What correlations mean for individual people

Citation for published version:
Mottus, R 2022, 'What correlations mean for individual people: A tutorial for researchers, students and the public', *Personality Science*, vol. 3, e7467, pp. 1-27. https://doi.org/10.5964/ps.7467

Digital Object Identifier (DOI):
10.5964/ps.7467

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Personality Science

PERSONALITY Science

PsychOpen GOLD

Tutorial

Check for updates

What Correlations Mean for Individual People: A Tutorial for Researchers, Students and the Public

René Mõttus [1,2]

[1] *Department of Psychology, University of Edinburgh, Edinburgh, Scotland.* **[2]** *Institute of Psychology, University of Tartu, Tartu, Estonia.*

Personality Science, 2022, Vol. 3, Article e7467, https://doi.org/10.5964/ps.7467

Received: 2021-09-06 • **Accepted:** 2022-05-18 • **Published (VoR):** 2022-12-16

Handling Editor: John F. Rauthmann, Universität Bielefeld, Bielefeld, Germany

Reviewing: *Round 1* - Daniel Leising; Julia M. Rohrer; Anonymous #1; *Round 2* - Daniel Leising; Julia M. Rohrer; Anonymous #1. Open reviews are available. [see Index of Supplementary Materials]

Corresponding Author: René Mõttus, 7 George Square, EH8 9JZ, Edinburgh, Scotland. E-mail: rene.mottus@ed.ac.uk

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Trisecting and cross-tabulating (TACT) two related variables shows what their correlation means for individual people. For example, knowing an individual's conscientiousness (lowest, medium or highest third among other people) improves the accuracy of predicting their health by 1.4, their child's conscientiousness by 4.2, and their job performance by 7.2 percentage points, compared to the random-guess accuracy of 33.3%. There's a 35% probability that they will score differently in a few years and a 50% probability that their partner would rate their conscientiousness differently. For typical correlations in psychology, about 40% of individuals with a low or high value in one variable have a similar value in the other variable, while medium values carry almost no predictive information. Hence, correlations' intuitive interpretations like "someone high in x is likely to be high in y" are almost always incorrect. An R package is provided for calculating and visualising TACT.

Keywords

correlation, prediction, individual, population, information, accuracy

Relevance Statement

Much of psychological research reports population trends, often expressed as correlations. A simple tool for researchers, students, and the public, TACT, shows how to (not) use correlations to say something about individuals.

Key Insights

- Correlations are often used to make statements about individuals.
- TACT helps to intuitively assess the accuracy of these statements.
- The accuracy can be expressed as a percentage, compared to a random guess.
- Most correlations don’t allow for meaningful statements about individuals.
- Phrases like “someone high in x is likely to be high in y” are usually incorrect.

Many psychological research findings represent statistical trends in the population, showing how two variables tend to vary together among people. The strengths of these trends are often expressed using correlation coefficients, the absolute value of which can range from 0 (no relation at all) to 1 (one variable is perfectly predictable from the other). Insofar as we assume psychology to be about individuals rather than populations, we expect these trends to tell us something meaningful about individual people. For example, from a correlation between income and happiness, we may conclude that an individual with high earnings (say, Kati) is probably happy, while someone with an average income (say, Mati) probably has about average happiness. This is how research findings are often interpreted in the (social) media and everyday conversations by researchers and the public alike. In clinical assessment, educational, or hiring settings, correlations can inform real-life decisions about individuals. Even psychologists themselves often admit to choosing their field—which advances by documenting correlations—to understand themselves better.

But how much can we trust statistical trends in the population to tell us something meaningful about actual individuals like Kati and Mati? Most people know that applying population trends to individuals entails uncertainty, so any conclusion is only correct to some degree. Here, I describe an intuitive way to think about and communicate this degree, based on grouping individuals into "low", "medium", and "high" groups in both variables and calculating the probabilities that similar values match. In contrast to clinical diagnostics with its tools for expressing binary outcomes’ probabilities, such simple trisecting and cross-tabulating (TACT) is particularly useful for variables that vary on arbitrarily defined continuous scales, as is very common in psychology. For example, given a .25 correlation between the personality trait of conscientiousness and supervisor-rated job performance, how likely is it that a highly conscientious individual performs highly at their job while a person with a medium conscientiousness is a medium-performer?

TACT makes research findings interpretable without specialist knowledge (e.g., what correlations are typical in the field?) or abstract statistical concepts (e.g., standard deviation). Characterising people as having low, medium, or high values makes continuous variables easily interpretable and aligns with how many people intuitively think of them. For example, like some of my colleagues, I am neither low nor high in talkativeness but somewhere in the middle, whereas some of our colleagues are distinctly more and others less talkative than us, the medium-talkativeness people. The distinction between medium and more extreme values is also important because the former are often less informative, as I demonstrate below.

I show how a range of commonly observed correlations can be interpreted using TACT, hoping that this helps researchers, students, and laypeople better grasp the implications of these and many other common research findings across psychology. Experimenting with TACT has reshaped my own intuition about the meaning of correlation coefficients and made me more careful about drawing conclusions from common research findings. I only wish this experimenting had happened as a part of my training, and I hope that this tutorial is useful to others, regardless of their career stage, as well as the public.[1]

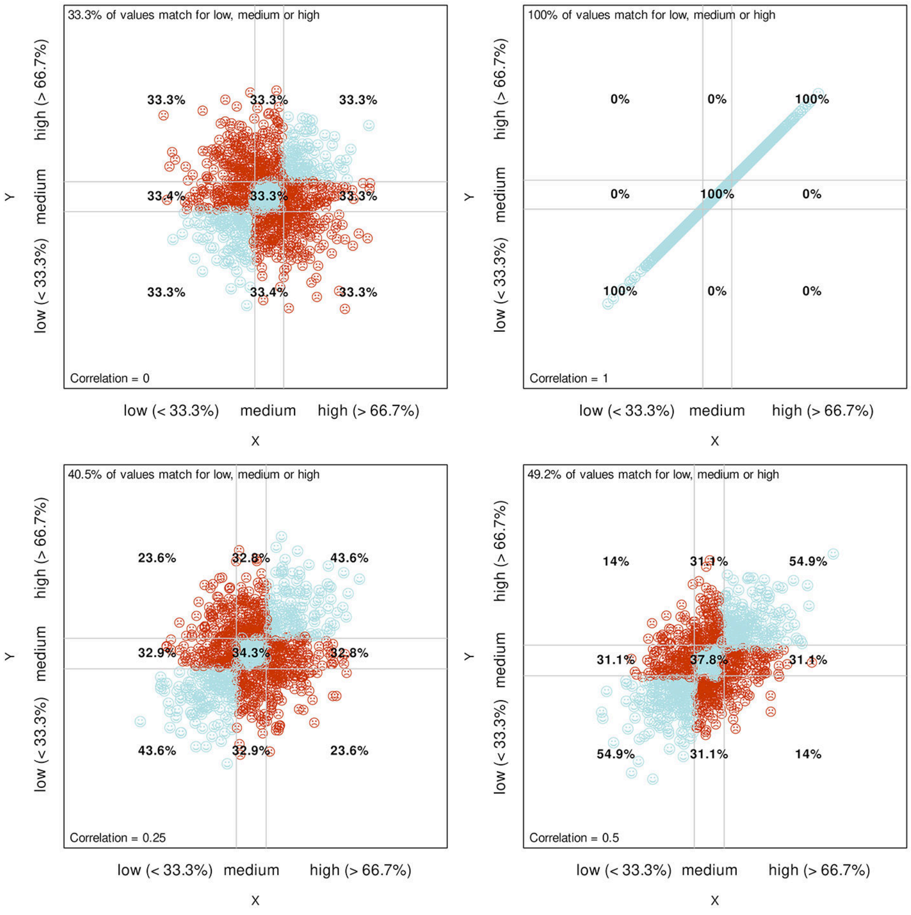
Trisect and Cross-Tabulate (TACT) Variables

