	-0.04	[-0.15, 0.07]	.00	[-.00, .00]	.70**	
								R^2 = .755**
								95% CI[0.70, 0.81]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. ** indicates $p < .01$.

Non-excluded

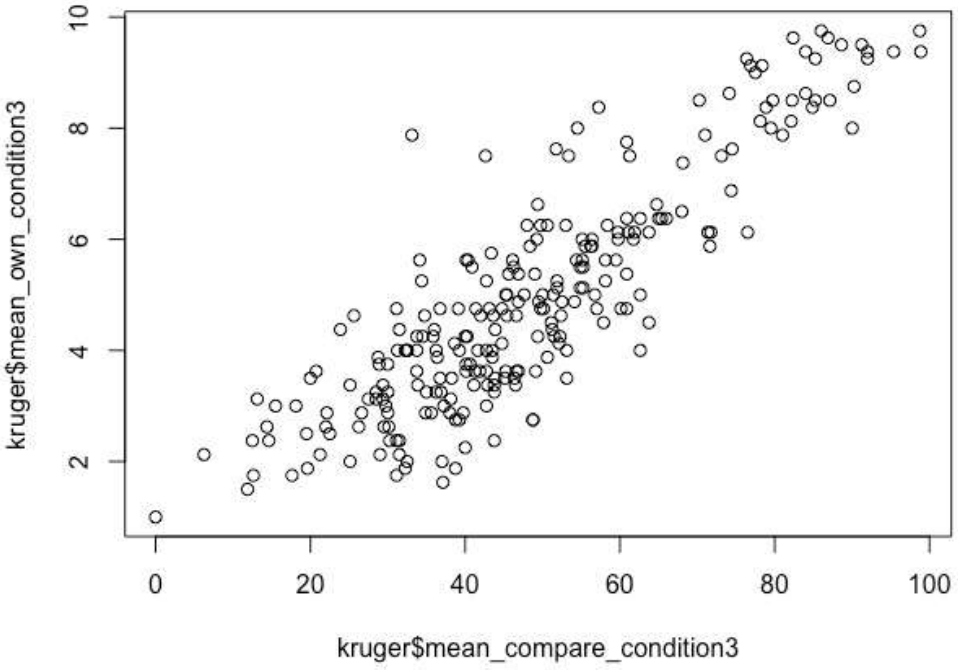


Figure 4.3. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the difficult domain group.

Excluded

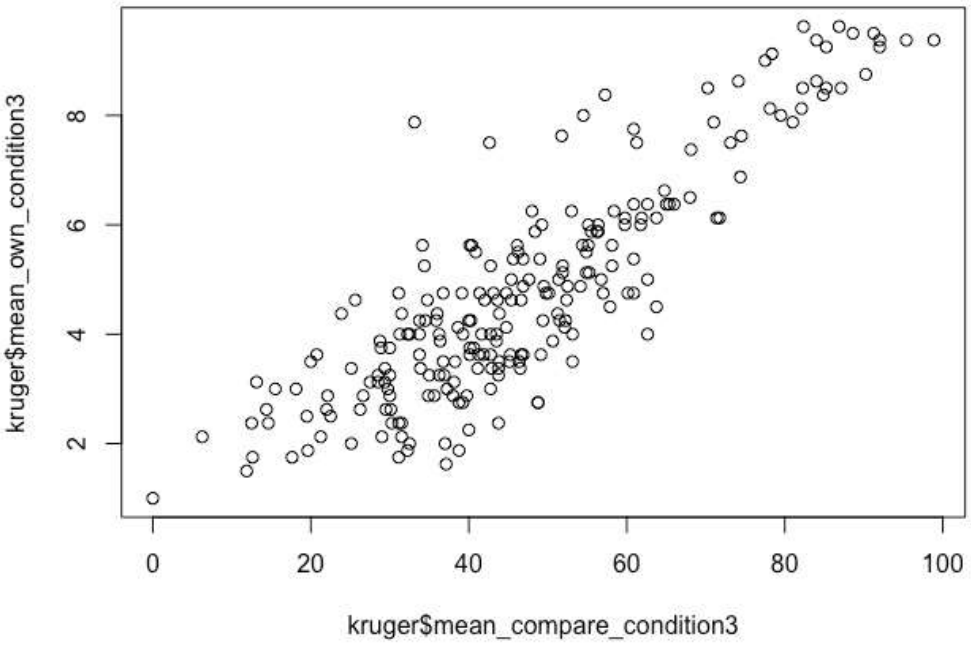


Figure 4.4. Scatterplot for the correlation between mean comparative ability and mean own absolute ability ratings in the difficult domain group.

Non-excluded

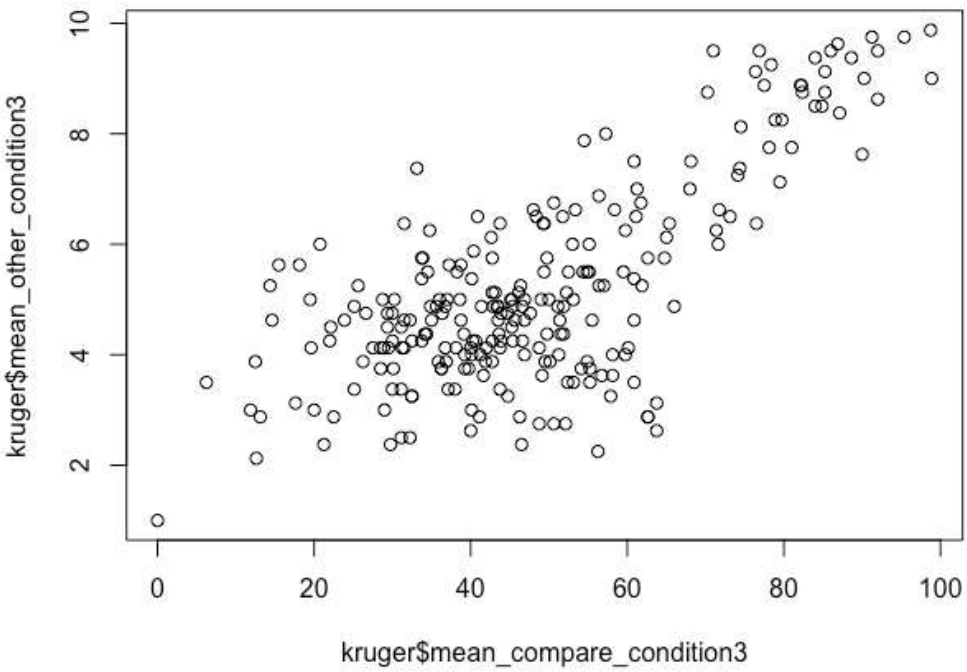


Figure 4.5. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the difficult domain group.

