Untrustworthy Reporting of Results in I-O (& Psychology in general)

From a psychologist writing in a popular magazine:
"Although IQ tests are very unpopular, they have been known to predict important outcomes for decades. For example, **children's IQ scores predict not only how well they will do at school and college, but also how long they will live** (even after controlling for socio-economic status)."

The stand-out message is that your IQ predicts how long you will live.

It takes many hours to find, read, digest, and compute additional indices from the actual scientific evidence …

From: Borghans, L., Golsteyn, B.H.H., Heckman, J.J., & Humphries, J.E. (2011). Identification problems in personality psychology. *Personality and Individual Differences*, 51, 3, 315-320
"Borghans, Golsteyn, Heckman, and Humphries (2011) examine the predictive power of grades, IQ and achievement tests measured in the adolescent years for a variety of life outcomes past age 30 *(the outcomes include wages, income, hours worked, depression, smoking, physical activity, health, voting, divorce and unemployment*.) **The R^2 of most relationships is below 0.10**", p. 317.

From: Deary, I.J., Weiss, A., & Batty, D. (2011). Outsmarting mortality. *Scientific American Mind*, July/Aug, 48-55.
"The findings are unequivocal, although few health practitioners are aware of them. The lower a person's measured intelligence, **the greater that individual's risk of living a shorter time**, developing both mental and physical ailments later in life and dying from cardiovascular disease, suicide or an accident. More surprising still is that low intelligence is a stronger predictor than several better known risk factors for illness and death, such as obesity and high blood pressure. " p. 50

From: Deary, I.J., Weiss, A., & Batty, G.D. (2010). Intelligence and Personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest*, 11, 2, 53-79.
"In this study, 938 participants from the Midspan prospective cohort studies, initiated in the 1970s, were, on the basis of their birth date, linked to their intelligence test scores at age 11, as captured using the Scottish Mental Survey 1932 (Hart et al., 2004). After approximately 3 decades of mortality and morbidity surveillance, a 1-SD disadvantage in intelligence at age 11 was related to an **11% increased risk** of hospital admission or death due to cardiovascular disease. This observation has been replicated in other cohorts drawn from Scotland (Deary, Whiteman, Starr, Whalley, & Fox, 2004) and Sweden (Hemmingsson, Melin, Allebeck, & Lundberg, 2006). ", p. 62.

Digging into some primary data …
Batty, G.D., Deary, I.J., & Gottfredson, L.S. (2007). Premorbid (early life) IQ and Later Mortality Risk: Systematic Review. *Annals of Epidemiology*, 17, 4, 278-288, looking at the study-data and effect sizes reviewed in Table 1, pp. 280-281, it's possible to compute a few useful probabilities from the Hazard Ratios presented in the table.

Scottish Mental Survey (1932) (24) (B)	Scottish retrospective cohort study from 1932–1997	Version of the Moray House Test No. 12 at 11 years (principally assesses verbal reasoning; 71 items concerning general, spatial and numerical reasoning) (33)	633 deaths in 1167 men; 438 deaths in 1050 women (national death registers)	Hazards ratios for *survival* (men)/1-SD decrease in IQ: 0.79 (0.75–0.84) Odds ratios (men [77]) Quartile 1: 1.0 (referent) Quartile 2: 1.52 (1.08–2.14) Quartile 3: 1.59 (1.13–2.23) Quartile 4: 1.81 (1.29–2.52)

The Hazard Ratio (HR: "a hazard is the rate at which events happen, so that the probability of an event happening in a short time interval is the length of time multiplied by the hazard. Although the hazard may vary with time, the assumption in proportional hazard models for survival analysis is that the hazard in one group is a constant proportion of the hazard in the other group. This proportion is the hazard ratio" http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/what_are_haz_ratios.pdf)
can be expressed as a probability of occurrence of the event in question, given one or more differentiating factors (here, the probability of mortality given amount of IQ)

$$p = \frac{HR}{1 + HR}$$

So an HR of 0.79 for a 15-point decrease in IQ at age 11 years translates to a probability of 0.44 that the individual will die sooner by a certain time than members of a group 15 points or more higher. But this is a relative risk of dying, not the prediction of death itself.