TACT means little more than paying attention to scatterplots that every psychologist is already trained to look at and that are familiar to most people with at least a high school education. A scatterplot is a cloud of points showing the relationship between two variables, with each point representing one individual's position in both variables (Figure 1). For TACT, simply lay a 3-by-3 grid over the scatterplot so that the lines trisect the distributions of both variables. For the variable on the horizontal (x) axis, a third of individuals fall in the centre of the grid, representing medium levels of the variable, and the two remaining thirds fall on each side, representing low and high values. For the variable on the vertical (y) axis, likewise, a third of individuals fall in the middle of the grid, representing medium levels of that variable, whereas the two remaining thirds fall on the top and bottom of the grid, respectively, representing the low and high values.

[1] I note that others have also thought about the meaning of statistical trends for individuals, such as in the context of clinical interventions (e.g., Jacobson & Truax, 1991), but not using a simple and general-purpose tool such as TACT, to my knowledge. A similar tool is known as Binomial Effect Size Display and will be discussed later in the article.

Figure 1

TACTs of Correlations 0, 1, .25, and .50



With no relationship between the two variables, all nine grid slots contain the same number of individuals, at least with a sufficiently large sample (Figure 1, top-left). In contrast, for a perfect positive correlation, all individuals fall into the three grid slots on the diagonal from the bottom-left to the top-right so that all low, medium and high values on one variable match the corresponding values of the other variable (Figure 1, top-right). (All values would lie on the other diagonal of the grid with a perfect negative correlation.) In most cases, however, the absolute values of the correlation are somewhere between 0 and 1.0, resulting in more varied patterns in where the individuals fall (Figure 1, bottom). In psychology, for example, correlations much above .30 are rare, so most TACT plots look like the first column of Figure 1 (Funder & Ozer, 2019).

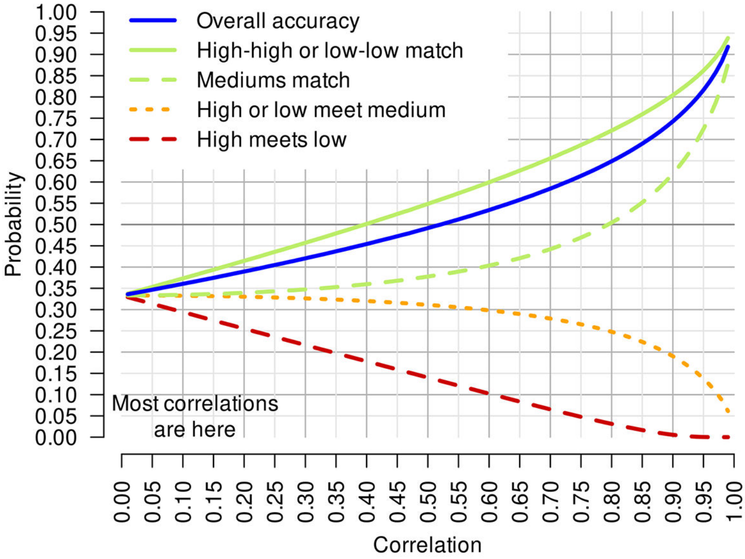
To show how many individuals match or do not match in their levels of the two variables, we can calculate their proportions so that they add up to 100% in each grid column. This means estimating how many individuals with low x values are expected to have low, medium, or high values on y, and the same for the medium and high values of x. For example, we can then say: "*If Mati scores high in conscientiousness, he has* p *probability of performing highly at work*". This *p* probability has to be compared to the random-guess value of 33.3%, expected when there is no correlation between the two variables (Figure 1, top-left).

Almost all psychological variables have many causes and subcomponents, so they have bell-shaped distributions in the population: most people are around the average, whereas more extreme values are increasingly less common. The middle grid slots are narrower both vertically and horizontally for such variables, as shown in Figure 1. In contrast, they would be equally wide for uniformly distributed variables, whereas the narrower slots would be on the sides, at the top or at the bottom for variables with skewed distributions.

In the following two sections, I apply TACT to several well-established correlations and provide rules of thumb for interpreting typical correlations. TACT probabilities for all other correlations can be seen in Figure 2 or in the Supplementary Materials.

Figure 2

TACT Probabilities for Correlations .01 to .99



Note. Overall accuracy means that similar values are matched regardless of whether they are low, medium or high.

In my calculations, I assume normal (bell-like) distributions for both variables (x and y), simulate a range of correlations between them among 10^8 individuals, apply the 3-by-3 grid on the resulting scatterplots, and calculate the three proportions of y values for each of the three levels of x. For this, I apply a companion R package TACT (see Supplementary Materials) that allows "TACTing" variables with any correlation (and varying other parameters that will be discussed later). For example, to TACT a .25 correlation between two normally distributed variables I use:

```
> TACT(r = .25, distribution = "normal").
```

The package can also TACT empirical correlations in raw data by being supplied with two variables rather than a correlation coefficient.

Here, I only consider positive correlations, although the logic is identical for negative ones if we swap the low and high labels for either variable.

In a subsequent section, I discuss the implications of and alternatives to my choices (e.g., different cut-offs for low and high values and alternative variables' distributions).

TACT in Action

Does my Low Conscientiousness Mean That I Am Unhealthy?