Excluded

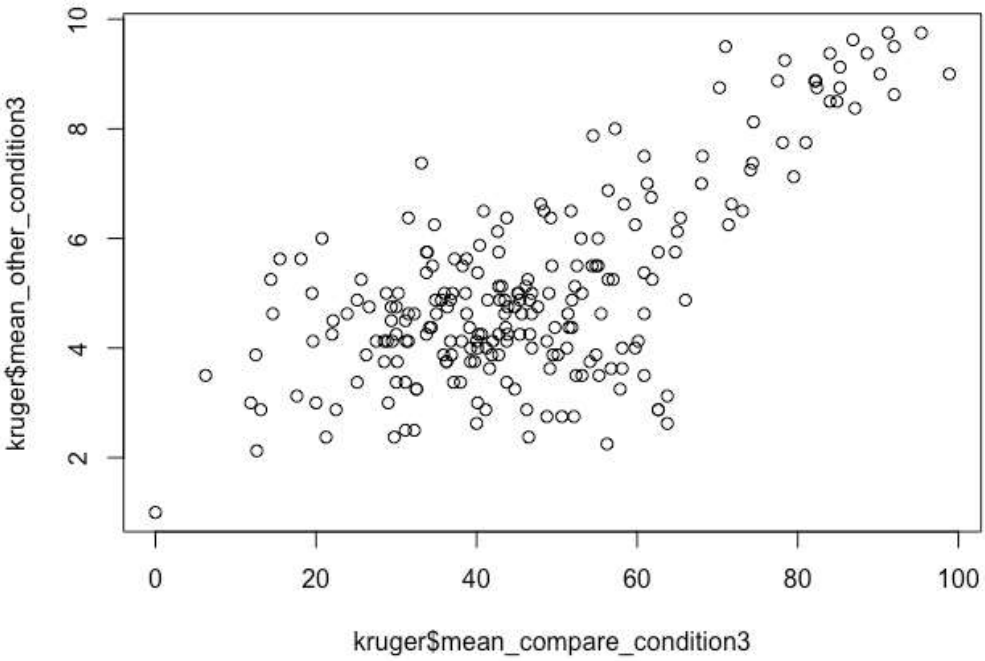


Figure 4.6. Scatterplot for the correlation between mean comparative ability and mean others' absolute ability ratings in the difficult domain group.