If we take the odds ratio of 1.81 (comparing the highest scoring IQ quartile with the lowest quartile), and re-express that as a correlation:

$$d = \ln(OR) * \frac{\sqrt{3}}{\pi}$$

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

we can compute a correlation effect size as:

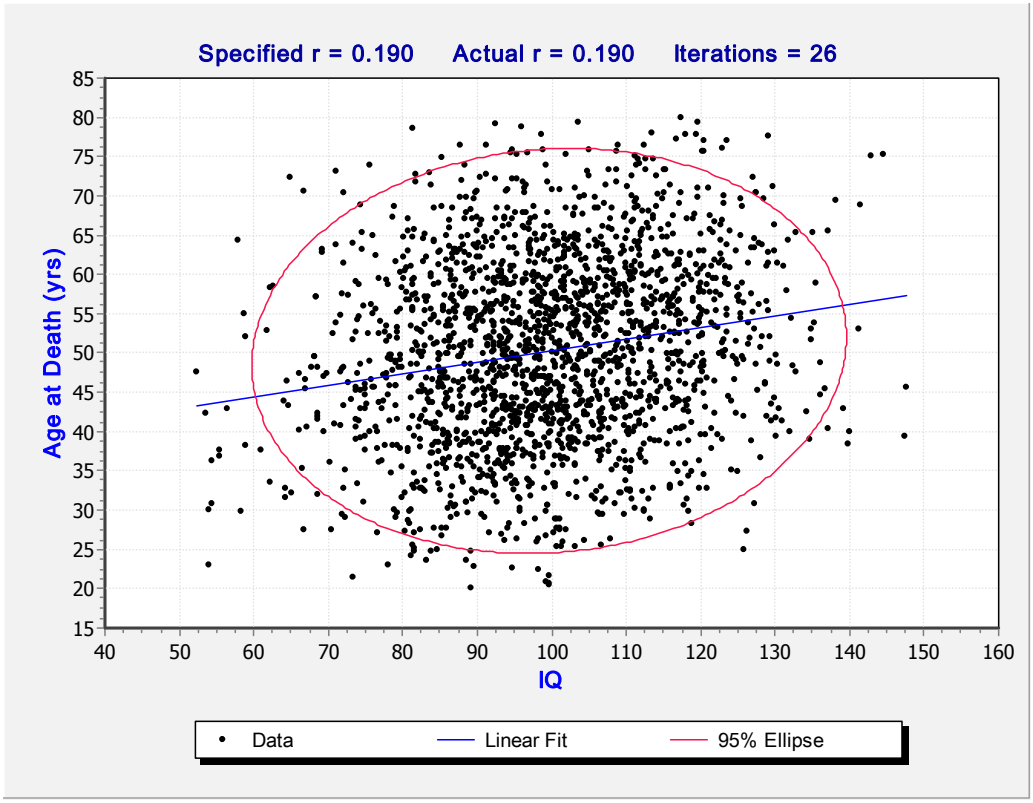
$$OR := 1.81 \qquad d := \ln(OR) \cdot \left(\frac{\sqrt{3}}{\pi}\right) = 0.327 \qquad r := \frac{d}{\sqrt{(d^2 + 4)}} = 0.161$$

In the end, the article by Whalley and Deary provides the clearest expression of a predictive function (also described in a later article: Deary, I., Whiteman. M.C., Starr, J.M., Whalley, L.J. &, & Fox, H.C. (2004). The impact of childhood intelligence on later life: following up the Scottish Mental Health Surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, 86, 1, 130-147).

Whalley, L.J., & Deary. I.J. (2001). Longitudinal cohort study of childhood IQ and survival up to age 76. *British Medical Journal*, 322, 819-822 states in its abstract:
"Childhood mental ability was positively related to survival to age 76 years in women (0.978 (0.971 to 0.984), P < 0.0001) and men (0.989 (0.984 to 0.994), P < 0.0001). A 15 point disadvantage in mental ability at age 11 conferred a relative risk of 0.79 of being alive 65 years later (95% confidence interval 0.75 to 0.84); a 30 point disadvantage reduced this to 0.63 (0.56 to 0.71)."

But of real significance is the reported correlation between Age at death and IQ … which directly addresses the claim: children's IQ scores predict not only how well they will do at school and college, but also how long they will live. **The correlation between IQ at age 11 and age at death after father's occupation and overcrowding were controlled for was 0.19** (P < 0.001), p. 821.

I generated 2000 cases of data for two variables which correlate at 0.19, which looks like:



That there is a relation at all is intriguing, and as a mortality 'risk' it is clearly of interest. So why exaggerate/spin the facts?

A child's IQ score does not predict how long they will live. But there is a replicable, albeit small relationship between IQ and mortality.

---

2 A published article:

Abstract

This paper develops and tests links between the reinforcement sensitivity theory of personality and senior-executive job performance, hypothesising that the theory's personality traits, known as 'BIS' and 'BAS', will interact to predict performance. Structural equation modeling showed that while BAS has no main effect and BIS has a marginally significant effect on performance (p = 0.07), **BIS and BAS interact to predict performance (p = 0.01)**, the optimal scenario being a combination of high BAS and low BIS. These results show the importance of testing traits' interactions in applied personality research.