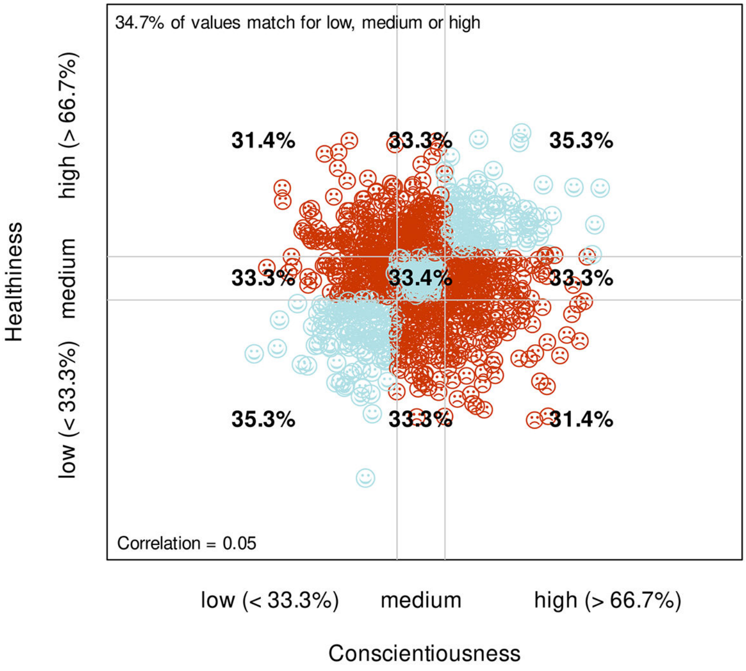
Given how conscientiousness correlates with multiple physical health markers such as body mass (Vainik et al., 2019), inflammation markers (Sutin et al., 2010), diabetes (Jokela et al., 2014) or longevity (Graham et al., 2017), we can hypothesise that its correlation with a hypothetical underlying trait of physical healthiness is about .05. This estimate is consistent with a recent meta-synthesis of 18 meta-analyses of conscientiousness-health associations (Strickhouser et al., 2017). If a journalist, for example, asked us to explain this .05 correlation in intuitive terms, we can try the following.

People differ widely in their general healthiness, but for simplicity, we can categorise them into three equally sized health status groups: low, medium, and high. If we know nothing about a person, we can only guess that their likelihood of being in any of the three groups is equally 33.3%. This is like throwing a fair three-sided die to guess how healthy that person is. But knowing whether that person belongs to the bottom, medium, or top third of the population in conscientiousness increases our accuracy in predicting their health status to 34.7%—it's like the die is now slightly loaded. On a closer look, however, how much we can learn about the person's health depends on their particular level of conscientiousness. If that person has high conscientiousness, they are also likely to have high health status with a 35.3% probability, and the same probability goes for low conscientiousness and low health (Figure 3). But with medium conscientiousness, they only have a 33.4% probability of having medium health status, which is very similar to the random-guess accuracy.

Suppose a layperson completes a personality questionnaire and receives feedback that they have a medium conscientiousness level. This information cannot make them much wiser regarding their health, being almost equally likely to have any health status. But having received feedback that their conscientiousness is low, they are about two percentage points likelier to have a low rather than medium health status and about four percentage points likelier to have a low rather than high health status. This way of presenting the information can increase the chances that individuals meaningfully apprehend the implications of *their* trait level for *their* physical health—or rather the lack of such implications. For example, being told about the link between conscientiousness and health without a proper explanation of its strength, someone may justifiably worry that their low conscientiousness means their health is worse than they thought. Explained properly, however, they will probably realise that this population trend is too small to mean much to any one individual. Although low conscientiousness could mean poor health, it still goes with either medium or even good health in most cases.

Figure 3

TACT of the .05 Correlation Between Conscientiousness and Healthiness



Put differently, although the best guess is that someone's health status is similar to their conscientiousness, it still remains incorrect nearly two times out of three, which is quite similar to guessing randomly. One may ask, then, what do researchers mean when they say that conscientiousness is an important predictor of health (e.g., Bogg & Roberts, 2004)? The answer is that instead of individuals, they typically think of large groups of people. Even correlations that have little relevance for individuals can make a difference at the level of very large groups (Funder & Ozer, 2019). We can also use TACT to convey this idea.

Hypothetically, suppose that the conscientiousness-health link is directly causal, and public health officials deploy a cheap yet at least modestly efficient intervention to a million people with currently low conscientiousness (or three million randomly selected people of whom a million have a low level of the trait), increasing the trait among 10% of them to a level that is currently considered medium. Among these people, 35.3% currently have poor health; after the intervention, this could decrease to 33.3%. This means that nearly 2,000 currently low-health people could end up with health that is presently considered medium. Moreover, almost 2,000 people will move from what is currently a medium health status to a level that is presently regarded as high. Although the lows, mediums, and highs will need to be recalculated after the intervention and, in comparison to others, most low-health people will remain low-health people, the absolute increase in their underlying health status trait may be substantial enough to mean that fewer people get sick and die. Although the intervention does not help a vast majority of people, it will help some, and public health officials can calculate whether the reduction in treatment costs and increased productivity are sufficient to cover the intervention's costs.

Like Father, Like Son?

Both the public and scientists know that parents and children tend to be similar in their traits, primarily due to their partly shared genes. Empirically, parent-child correlations in personality traits tend to be around .15 (Loehlin et al., 2005). Given this correlation, how accurate is a, say, modestly conscientious father in predicting that their child's trait level will match theirs?

Given the .15 parent-child correlation, the overall probability that a parent in the bottom, medium or top third of the population in a personality trait has a child in the same third is 37.5%; for comparison, any two randomly compared people have a 33.3% probability of being in the same third. More specifically, a parent with a high or low value on the trait can expect their child to match their trait level with 39.4% probability, whereas the child has a medium or even the opposite trait level with 33.2% and 27.5% probabilities, respectively. But a child of a parent with a medium trait level is almost equally likely to have a low (33.2%), medium (33.7%), or high (33.2%) trait level (Table 1).

Table 1

TACT of the Correlation .15 Between Parents' and Children's Personality Traits

	Child's trait (%)		
Parent's trait	***Low***	***Medium***	***High***
High	27.45	33.16	**39.38**
Medium	33.16	**33.69**	33.16
Low	**39.38**	33.16	27.45

This means that over three in five sons (or daughters) are *not* like their father (or mother) in any given personality trait, when we think of that trait in terms of low, medium, and high values. So, for example, worrying that a potential partner's low level of a desired trait will show up in their child would not be a compelling reason for rejecting their marriage proposal. Likewise, guessing someone's trait level from their mother's or father's is only marginally more accurate than a random guess.

Will a Highly Conscientious Applicant Be a High-Performing Employee?

The personality trait of conscientiousness is known to be among the predictors of various indicators of job performance and is often used for selecting suitable job applicants. Across many studies, the conscientiousness-job performance correlation is around .25 (e.g., Judge et al., 2013). How could a psychologist explain this correlation to an HR manager, hoping to convince them to start testing job applicants?

Knowing whether a job applicant is in the bottom, middle, or top third of conscientiousness among other applicants, we can predict their job performance level—low, medium, or high—with 40.5% accuracy; for comparison, by just throwing a die, we could achieve 33.3% accuracy. More specifically, it is a 43.6% probability that a person in the top third in conscientiousness will also be in the top third in job performance, against a 32.9% probability of being a medium performer and a 23.6% probability of having a low-third performance level (Table 2).