Non-excluded

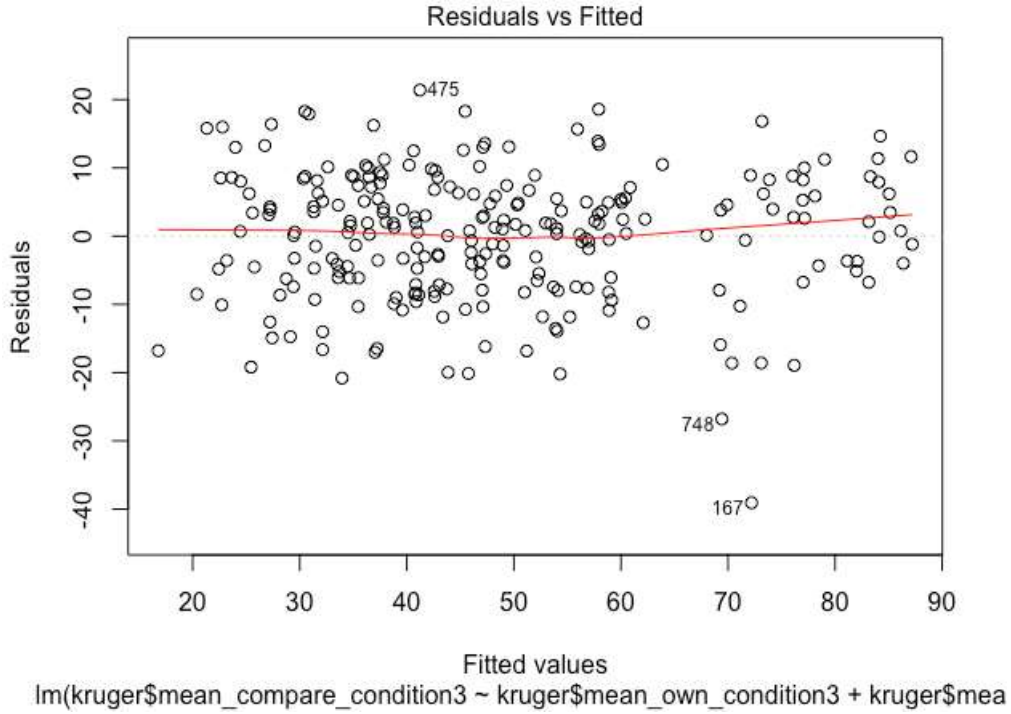


Figure 4.7. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

Excluded

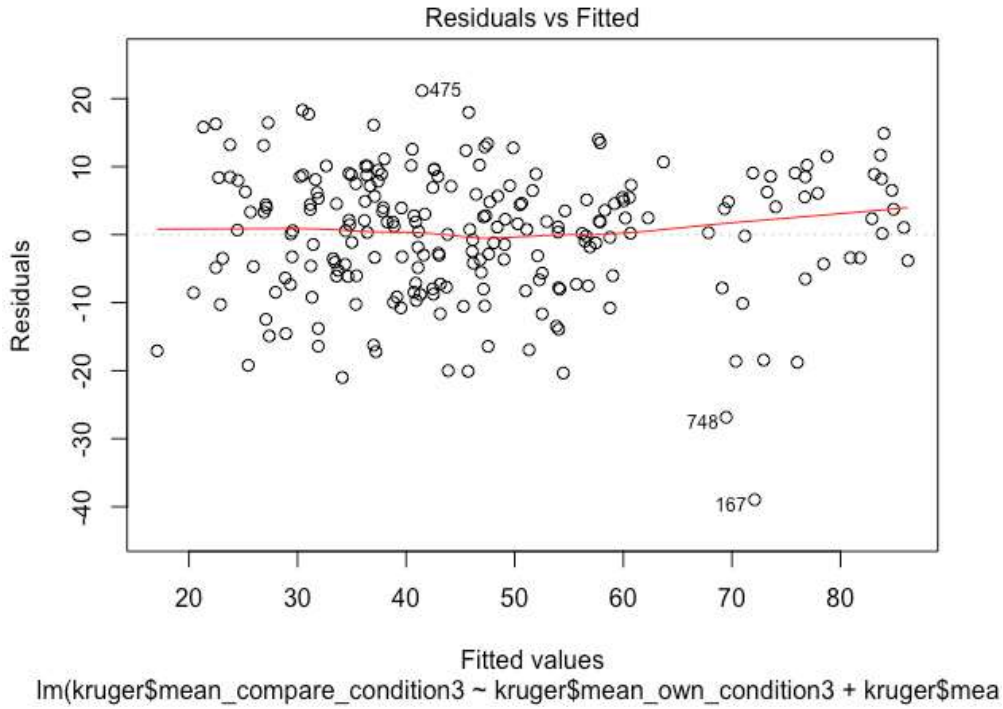


Figure 4.8. Residuals versus fitted plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

Non-excluded

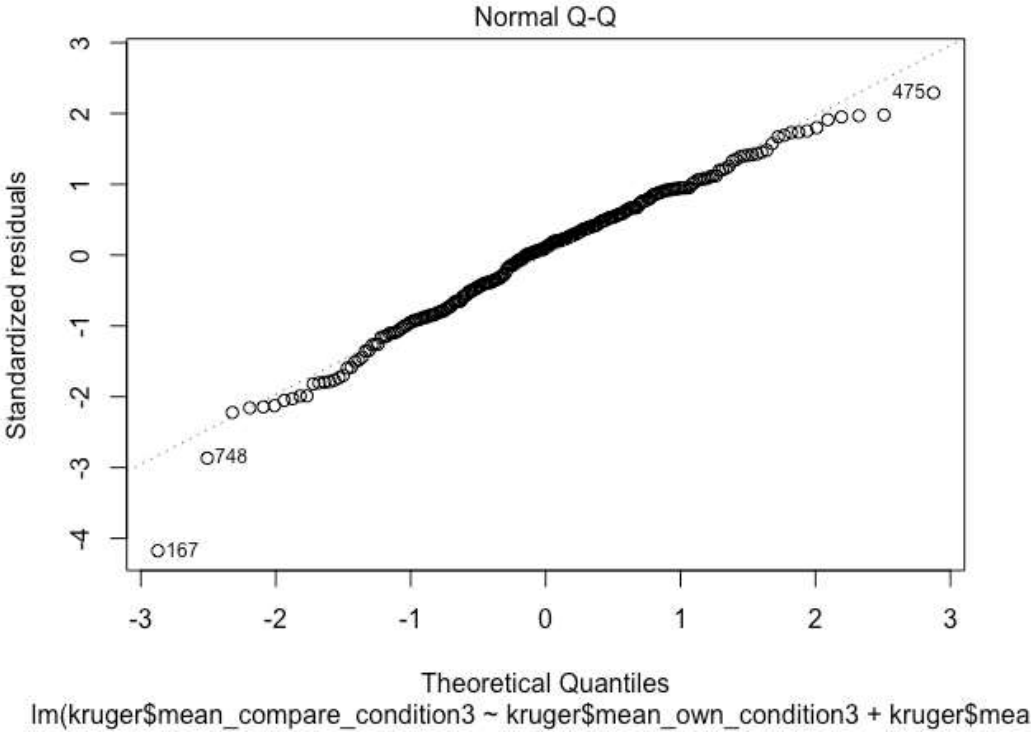


*Figure 4.*9. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

Excluded

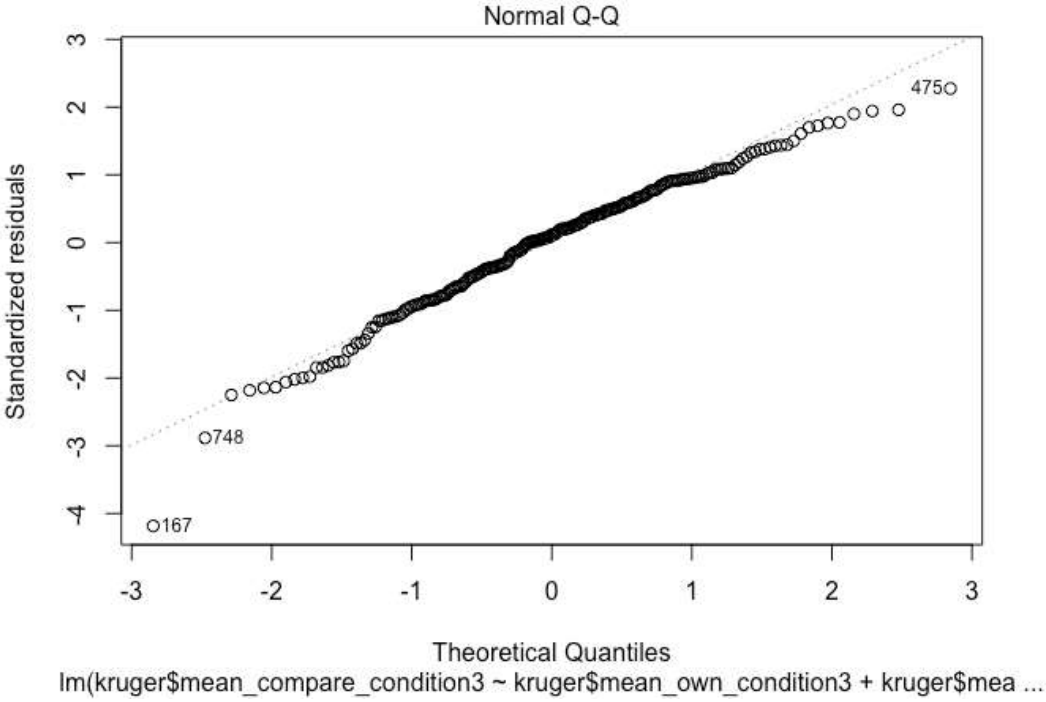


*Figure 4.*10. Normal Q-Q plot for mean comparative ability predicted from mean own and others' ability in the difficult domain group.