Two issues here:

❶ The results appear to show the opposite to: "the importance of testing traits' interactions in applied personality research."

❷ And even if you ignored #1, the model explains just 9% of job performance.

❶ The best-fitting (one BAS item removed) measurement model consisting of BAS, BIS, and Job Performance items loading on each of their respective latent variables fitted with a **chi-square of 354.16**, with **230 df**. I'm assuming this is just a model with three latent variables, no paths between them, just testing that the respective items are associated with their respective latent variable.

The 'structural' model tested now introduced a new interaction latent variable, along with separate BAS and BIS latent variables, all predicting an overall Job Performance variable.

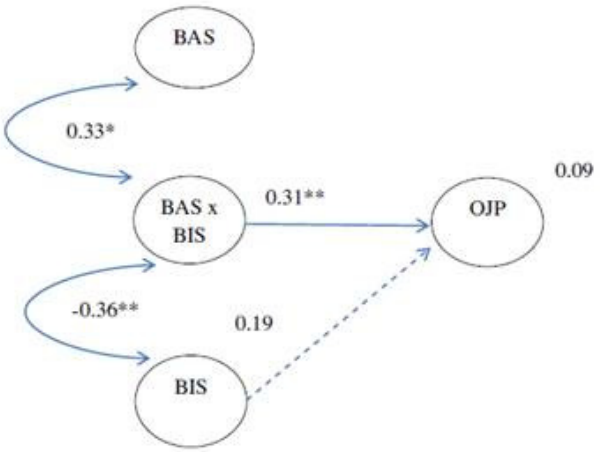


Fig. A.1. Final model showing main and interactive effects of traits BIS and BAS on overall job performance. Latent variables of trait BAS, trait BIS, and their interaction, predicting a dependent latent variable, overall job performance. Curved arrows show the correlations between BAS and BIS and the interaction term, while straight arrows show standardised regression estimates. BAS did not significantly predict OJP, BIS had a marginally significant relationship with OJP, and the interaction had a significant relationship. $^{*}p < .05$, $^{**}p < .01$; OJP = overall job performance.

This structural model fit with **chi-square of 369.25**, with **248 df**.

Assuming the models are nested, a **chi-square difference test** between these two models would be: (369.25 - 354.16) = **15.09** with (248-230) =**18 df** which indicates a probability of **0.66**, meaning there is no statistically significant difference between the 'interaction' model vs one where all three latents are completely independent from one another (the measurement model).

I was in contact with the authors about this issue prior to final publication (I saw the earlyview article). I'm not going to go into any details except to say they relied upon an interpretation of what was in Hair et al's textbook on multivariate analysis:
"My understanding (from Hair et al.'s 2010 textbook) is that the measurement model will always be the best fit, once you've got that measurement model in place you add your structural parameters and as long as the model still meets the minimum fit criteria you can interpret the resulting estimates as findings."

My response:
As I see it, the issue simply comes down to:
[Authors]: My understanding (from Hair et al.'s 2010 textbook) is that the measurement model will always be the best fit, once you've got that measurement model in place you add your structural parameters and as long as the model still meets the minimum fit criteria you can interpret the resulting estimates as findings.
vs
[Me]: Assuming the models are nested, a chi-square difference test between these two models would be: (369.25 - 354.16) = **15.09** with (248-230) =**18 df** which indicates a probability of **0.66**, meaning there is no statistically significant difference between the 'interaction' model vs one where all three latents are completely independent from one another (the measurement model).

In the former statement, you (via Hair et al) rely upon the model fit indices for your new model fitting the data according to whatever criteria you choose. If the new model fits, it is a candidate model. If the hypothesis driven model fits, it is your chosen model.

In the latter statement, I am computing the significance of the difference between the fitting models. If there is no difference (as the significance test indicated), then the more complex model (the interaction one) is not statistically significantly different from one where three independent latents account for the covariance within the same model-implied covariance matrix i.e. the residuals between the sample and model-implied covariance matrices do not differ between the fitted models. So, you have the awkward situation where two models fit your data equally well (statistically), which begs the question how you now might choose between them!

A more formal statement of the need for formal comparison of models is given in a small target article at:
http://www.psychologie.uzh.ch/fachrichtungen/methoden/team/christinawerner/sem/chisquare_diff_en.pdf

I've never seen that statement "the measurement model will always be the best fit" before … because the measurement model is invariably a CFA with what are effectively orthogonal latents (or at least 'unmeasured' interconnecting paths). So, this model (like the null model) should always fit more poorly than one with structural paths added. That's the whole point of model-building ... to move beyond the simple measurement model by attaining models which better fit the asymptotic model-based covariance matrix. The chi-square difference test evaluates the fit relative to the number of extra parameters used to attain a better fit.