Table 2

TACT of the Correlation .25 Between Conscientiousness and Job Performance

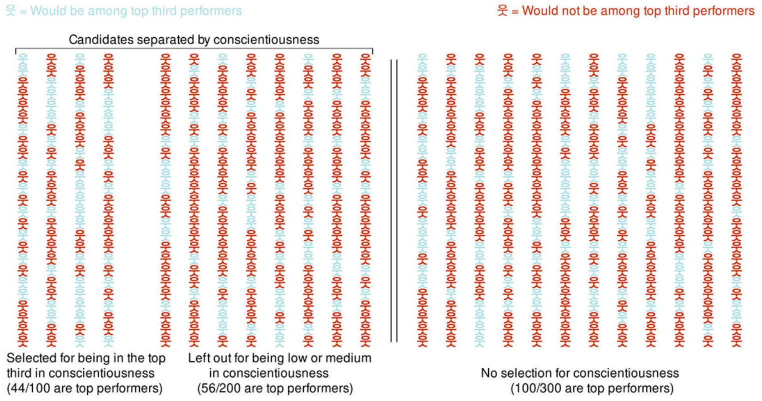
	Conscientiousness (%)		
Job Performance	***Low***	***Medium***	***High***
High	23.57	32.85	**43.58**
Medium	32.85	**34.30**	32.85
Low	**43.58**	32.85	23.57

In other words, by picking an applicant from the highest third in conscientiousness rather than randomly, we decrease the chances of *not* getting a high-performer from 66.6% to 56.4%, and the chances of getting a low-performer decrease from 33.3% to 23.6%. But the chances of ending up with the medium-performer remain virtually unchanged regardless of whether we pick the applicant randomly or use their conscientiousness score.

This information could be helpful for employers who want to quickly filter out a majority of applicants from a larger applicant pool while over-saturating the remaining pool with high-performers (43.6%) and ensuring that among those removed from that pool, the proportion of high-performers is lower (28.2%, the average of 32.9% and 23.6%). However, these employers have to be careful with providing feedback to the applicants. For example, a statement like "*your personality trait scores suggests that you are (un)likely to be among the best candidates*" would be misleading because most candidates are not high-performers at any level of conscientiousness (Figure 4). Also, employers averse to losing outstanding applicants should heed filtering out more than half (56.4%) of the high-performers due to their medium or low conscientiousness. (Later in the article, I discuss defining high performance differently than being the top third, which may be more useful in specific hiring situations.)

Figure 4

Proportions of High-Performing Job Applicants Either When Selected or not Selected for High Conscientiousness



Note. Assuming a correlation of .25 between conscientiousness and job performance, the proportions of high-performers among 300 hypothetical job applicants, if they were selected for being in the highest third (left) of conscientiousness or filtered out because of not being in the highest conscientiousness third (centre), or not selected based on conscientiousness at all (right). In any case, high-performers remain a minority among applicants, including among those selected for high conscientiousness.

Would my Friend Agree With me on my Personality Traits?

When people (targets) complete well-established personality questionnaires about themselves and have others who know them well (informants) also complete that test about them, the scores often correlate around .50 (McCrae et al., 2004; Mõttus et al., 2014). This is an unusually high correlation by psychological research standards and even beyond; for example, parents and their children's heights correlate about .50 (Luo et al., 1998).

If people were randomly filling out the questionnaire, 33.3% of the targets scoring low, medium or high in a self-reported personality trait would score similarly in the informant-reports. However, assuming the .50 correlation, 49.2% of targets are expected to have a similar trait level in both self- and informant-reports. More specifically, of those scoring high or low in self-reports, 54.9% would score similarly in informant-reports, whereas 31.1% would have a medium and 14% an opposite score (Table 3). Among medium-scorers in self-reports, only 37.8% would have the same trait level in informant-reports.

Table 3

TACT of the Correlation .50 Between Self- and Informant-Reports of a Personality Trait

	Informant-reports (%)		
Self-reports	***Low***	***Medium***	***High***
High	14.02	31.12	**54.87**
Medium	31.12	**37.77**	31.12
Low	**54.87**	31.12	14.02

Suppose we conducted a study where people received feedback on their self-reported personality traits (low, medium, or high compared to other self-reports) and that each person was also rated by an informant who independently received similar feedback about the target (low, medium, or high compared to other informant-reports). If the participants and their informants could compare their feedback, about half of the target-informant pairs would find that they similarly described the target in any given trait.

Will I Get a Second Opinion on my Personality Traits?

Suppose a person completes a personality trait questionnaire and receives feedback that they are in the medium third in neuroticism compared to other people; they are unsure about this feedback and want to complete another questionnaire to get a “second opinion”. How likely would they get a different result?

The correlations among different neuroticism scales vary but average somewhere around .70 (Pace & Brannick, 2010; Soto & John, 2017; Thielmann & Hilbig, 2019; ipip.ori.org). Given this correlation, a person is likely to get similar feedback—being in the lowest, medium or highest third in the trait—in two different questionnaires with a 58.4% probability. More specifically, a high or low scorer would receive similar feedback in 65.6% of cases, whereas that probability is 44.2% for a medium scorer (Table 4). So, a person uncomfortable with their medium feedback is indeed more likely to get a different result from another test than receive a medium score once again.

Table 4

TACT of the Correlation .70 Between the Scores of two Questionnaires Measuring the Same Personality Trait

	Score in the second test (%)		
Score in the first test	***Low***	***Medium***	***High***
High	6.55	27.90	**65.56**
Medium	27.89	**44.21**	27.90
Low	**65.56**	27.89	6.55

How Much Can I Trust my Test Score?

TACT offers a way to think about and explain what measurement (im)precision means for the measurements of individual people. As shown, about five in ten people will be in the same third in a personality trait according to two different sources of information—their own ratings and ratings by someone who knows them well—and nearly six in ten will be in the same third when completing two different tests of the same trait. To complement that, over seven out of ten people will be in the same third when completing the same test twice.

Specifically, scores of well-established psychometric tests taken twice over two weeks typically correlate close to .90 (e.g., Henry et al., 2022). Using TACT, when people complete such a test and receive feedback, scoring either low, medium or high relative to others, and then do it again in two weeks, 74.3% of them will get the same feedback. In particular, among those scoring either high or low on the first occasion, 80.4% are expected to receive similar feedback again, whereas 19% score medium on the second occasion, and only 0.6% flip from one extreme to the other (Table 5). But among medium-scorers on the first occasion, only 62.0% should expect similar feedback again.

Table 5

TACT of the Correlation .90 Between the Scores of the Same Test Taken Two Weeks Apart

	Score in two weeks (%)		
Score in the first testing	***Low***	***Medium***	***High***
High	0.56	19.02	**80.42**
Medium	19.00	**61.97**	19.02
Low	**80.42**	19.00	0.56

For psychologists, correlations between the scores from two testing occasions show how reliable the scores are, often abstractly defined as the proportion of "true score" variance in them (McCrae & Mõttus, 2019). TACT can help to make this concept more meaningful for individual test-takers. With a comparatively good test, for example, there is nearly three out of four chance that a person would get a similar score if they did that test once again, although they can trust high and low scores more and medium scores less. Yet getting a different result the second time is not uncommon either, happening to about every fourth test taker; when this happens, people should not worry about having a split personality.