Correlation Matrices Condition 3

Table 7.5
Person's r for mean values (across abilities) in the difficult (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	1.00					
Difficulty	0.86	1.00				
Own ability	-0.13	-0.13	1.00			
Peers' ability	0.70	0.82	-0.03	1.00		
Desirability	0.20	0.13	0.27	0.22	1.00	
Ambiguity	-0.03	-0.09	0.27	0.01	0.43	1.00

Table 7.6
P- values for correlations between mean values (across abilities) in the difficult (extension) condition

	Comparative Ability	Difficulty	Own ability	Peers' ability	Desirability	Ambiguity
Comparative Ability	0					
Difficulty	<.0001	0				
Own ability	0.0498	0.0595	0			
Peers' ability	<.0001	<.0001	0.6070	0		
Desirability	0.0025	0.0467	<.0001	0.0007	0	
Ambiguity	0.6860	0.1850	<.0001	0.8370	<.0001	0

Comparisons between the three conditions

Excluded

Table 8.2.

Student's t-tests comparing ratings between the replication and easy or difficult domain groups

Variable	*n*	*Mean*	*SD*	*t-statistic*	*df*	*p*	*d*	95% confidence intervals	
								Lower	Upper
Replication and easy domain conditions									
Domain difficulty (replication)	240	6.05	1.15	6.38	463	<.001	0.59	0.40	0.78
Domain difficulty (easy domain)	225	5.22	1.63						
Ambiguity (replication)	240	8.00	1.24	-2.79	463	.005	0.26	0.08	0.44
Ambiguity (easy domain)	225	8.32	1.23						
Replication and difficult domain conditions									
Domain difficulty (replication)	240	6.05	1.15	-12.27	464	<.001	1.15	0.95	1.35
Domain difficulty (difficult domain)	226	7.39	1.19						

Ambiguity (replication)	240	8.00	1.24	-1.92	464	.055	0.18	-0.00	0.36
Ambiguity (difficult domain)	226	8.24	1.43						

One-sample Wilcoxon-tests

Table 9.1

One-sample Wilcoxon tests testing median comparative ability scores against the scale mid-point in the original (replication) condition

		90% CI			
Item	Median	Lower	Upper	P-value	Effect size *r*
Using mouse	75.50	73.00	78.00	<.001	0.81
Driving	70.50	67.70	73.00	<.001	0.61
Riding bicycle	64.50	61.00	67.50	<.001	0.52
Saving money	64.50	61.50	67.50	<.001	0.53
Telling jokes	53.00	50.00	56.50	0.0569	0.13
Playing chess	40.00	36.00	44.00	<.001	0.30
Juggling	28.00	24.00	33.50	<.001	0.55
Computer programming	40.00	36.00	44.00	<.001	0.30

Table 9.2

One-sample Wilcoxon tests testing median comparative ability scores against the scale mid-point in the easy domain (extension) condition

		90% CI			
Item	Median	Lower	Upper	P-value	Effect size *r*
Using mouse	76.50	74.00	79.50	<.001	0.76
Driving	71.00	67.50	74.00	<.001	0.62
Riding bicycle	69.00	65.60	70.50	<.001	0.63
Saving money	66.50	62.50	70.50	<.001	0.49
Telling jokes	62.00	58.50	65.50	<.001	0.43
Playing chess	47.50	43.50	52.00	0.26	0.06
Juggling	46.00	41.00	50.50	0.085	0.08
Computer programming	49.50	46.00	53.50	0.85	0.01

Table 9.3

One-sample Wilcoxon tests results testing against the scale mid-point for comparative ability in difficult domain (extension) condition

		90% CI			
Item	Median	Lower	Upper	P-value	Effect size *r*

Using mouse	59.00	55.50	62.50	<.001	0.33
Driving	39.50	34.50	44.00	<.001	0.30
Riding bicycle	49.50	45.00	53.50	0.69	0.02
Saving money	66.50	62.50	70.50	<.001	0.46
Telling jokes	40.00	35.50	44.00	<.001	0.30
Playing chess	40.50	36.00	44.50	<.001	0.25
Juggling	37.50	33.00	42.00	<.001	0.35
Computer programming	45.00	40.50	49.00	0.01	0.16

Table 9.4
One-sample Wilcoxon-tests results testing against the scale mid-point for desirability in the easy domain (extension) condition

		90% CI			
Item	Median	Lower	Upper	P-value	Effect size *r*
Using mouse	9.00	8.50	9.00	<.001	0.88
Driving	9.50	9.00	9.50	<.001	0.89
Riding bicycle	8.00	8.00	8.50	<.001	0.85
Saving money	9.50	9.00	9.50	<.001	0.89
Telling jokes	7.50	7.50	8.00	<.001	0.80
Playing chess	7.50	8.00	8.50	<.001	0.80
Juggling	6.50	7.00	7.50	<.001	0.56
Computer programming	9.00	9.00	9.00	<.001	0.87

Table 9.5
One-sample Wilcoxon-tests results testing against the scale mid-point for desirability in easy domain (extension) condition

		90% CI			
Item	Median	Lower	Upper	P-value	Effect size *r*
Using mouse	8.50	8.50	9.00	<.001	0.84
Driving	9.00	9.00	9.50	<.001	0.87
Riding bicycle	8.00	8.00	8.50	<.001	0.78
Saving money	9.00	8.50	9.00	<.001	0.87
Telling jokes	8.00	7.50	8.00	<.001	0.81
Playing chess	8.00	8.00	8.50	<.001	0.76
Juggling	7.00	6.50	7.50	<.001	0.57
Computer programming	8.50	8.50	8.50	<.001	0.84

Table 9.6
One-sample Wilcoxon-tests results testing against the scale mid-point for desirability in difficult domain (extension) condition