Will I Still Be Average in Neuroticism in a few Years?

When psychologists discuss personality with a lay audience, one question that often comes up is about personality stability: *How set are people in their ways?* A common but vague answer may be: *Personality traits are not fixed, but generally people tend to maintain their trait levels relative to their peers.* Indeed, some of the strongest statistical trends in psychology are seen when the same traits are measured multiple times, even over many years. For example, among adults, two measurements of a personality trait, separated by a few years, correlate at around .70 although studies with presumably better assessment methods have reported estimates close to .80 (Briley & Tucker-Drob, 2014; Mõttus et al., 2019; Terracciano, Costa, & McCrae, 2006). But how can we explain this in a way that people could meaningfully relate to?

A correlation of .80, for example, would mean that an individual with a low, medium or high level of the trait on the first testing occasion can expect the same result in a few years with a 64.9% probability. At a closer look, someone with a high or low trait score has a 72.1% probability of scoring similarly, 24.8% probability of having a medium score and only 3.1% probability of having an opposite trait score after a few years (Table 6). But of those with a medium score, only about half (50.4%) will get the same score again.

Table 6

TACT of the Correlation .80 Between Trait Scores Measured Six Years Apart

	Score a few years later (%)		
Score in the first testing	***Low***	***Medium***	***High***
High	3.13	24.78	**72.09**
Medium	24.79	**50.44**	24.78
Low	**72.09**	24.79	3.13

So, about two-thirds of people will retain their relative trait level in the population within a few years, while every third person changes, most often by moving to the adjacent trait level. Even with such strong population-level stability, then, personality trait change is still very common among individuals, especially among those with medium trait levels.

Why Are Medium Values Less Informative?

Variables' medium values carry less predictive information about other variables than more extreme values—at least when the baseline probabilities of low, medium and high values are equal (see below for different examples)—because they have two equally probable adjacent levels on the other variable that they can deviate to. However, more extreme values have only one adjacent value to deviate to, whereas deviating to the other extreme is less likely.

Rules of Thumb for Interpreting Typical Correlations

In psychological research, correlations between .15 to .25 represent the most common statistical associations between variables (Gignac & Szodorai, 2016; Richard et al., 2003). When we think of variables as having low, medium and high values, such correlations mean that

- around 40% of those with a low or high value on one variable are likely to have a similar value on the other variable, whereas 60% are likely to have different values on the two variables;
- compared against the random-guess baselines of 33% and 67%, respectively, typical correlations mean a seven percentage points improvement in the accuracy of predicting low or high values in one variable from the other variable;
- medium values on one variable carry virtually no information about the other variable.

Innocent-looking statements like "*a person high in x is likely to be high in y*" should generally be avoided because

- it takes a correlation over .40 for it to be even marginally correct, being valid for at least over half of the people;
- it takes a correlation over .80 for it to be accurate to a compelling extent, applying to at least two in three people.

Such statements would typically be misleading at best or incorrect at worst because correlations that high are rare (Funder & Ozer, 2019) unless we consider associations among multiple measurements of essentially the same phenomenon.

So, we should avoid categorical conclusions about any real person based on typical statistical trends among people. Instead, one solution is to confine our conclusions to vague population-level statements such as "*in a large group of people, higher levels of x tend to be associated with higher levels of y*". Alternatively, conclusions can be articulated using numeric probabilities about individuals like those provided by TACT and illustrated above. This should make most correlations' limited implications for individuals objectively obvious.

News articles often present research findings as having implications for the reader: unfortunately, this almost always means misleading the reader because the vast majority of correlations are not even nearly sufficiently strong for that.

Using TACT in More Ways

Primarily, I present TACT as a simple and general-purpose tool for understanding and communicating correlations' magnitudes. Admittedly, it involves arbitrary assumptions and simplifications, some of which I address in this section. Yet it can be adapted to more specialised needs to meet different assumptions, also described here. I also discuss some of TACT's overlaps with the rich but sophisticated toolbox of clinical diagnostics. Those readers interested in TACT as a simple and general-purpose tool may find this section too technical and wish to skip it in full or partly.

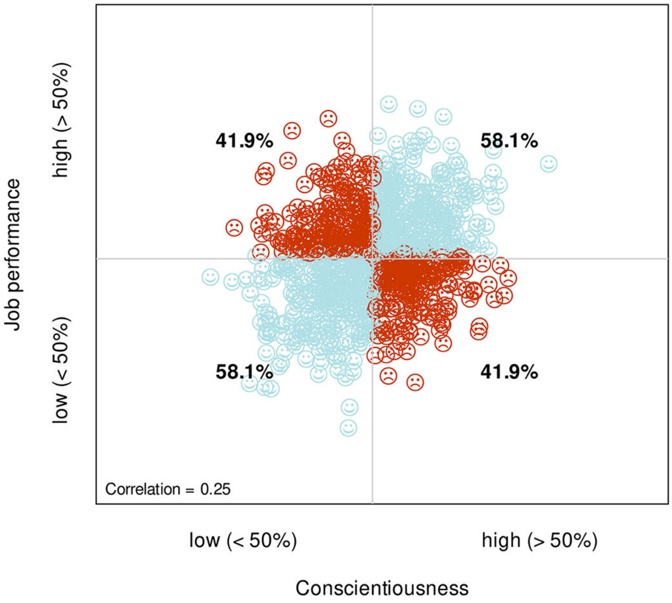
Why Medium, Low and High?

For example, why not use two categories instead of three, thus *bi*secting and cross-tabulating continuous variables? In fact, this is sometimes done, known as the Binomial Effect Size Display (BESD; Rosenthal & Rubin, 1982). For continuous variables, BESD involves splitting variables at the mean or median and then cross-tabulating them. The TACT R function also allows doing this by setting the cut-offs of both variables equal at the 50% percentile, so the correlation of .25 between job success and conscientiousness could be calculated and presented as follows, producing Figure 5:

```
> TACT(r = .25, distribution = "normal", cutoffsx = c(.5,.5), cutoffsy =
c(.5,.5))
```

Figure 5

Illustration of a Correlation Using Binomial Effect Size Display



Given the .25 correlation, those above the median in conscientiousness have a 58.1% probability of also being above the median in job performance, against the 50% random-guess rate.[2] Using BESD, thus, this correlation entails an eight percentage points increase in our accuracy in identifying a comparatively higher-performing individual. For comparison, TACT provides a seven percentage points increase in accuracy with a .25 correlation (40.5% vs 33.3%). TACT's accuracy is somewhat lower because it makes a more precise prediction about individuals' positions on the variables, which makes the predictions slightly riskier and automatically less likely accurate.

Labelling people as low *versus* high on a variable may seem simpler than also considering medium levels, but it has some important limitations when applied to continuous variables. First, even heuristically, people do not fall into just high and low groups for the multiply determined variables (constructs) that interest psychological researchers. At the very least, many score moderately in them because they are not consistently one way or

[2] For continuous variables, BESD is sometimes incorrectly calculated, even in popular textbooks (e.g., Funder, 2019). This is because these calculations do not take into account the reduction in correlation due to dichotomising the variables (Hunter & Schmidt, 1990). The TACT R function provides a more accurate BESD because it actually cross-tabulates the observations above and below mean values.

another, be it across occasions or different aspects of the constructs. With many variables obeying bell-shaped normal distributions, most people fall into a narrow range of values around the middle points of their distributions, thus being close to the border between what would appear high and low values in BESD. So, highs are often more similar to lows than to many of their fellow highs, and vice versa.

Second, with continuous variables BESD masks the omnipresent phenomenon of regression to the mean: high (or low) values in one measurement are statistically expected to match relatively lower (or higher) values in another measurement, even when the measurements are correlated. Within high or low groups, people at the more extreme end in one variable are more likely to be closer to the middle of the distribution of the other variable. Moreover, the more extreme value they have on one variable, the more they are expected to regress towards the mean on the other variable. For example, a highly conscientious person is likely to regress towards the average on any other measure correlated with conscientiousness, even if they still remain (just) above the average slightly more than half of the time. Because the high and low groups are so broad when using BESD for continuously distributed variables, this trend remains masked (e.g., someone can be on the 5th percentile on one variable but on the 49th percentile on the other, yet counting as similarly high on both). In TACT, the groups are narrower, so the regression to the mean is comparatively less likely to go unnoticed—people can actually regress from high or low to the medium group.

Third, BESD masks medium values' tendency to be less informative about other variables than more extreme values. In fact, with typical correlations, medium values carry virtually no predictive information. Admittedly, I had never thought about it before experimenting with TACT. As the TACT examples showed, this can often have important implications for interpreting research findings at the level of single individuals (e.g., for feedback).

So, I consider TACT an improvement over BESD, because it is more consistent with how people think of continuous variables, better addresses regression to the mean and shows the different predictive values of the medium and more extreme scores. Of course, we could also categorise people into four or five groups, but this would inevitably make the interpretation of correlations more complex since there would be 16 or 25 slots on the scatterplot.

What if More People Are Medium Than High or Low?

Choosing any cut-offs between low, medium and high values is arbitrary, but making the three groups equal in size is arguably the least arbitrary and most generally applicable and intuitive solution. It means that the random guess is equally accurate at all levels of the variables involved—that is, the die is not *a priori* loaded in any way.

However, in specific circumstances, other cut-offs may be more practical such as, for example, half of the individuals or those within two standard deviations around the mean categorised as medium. The effects of any such cut-offs can be tested using the TACT R function by setting the cut-offs to, say, 25th vs 75th percentiles (half of the people now being medium) or 16th vs 84th percentiles (two standard deviations around the mean now being medium, for normally distributed variables). As a general rule, the overall probability that low, medium and high values on both variables match remains similar for all cut-offs, but larger medium groups mean higher probabilities that variables' medium values match at the expense that the probabilities of matching low (or high) values match. Of course, when more people are set to have medium values, it is *a priori* likelier that any given individual has a medium value as well—that is, the die becomes loaded in favour of medium values even before we introduce any other information about the individuals.

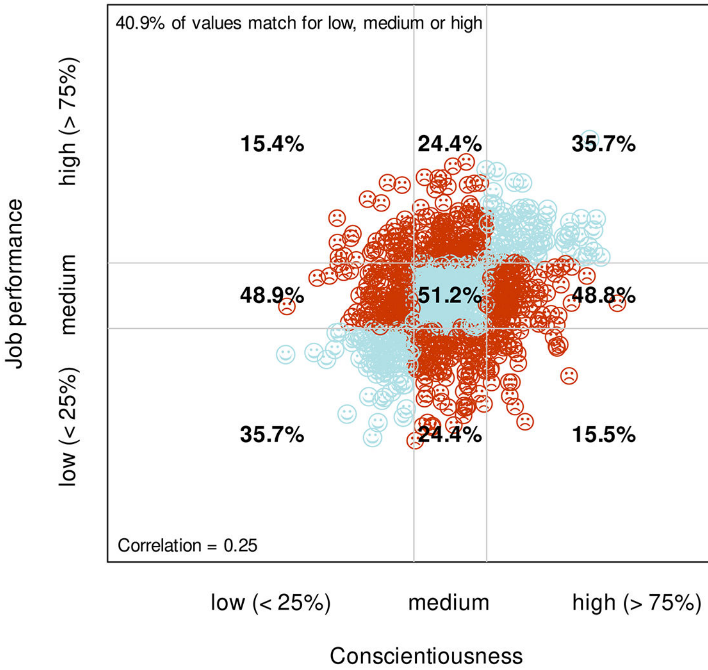
In other words, the more precise a conclusion we want to draw about someone's level in one variable based on their score in another variable, the more likely we are to be incorrect because riskier predictions are always less probable, all else equal. (I already showed this in relation to bisecting variables.) Suppose we want to identify the top quarter of job performers by only keeping the applicants in the top quarter in conscientiousness instead of the top third. We will find that the top-quarter cut-off would entail a lower accuracy in identifying a top-performer (35.7% vs the 25% random-guess baseline) than the top-third cut-off (43.6% vs the 33.3% baseline), and the same is true for even higher cut-offs. That is, the increase in accuracy over the random-guess baseline remains comparable, but the absolute accuracy decreases because the random-guess baseline decreases.

To test the effect of assigning half of people to the medium category, this code can be used, producing Figure 6:

```
> TACT(r = .25, distribution = "normal", cutoffsx = c(.25,.75),
cutoffsy = c(.25,.75))
```

Figure 6

TACT of an Association With Half of the People Having Medium Values, Instead of a Third Having Each of the Three Values



Cut-Offs for Diagnostics

Bisecting variables and varying the high-low cut-offs connects TACT with the literature on clinical testing and diagnostics (e.g., Lalkhen & McCluskey, 2008) where the goal is to identify those having (a risk for) a certain outcome as accurately as possible while not "flagging" those who don't have it. Suppose the binary outcome we care about here is being in the top quarter in job performance, thus bisecting this variable at a fixed cut-off of 75th percentile. Then, a conscientiousness test's *sensitivity* (also called recall or hit rate, among other terms) shows how many of those in the top quarter in job performance we can identify by bisecting its scores at a certain cut-off, while its *specificity* (or selectivity) shows how many of the lower-performing candidates can this cut-off filter out. In contrast, the test's *positive predictive value* (PPV) shows how many of those identified as potential top-performing candidates by bisecting the conscientiousness scores are actually top-performers, whereas its *negative predictive value* (NPV) shows how many of those filtered out by the conscientiousness test are actually lower-performers. I have presented TACT along the lines of PPV and NPV, assuming that the TACT scatterplot grid's columns sum to 100%, rather than in terms of sensitivity and specificity, which assume that the grid's rows sum to 100%. Ideally, all of these four indicators should be as high as possible, but unfortunately, they are in tension. For example, by lowering the conscientiousness cut-off and selecting more people as potential top candidates, we miss fewer good candidates (higher sensitivity) but pass more lower-performing candidates (lower specificity and PPV). To find a balance between them that works for a particular situation, we can try different conscientiousness cut-offs, thereby decoupling them from the fixed cut-off for job performance.