Item	Median	90% CI		P-value	Effect size *r*
		Lower	Upper		
Using mouse	7.00	8.50	9.00	<.001	0.61
Driving	8.50	9.00	9.50	<.001	0.81
Riding bicycle	8.00	8.00	8.50	<.001	0.78
Saving money	8.50	8.50	9.00	<.001	0.84
Telling jokes	8.00	7.50	8.00	<.001	0.74
Playing chess	8.50	8.00	8.50	<.001	0.79
Juggling	8.00	6.50	7.50	<.001	0.67
Computer programming	8.00	8.50	8.50	<.001	0.72

One-sample t-tests

Table 10.1

One-sample t-tests results testing against the scale mid-point for comparative ability in original (replication) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	71.20	18.34	1.6961E-47	239	68.92	73.47
Driving	65.17	10.64	6.7221E-22	239	62.36	67.97
Riding bicycle	61.01	8.33	6.3572E-15	239	58.41	63.62
Saving money	62.88	9.45	3.1702E-18	239	60.19	65.56
Telling jokes	52.42	1.66	0.0987806	239	49.54	55.30
Playing chess	40.98	-5.18	4.8043E-07	239	37.55	44.41
Juggling	31.98	-10.09	3.5936E-20	239	28.46	35.50
Computer programming	40.73	-4.92	1.6472E-06	239	37.01	44.44

Table 10.2

One-sample t-tests results testing against the scale mid-point for comparative ability in easy domain (extension) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	71.27	15.56	1.72E-37	224	68.58	73.97
Driving	66.32	10.77	4.63E-22	224	63.34	69.31
Riding bicycle	65.76	10.78	4.26E-22	224	62.88	68.64
Saving money	63.63	8.00	6.58E-14	224	60.27	66.98
Telling jokes	59.74	7.11	1.56E-11	224	57.04	62.44
Playing chess	47.81	-1.19	0.23413041	224	44.19	51.43
Juggling	46.76	-1.75	0.08125851	224	43.11	50.41
Computer programming	49.57	-0.25	0.80363761	224	46.20	52.95

Table 10.3

One-sample t-tests results testing against the scale mid-point for comparative ability in difficult domain (extension) condition

Item	estimate	statistic	p-value	df	Lower CI	Upper CI
Using mouse	55.79	4.12	5.2822E-05	225	53.02	58.56
Driving	40.63	-4.81	2.7508E-06	225	36.79	44.47
Riding bicycle	48.90	-0.60	0.54982103	225	45.29	52.51
Saving money	62.68	7.40	2.7397E-12	225	59.30	66.06
Telling jokes	40.82	-5.10	7.0415E-07	225	37.28	44.37
Playing chess	41.86	-4.49	1.1181E-05	225	38.29	45.43
Juggling	39.67	-5.61	5.7662E-08	225	36.04	43.29
Computer programming	45.36	-2.68	0.00792618	225	41.95	48.77

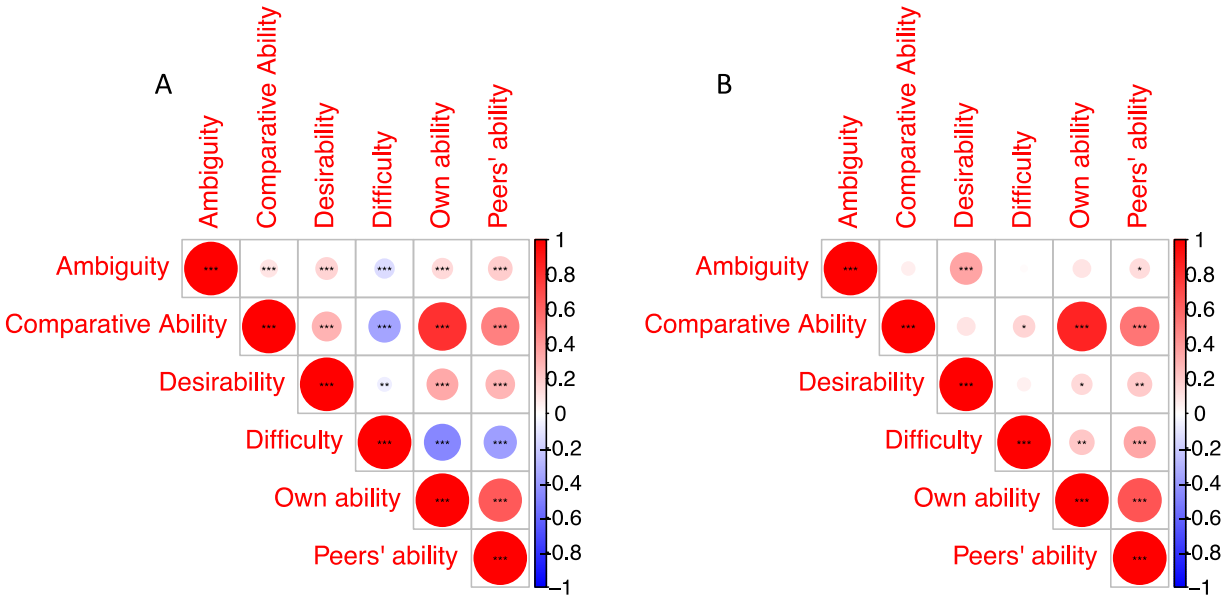
Table 10.4

One-sample t-tests results testing against the scale mid-point for desirability in original (replication) condition

Item	estimate	statistic	p-value	df	d	Lower CI	Upper CI
Using a computer mouse	8.654	34.058	1.11E-93	239	2.203	1.968	2.435
Driving	9.146	47.234	3.34E-123	239	3.055	2.750	3.354
Riding bicycle	7.996	27.992	2.14E-77	239	1.811	1.604	2.016
Saving money	9.129	45.568	7.60E-120	239	2.948	2.651	3.237
Telling jokes	7.500	20.858	9.87E-56	239	1.349	1.173	1.523
Playing chess	7.613	20.269	7.96E-54	239	1.311	1.138	1.483
Juggling	6.529	10.278	9.33E-21	239	0.665	0.524	0.804
Computer programming	8.663	33.998	1.58E-93	239	2.199	1.964	2.431