With a .25 correlation, an accurate prediction about an individual—that they are in the top quarter in performance because of their high conscientiousness—will always remain less likely than a wrong prediction, but it does increase somewhat with higher conscientiousness cut-offs. For example, with a 50% conscientiousness cut-off, the probability of being accurate is 31.4% (against the 25% random guess baseline), which increases to 35.7%, 40.4% and 43.5% with 75%, 90% and 95% cut-offs, respectively. However, the probability of correctly identifying those not in the top quarter of job performance, or NPV, *decreases* with increasing PPV. For these four conscientiousness cut-offs, the respective NPVs decline from 81.4% to 78.6%, 76.7% and 76.0%, against the 75% random guess baseline. Sensitivity and specificity also move in opposite directions for these cut-offs, from 62.8% to 35.7%, 16.2% and 8.7%, and from 54.3% to 78.6%, 92.1% and 96.2%, respectively. So, there is an inherent trade-off: tightening the selection criteria increases the accuracy in keeping high-performers in the applicant pool, but it also increases the chances of incorrectly eliminating high-performing candidates. When the correlations between the predictor and the outcome are considerably stronger than .25, it is easier to find a useful cut-off because then sensitivity decreases slower with higher cut-offs than specificity increases.[3] Unfortunately, these stronger correlations are quite rare in psychology.

Such scenarios, already discussed by Taylor and Russell (1939), can be experimented with the TACT function, for example:

```
> TACT(.25, "normal", cutoffsx = c(.90,.90), cutoffsy = c(.75,.75))
```

These scenarios can provide helpful solutions for specialised needs, but they also show the complexities of using correlations for diagnostic decisions about real individuals. Here, I emphasise again that the main idea of TACT is to introduce a simple general-purpose way of thinking about and communicating the meaning of correlations for individuals. For this, the default approach of trisecting variables probably works the best.

[3] The ratio of sensitivity to the opposite of specificity (100% - specificity) is often called the *likelihood ratio* and the plot of different likelihood ratios resulting from different cut-offs is often called the *Receiver Operator Characteristic (ROC) Curve*. The area under the curve (in relation to the total area of the graph; AUC) represent the test's accuracy. Test developers often calculate the ROCs for different cut-offs, looking for the higest achievable AUC.

What if the Variables Are Distributed Differently in the Population?

As presented so far, TACT has been based on variables with bell-shaped population distributions. But the TACT R function can also be used to assess correlations between variables that have uniform (all values are equally likely) or skewed (values in one extreme are more likely than medium values and especially values in the other extreme) distributions.

Generally, the TACT is relatively robust to how the variables are distributed. However, the accuracy in applying correlations to individuals—the probabilities of values in one variable matching similar values in another—tends to be somewhat smaller the more uniformly the variables are distributed. For example, this can be experimented with:

```
> TACT(r = .25, distribution = "uniform").
```

What if the Correlation is Already About Individuals?

Increasingly, researchers measure their participants at many time points and the resulting time-series data allow for calculating correlations between variables' values at different time points. These correlations describe variance trends within, not between, individuals. They can be unique to each individual, although they are often aggregated to sample-level estimates representing an average individual and becoming another population-level trend ("fixed effect"). TACT can also be applied to these correlations, with the only difference being that individuals are swapped for measurement occasions.

For example, using such a within-individual design Wieczorek and colleagues (2022) found that perceptions of how expressive people are in social interactions have a .60 correlation with how satisfied they are with these interactions—this is an unusually strong relation in the psychological research context. Using TACT, we can then say that the average participant's satisfaction with a given interaction can be predicted from their expressiveness in this interaction with 53.4% accuracy against the 33.3% random-guess baseline. Specifically, if the average participant is more expressive than they are in two-thirds of their interactions, then there is a 59.9% probability that they are also experiencing commensurate satisfaction; for medium (typical to the person) expressiveness to match medium (typical to the person) satisfaction, the probability is 40.4%, all against the 33.3% chance level.

What About Measurement Error?

No variable is measured with perfect accuracy. Random measurement error biases correlations downwards, so our ability to say something about individuals based on correlations could be greater if we somehow fixed the correlations. The best way to do this is to reduce measurement error in the first place. *Post hoc* adjustments of correlations for measurement error do not always help because the uncertainty about where individuals are located in the scatterplot remains unaddressed. The same applies to systematic measurement error, which could both inflate and deflate correlations.

However, in cases where one variable is clearly the predictor and the other being predicted, it could make sense to adjust the correlations for measurement error in the latter before TACTing the correlation. This is because it is the hypothetical true value that is being predicted rather than its imperfect measurement. Also, when the association is interpreted in relation to variables' hypothetical true values rather than their measured values, it may be useful to adjust the correlations for measurement error. For example, when we estimate individuals' hypothetical trait scores' stability over time rather than the stability of the trait's measurements—as was discussed above—we may adjust the rank-order stability for measurement error before TACTing it.

What About Non-Linear Associations?

TACT is not suitable for interpreting linear correlation coefficients (incorrectly) calculated for relations that are actually non-linear. However, TACTing the scatterplot for these variables' associations, when raw data are available, can be particularly useful for illustrating the non-linearity. This can be achieved with the TACT function of the TACT R package.

Conclusion

Typical correlations that emerge from psychological research can be useful for showing trends in the population. Provided that necessary conditions are met, these trends may inspire cost-effective population-level interventions that could benefit small but occasionally worthwhile proportions of these populations (Funder & Ozer, 2019). But using these correlations to draw meaningful conclusions about particular individuals (e.g., "*the person is likely to score high in X as they score high on Y*") can be more difficult than many expect, and this should generally be avoided altogether. Statistically speaking, such conclusions are often inaccurate (most people are not high in X, even given high Y) or misleading (given high Y, someone is only marginally more likely to be high than medium or low on X, making little noticable difference). TACT can help to think about the implications of correlational research findings for individuals and how to communicate these to the public.

Funding: The author has no funding to report.

Acknowledgments: I am grateful to Wendy Johnson, Michelle Luciano, Lisanne de Moor, Yavor Dragostinov, Sam Henry, Ross Stewart, Manuel Maldonado, Marco Del Guidice, Kadri Arumäe and Emma Mõttus for their comments on the drafts of this article.