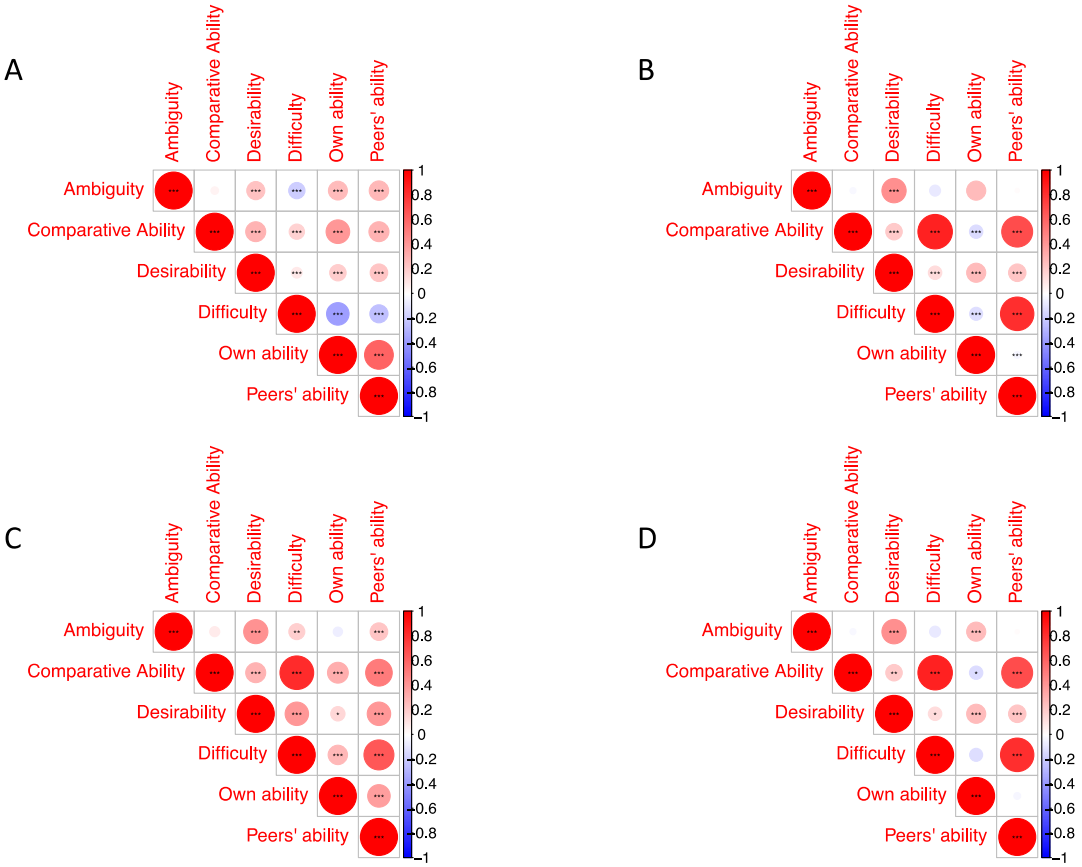
Correlation Matrices for all DV's Across Conditions

Figure 2.11
Correlation matrix for domains in the replication condition.



Note: (Person's r) for variables across abilities in the replication condition. Panel A: Absolute value correlations. Panel B: Mean value correlations. See tables 5.5-5.8 in the supplementary for exact *r*- and *p*-values. *p <.05. **p <.01. ***p <.001.

Figure 2.12
Correlation matrices across abilities in the extension conditions.



Panel A: easy extension condition absolute scores. Panel B difficult extension condition absolute scores. Panel C: easy extension condition mean scores. Panel D difficult extension condition mean scores.

*p <.05. **p <.01. ***p <.001

Note: Bigger circles indicate stronger correlations. See supplementary Tables 6.5-6.8 and 7.5-7.8 for exact r and p-values.

Power Simulation for Exploratory Analysis

Table 10.5
*Power Simulation for Main and Interaction effects for 3(Condition)*2(Difficulty) mixed design*

Effect	Power	Effect Size
Condition	100	0.12353699
Difficulty	100	0.28369743
Interaction	99.7	0.04595521

Power simulations in R using the "Superpower" package (Lakens & Caldwell, 2021) showed that using our sample of *n* = 691 (with sample size by cell/between factor: replication group n = 240, easy extension n = 225, difficult extension n = 226), near 100% power was reached to examine main & interaction effects. For some of the multiple comparisons we have however power close to 0%

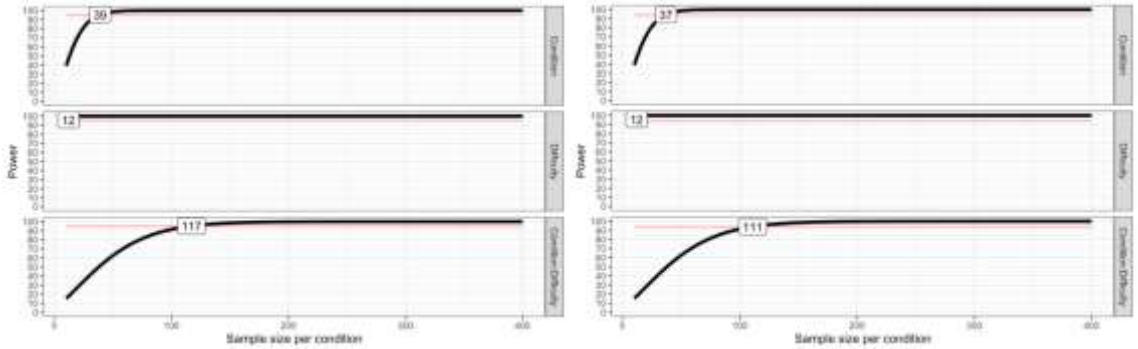
Comparison	Power	Effect Size
Condition_Replication_Difficulty_Easy VS Condition_Replication_Difficulty_Difficult	100	-1.3760623
Condition_Replication_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Easy	**3.8**	**0.10723954**
Condition_Replication_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Difficult	100	-0.8515678
Condition_Replication_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Easy	100	-0.7868727
Condition_Replication_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	100	-1.2543963
Condition_Replication_Difficulty_Difficult VS Condition_Easy Extension_Difficulty_Easy	100	1.39440291
Condition_Replication_Difficulty_Difficult VS Condition_Easy Extension_Difficulty_Difficult	98.5	0.48786589
Condition_Replication_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Easy	99.2	0.53517156
Condition_Replication_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Difficult	**0.5**	**0.01912579**
Condition_Easy Extension_Difficulty_Easy VS Condition_Easy Extension_Difficulty_Difficult	100	-0.8971173
Condition_Easy Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Easy	100	-0.8359101
Condition_Easy Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	100	-1.2817852
Condition_Easy Extension_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Easy	**1**	**0.05121554**
Condition_Easy Extension_Difficulty_Difficult VS Condition_Difficult Extension_Difficulty_Difficult	94.7	-0.4397807
Condition_Difficult Extension_Difficulty_Easy VS Condition_Difficult Extension_Difficulty_Difficult	98.2	-0.4847737

Those are:

(1) Replication Condition Easy items compared to Easy Extension Condition Easy Items,
(2) Replication Condition Difficult Items vs Difficult Extension Difficult Items, and
(3) Easy Extension Difficulty Items vs Difficult Condition Easy Items.

Moreover, the sample size per cell was slightly too small to reach 95% power

Figure 2.13
*Power Curves for Main and Interaction effects for 3(Condition)*2(Difficulty) mixed design*



Left panel: 95% power. Right panel: 94% power.
Note: the sample size required to reach 95% power was n = 117 for each cell. Using this calculation, our collected sample was slightly too small for the extension conditions (easy extension: n = 225/2 = 112 each cell, and difficult extension: n = 226/2 = 113 each cell).

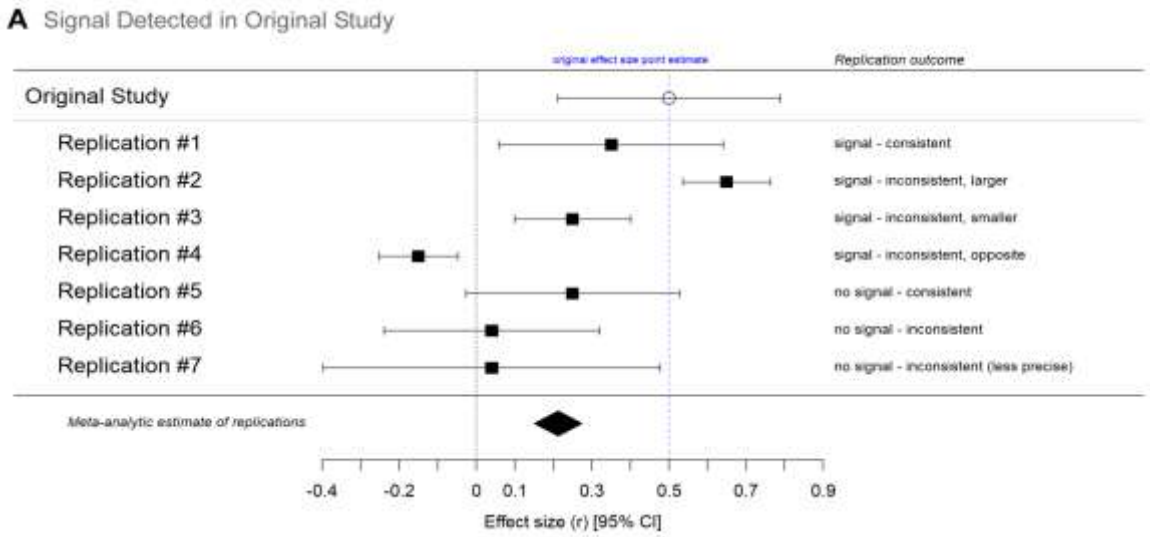
Equivalence Tests

EQ Test #	Correlation	Variable Controlled for	df	*r*diff	Upper 95% CI	Lower 95% CI	*p*	Condition
1	Comparative ability & domain difficulty (as in original)	desirability	5	0.08604	0.018	0.212	0.91	Replication
2		ambiguity	5	-0.0085	-0.012	-0.005	0.33	Replication
3	Comparative ability & domain difficulty (vector-wise)	desirability	1917	-0.0033	-0.016	0.007	0.55	Replication
4		ambiguity	1917	-0.0085	-0.016	-0.004	0.002	Replication
5	Comparative ability & domain difficulty (mean)	desirability	237	0.00527	-0.003	0.029	0.22	Replication
6		ambiguity	237	0.00054	-0.008	0.016	0.69	Replication
7	Comparative ability & domain difficulty (as in original)	desirability	5	0.09	0.02	0.159	0.005	Easy Extension
8		ambiguity	5	-0.084	-0.193	-0.011	0.03	Easy Extension
9	Comparative ability & domain difficulty (as in original)	desirability	5	0.04022	-0.238	0.126	0.67	Difficult Extension
10		ambiguity	5	0.03	-0.009	0.578	0.91	Difficult Extension
11	Comparative ability & domain difficulty (vector-wise)	desirability	1798	0.006	-0.007	0.018	0.32	Easy Extension

12		ambiguity	1798	-0.03	-0.041	-0.02	0.001	Easy Extension
13	Comparative ability & domain difficulty (vector-wise)	desirability	1806	0.023	0.015	0.034	0.002	Difficult Extension
14		ambiguity	1806	0.001	-0.0003	0.005	0.13	Difficult Extension
15	Comparative ability & domain difficulty (mean)	desirability	223	0.029	-0.0003	0.067	0.049	Easy Extension
16		ambiguity	223	-0.006	-0.028	0.002	0.15	Easy Extension
17	Comparative ability & domain difficulty (mean)	desirability	223	0.064	0.028	0.126	< .001	Difficult Extension
18		ambiguity	223	0.0025	-0.046	0.0349	0.93	Difficult Extension

Criteria for evaluation of replications

A simplified replication taxonomy for comparing replication effects to target article original findings by (LeBel et al., 2019)



Replication evaluation

We used the replication classification criteria by LeBel and colleagues' (2018) summarized in Table 6. We categorized the current replication as a "close replication" and provided details in Table 7. Variables and questions were the same as in the original, with the addition of extensions and adjustments to fit the MTurk sample, instead of Cornell university students.