Competing Interests: The author has declared that no competing interests exist.

Author Contributions: *René Mõttus*—Idea, conceptualization | Design planning | Visualization (data presentation, figures, etc.) | Data analysis | Validation, reproduction, checking | Writing | Feedback, revisions | Project coordination, administration.

Other Notes: The editor-in-chief (who also handled this paper as the acting editor) encouraged the author to submit a paper based on their post at https://www.ejp-blog.com/blog/2021/8/2/does-a-correlation-say-anything-about-an-individual-a-trinomial-effect-size-display

Related Versions: The manuscript has been available as a pre-print at https://psyarxiv.com/bpm9y

Data Availability: For this article, data is freely available (Mõttus, 2022).

Supplementary Materials

The supplementary material contains TACT propabilities for a range of correlation magnitudes and statistical (R) software for calculating these probabilities in similated and actual data (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Mõttus, R. (2022). *Supplementary materials to "What correlations mean for individual people: A tutorial for researchers, students and the public"* [Data]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.8275

Mõttus, R. (2022). *Supplementary materials to "What correlations mean for individual people: A tutorial for researchers, students and the public"* [Code]. GitHub. https://github.com/mottusrene/TACT

Personality Science. (Ed.). (2022). *Supplementary materials to "What correlations mean for individual people: A tutorial for researchers, students and the public"* [Open peer-review]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.8276

References

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130*(6), 887–919. https://doi.org/10.1037/0033-2909.130.6.887

Briley, D. A., & Tucker-Drob, E. M. (2014). Genetic and environmental continuity in personality development: A meta-analysis. *Psychological Bulletin, 140*(5), 1303–1331. https://doi.org/10.1037/a0037091

Funder, D. C. (2019). *The personality puzzle* (8th ed.). W. W. Norton

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Graham, E. K., Rutsohn, J. P., Turiano, N. A., Bendayan, R., Batterham, P. J., Gerstorf, D., Katz, M. J., Reynolds, C. A., Sharp, E. S., Yoneda, T. B., Bastarache, E. D., Elleman, L. G., Zelinski, E. M., Johansson, B., Kuh, D., Barnes, L. L., Bennett, D. A., Deeg, D. J. H., Lipton, R. B., . . .Mroczek, D. K. (2017). Personality predicts mortality risk: An integrative data analysis of 15 international longitudinal studies. *Journal of Research in Personality, 70*, 174–186. https://doi.org/10.1016/j.jrp.2017.07.005

Henry, S., Thielmann, I., Booth, T., & Mõttus, R. (2022). Test-retest reliability of the HEXACO-100—And the value of multiple measurements for assessing reliability. *PLoS One, 17*(1), Article e0262465. https://doi.org/10.1371/journal.pone.0262465

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *The Journal of Applied Psychology, 75*(3), 334–349. https://doi.org/10.1037/0021-9010.75.3.334

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jokela, M., Elovainio, M., Nyberg, S. T., Tabák, A. G., Hintsa, T., Batty, G. D., & Kivimäki, M. (2014). Personality and risk of diabetes in adults: Pooled analysis of 5 cohort studies. *Health Psychology, 33*(12), 1618–1621. https://doi.org/10.1037/hea0000003

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *The Journal of Applied Psychology, 98*(6), 875–925. https://doi.org/10.1037/a0033901

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain, 8*(6), 221–223. https://doi.org/10.1093/bjaceaccp/mkn041

Loehlin, J. C., Bowles, S., Gintis, H., & Osborne Groves, M. (2005). Resemblance in personality and attitudes between parents and their children: Genetic and environmental contributions. In S.

Bowles, H. Gintis, & M. Osborne Groves (Eds.), *Unequal chances: Family background and economic success.* (pp. 192–207). Princeton University Press. https://doi.org/10.1515/9781400835492.192

Luo, Z. C., Albertsson-Wikland, K., & Karlberg, J. (1998). Target height as predicted by parental heights in a population-based study. *Pediatric Research, 44*(4), 563–571. https://doi.org/10.1203/00006450-199810000-00016

McCrae, R. R., Costa, P. T., Martin, T. A., Oryol, V. E., Rukavishnikov, A. A., Senin, I. G., Hřebíčková, M., & UrbĆünek, T. (2004). Consensual validation of personality traits across cultures. *Journal of Research in Personality, 38*(2), 179–201. https://doi.org/10.1016/S0092-6566(03)00056-4

McCrae, R. R., & Mõttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science, 28*(4), 415–420. https://doi.org/10.1177/0963721419849559

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality, 52*, 47–54. https://doi.org/10.1016/j.jrp.2014.07.005

Mõttus, R., Sinick, J., Terracciano, A., Hrebickova, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability and utility of personality nuances. *Journal of Personality and Social Psychology, 117*(4), e35–e50. https://doi.org/10.1037/pspp0000202

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the “same” construct? A meta-analytic investigation. *Personality and Individual Differences, 49*(7), 669–676. https://doi.org/10.1016/j.paid.2010.06.014

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*(2), 166–169. https://doi.org/10.1037/0022-0663.74.2.166

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Strickhouser, J. E., Zell, E., & Krizan, Z. (2017). Does personality predict health and well-being? A metasynthesis. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association, 36*(8), 797–810. https://doi.org/10.1037/hea0000475

Sutin, A. R., Terracciano, A., Deiana, B., Naitza, S., Ferrucci, L., Uda, M., Schlessinger, D., & Costa, P. T. (2010). High neuroticism and low conscientiousness are associated with interleukin-6. *Psychological Medicine, 40*(9), 1485–1493. https://doi.org/10.1017/S0033291709992029

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *The Journal of Applied Psychology, 23*(5), 565–578. https://doi.org/10.1037/h0057079

Terracciano, A., Costa, P. T., & McCrae, R. R. (2006). Personality plasticity after age 30. *Personality and Social Psychology Bulletin, 32*(8), 999–1009. https://doi.org/10.1177/0146167206288599

Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. *Journal of Personality, 87*(3), 715–730. https://doi.org/10.1111/jopy.12428

Vainik, U., Dagher, A., Realo, A., Colodro-Conde, L., Mortensen, E. L., Jang, K., Juko, A., Kandler, C., Sørensen, T. I. A., & Mõttus, R. (2019). Personality-obesity associations are driven by narrow traits: A meta-analysis. *Obesity Reviews, 20*(8), 1121–1131. https://doi.org/10.1111/obr.12856

Wieczorek, L. L., Mueller, S., Lüdtke, O., & Wagner, J. (2022). What makes for a pleasant social experience in adolescence? The role of perceived social interaction behavior in associations between personality traits and momentary social satisfaction. *European Journal of Personality, 36*(5), 787–808. https://doi.org/10.1177/08902070211017745

eapp

Personality Science (PS) is an official journal of the European Association of Personality Psychology (EAPP).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.