Criteria for evaluation of replications by LeBel et al. (2018)

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication		Conceptual replication		
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
IV operationalization	Same/similar	Same/similar	Same/similar	Different	
DV operationalization	Same/similar	Same/similar	Same/similar	Different	
IV stimuli	Same/similar	Same/similar	Different		
DV stimuli	Same/similar	Same/similar	Different		
Procedural details	Same/similar	Different			
Physical setting	Same/similar	Different			
Contextual variables	Different				

A classification of relative methodological similarity of a replication study to an original study. "Same" ("different") indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. "Everything controllable" indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.).
"Similar" category was added to the LeBel et al. (2018) typology to refer to minor deviations, aimed to adjust the study to the target sample, that are not expected to have major implications on replication success.

Correlations per each condition

Correlation matrix for the replication condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.35*** [-0.39, -0.31]				
Own Ability	0.81*** [0.79, 0.82]	-0.46*** [-0.50, -0.43]			
Others' Ability	0.50*** [0.46, 0.53]	-0.37*** [-0.41, -0.33]	0.64*** [0.62, 0.67]		
Desirability	0.30*** [0.25, 0.34]	-0.07** [-0.11, -0.02]	0.34*** [0.30, 0.38]	0.29*** [0.25, 0.33]	
Ambiguity	-0.10*** [-0.14, -0.06]	0.13*** [0.09, 0.17]	-0.15*** [-0.19, -0.10]	-0.20*** [-0.24, -0.15]	-0.17*** [-0.21, -0.12]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Correlation matrix for the easy extension condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.27** [-0.31, -0.22]				
Own Ability	0.78*** [0.76, 0.80]	-0.37*** [-0.41, -0.33]			
Others' Ability	0.47*** [0.44, 0.51]	-0.24*** [-0.28, -0.20]	0.61*** [0.58, 0.64]		
Desirability	0.28*** [0.24, 0.32]	-0.02 [-0.06, 0.03]	0.36*** [0.31, 0.39]	0.34*** [0.29, 0.38]	
Ambiguity	-0.19*** [-0.23, -0.14]	0.19*** [0.15, 0.24]	-0.26*** [-0.30, -0.22]	-0.28*** [-0.32, -0.23]	-0.25*** [-0.29, -0.21]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Correlation matrix for the difficult extension condition

	Comparative ability	Domain Difficulty	Own Ability	Others' Ability	Desirability
Domain Difficulty	-0.31*** [-0.35, -0.27]				
Own Ability	0.78*** [0.76, 0.80]	-0.33*** [-0.37, -0.29]			
Others' Ability	0.45*** [0.41, 0.48]	-0.18*** [-0.23, -0.14]	0.56*** [0.52, 0.59]		
Desirability	0.13*** [0.08, 0.17]	0.14*** [0.09, 0.18]	0.14*** [0.09, 0.18]	0.15*** [0.10, 0.19]	
Ambiguity	-0.02 [-0.07, 0.02]	-0.03 [-0.08, 0.01]	-0.01 [-0.06, 0.04]	-0.01 [-0.06, 0.03]	-0.24*** [-0.28, -0.19]

* $p < .05$. ** $p < .01$. *** $p < .001$.

Replication condition: Item-wise correlations between domain difficulty and comparative ability ratings for each ability domain

			95% CI	
Ability domain	***p***	***r***	**Lower**	**Upper**
Using a computer mouse	.768	-0.02	-0.15	0.11
Driving	.006	-0.18	-0.30	-0.05
Riding a bicycle	.790	0.02	-0.11	0.14
Saving money	.030	-0.14	-0.26	-0.01
Telling jokes	.161	-0.09	-0.22	0.04
Playing chess	.020	-0.15	-0.27	-0.02
Juggling	<.001	-0.25	-0.37	-0.13
Computer programming	<.001	-0.23	-0.34	-0.10

Reliability for domains across conditions

Variable	Original domains condition (n = 240)	Easy domains condition (n = 225)	Difficult domains condition (n = 226)
Domain difficulty	.64 (.68, .48)	.76 (.76, .48)	.68 (.47, .61)
Comparative ability	.76 (.61, .72)	.77 (.67, .72)	.86 (.71, .83)
Own absolute ability	.71 (.51, .72)	.68 (.48, .69)	.85 (.65, .82)
Others' absolute ability	.74 (.59, .77)	.77 (.60, .76)	.90 (.78, .85)
Desirability	.69 (.56, .68)	.74 (.70, .65)	.77 (.67, .64)
Ambiguity	.70 (.62, .50)	.73 (.64, .60)	.81 (.75, .59)

Note: Reliabilities are Cronbach's α. Reporting structure is the following: full inventory (easy items, difficult items). Reliability met requirements (α ≥ .7, see Tavakol & Dennick, 2011).

Extension conditions: correlations between comparative ability and domain difficulty ratings for each ability domain

Ability domain	Easy domain condition				Difficult domain condition			
	p	*r*	95% CI		*p*	*r*	95% CI	
			Lower	Upper			Lower	Upper
Using a computer mouse	.647	0.03	-0.10	0.16	<.001	-0.2	-0.37	-0.12
Driving	.082	-0.12	-0.24	0.01	<.001	-0.27	-0.39	-0.15
Riding a bicycle	.071	-0.12	-0.25	0.01	<.001	-0.24	-0.36	-0.12
Saving money	.144	-0.10	-0.23	0.03	<.001	-0.36	-0.47	-0.24
Telling jokes	.935	0.01	-0.13	0.14	<.001	-0.23	-0.35	-0.11
Playing chess	<.001	-0.36	-0.47	-0.24	<.001	-0.27	-0.39	-0.15
Juggling	<.001	-0.26	-0.38	-0.13	<.001	-0.25	-0.37	-0.12
Computer programming	<.002	-0.22	-0.34	-0.09	.002	-0.20	-0.33	-0.08

References

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge. doi: https://doi.org/10.4324/9780203771587

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. Download PDF

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221–232. https://doi.org/10.1037/0022-3514.77.2.221

Lakens, D., Caldwell, A. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, **4**(1), 251524592095150. doi: 10.1177/2515245920951503

Lenhard, W., & Lenhard, A. (2016). *Calculation of Effect Sizes*. Retrieved from: https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): Psychometrica. DOI: 10.13140/RG.2.2.17823.92329

The Pennsylvania State University. (2020). *6.3 - Testing for Partial Correlation: STAT 505.* Retrieved from https://online.stat.psu.edu/stat505/lesson/6/6.3