I also found a handy 'in-depth' document which explains things in much more detail.
http://www.cob.unt.edu/slides/paswan/BUSI6280/Structural%20Equation%20Modeling_Nov122007.doc

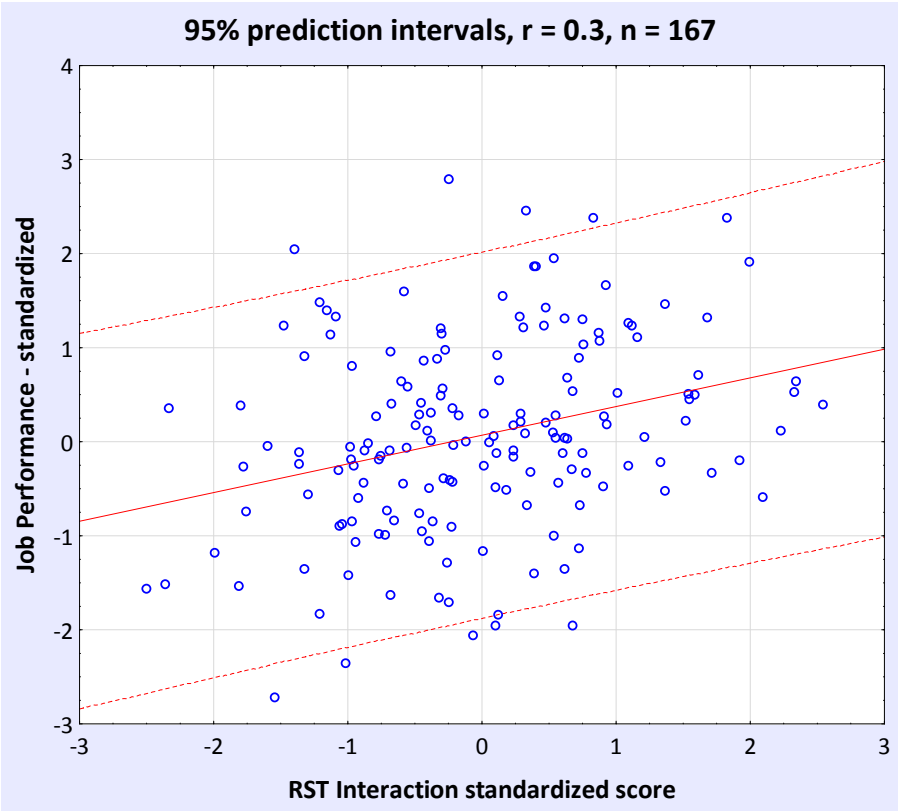
And slides 43 and 44 provide again the chi-square difference test logic …
http://www.psych.yorku.ca/cribbie/SEM%20Course%202012/Intro%20to%20SEM_day%201_nov2012.pdf

Do not Hair et al discuss the logic and rationale of the formal statistical testing of competing nested models, and the use of AIC/BIC differences for non-nested models? Every SEM book I have goes into this issue in some detail.

The simple analogy is comparing two groups' mean scores. A single sample t-test conducted on each of both means show they are both significantly different from zero; then, you subjectively choose the one that is higher than the other as per you hypothesis. But, surely the correct way to do this is to establish whether the two means actually do differ from one another statistically, with an appropriate independent-groups t-test?

Right now, because of this unresolved issue of comparative model fit, much of their discussion is rendered 'untrustworthy' because, statistically-speaking, the interaction model does not fit the data any better than the model not containing that interaction.

❷ Given a correlation of just 0.3 (9% explained variation), the interactive term is hardly a substantive predictor of anything. As with all such smallish-sample studies (n=167), the use of prediction intervals around the regression line predicting job performance from the interactive variable shows just how 'untrustworthy' the prediction is in reality:



As Dr. McCoy might have said to Captain Kirk in Star Trek ... “This is not prediction as we know it Jim”.

“Trustworthy” requires that results are reported accurately, with the implications of the smallish-sample size taken into account, and suitably qualified statements made about the degree of ‘predictive accuracy’ and how the current results might be understood.

From another article, assessing whether self-ratings of personality during adolescence predict CWBs in adulthood. The author/s claim:

"In sum, we found evidence that personality traits such as Agreeableness and Conscientiousness reliably predicted Interpersonal and Organizational CWBs two decades later."

The problem for the authors is that their data showed personality attribute scores assessed in adolescents are extremely *inaccurate* predictors of CWBs 20 years later.

If we were to use Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40, 5, 532-538, Table 1, as our guideline for how results might be described in text, we would conclude from their data that no personality attribute reaches even the recommended minimum practical effect size (0.20).

But, many might feel that relying upon recommendations concerning how to describe effect sizes is not a sound basis for making claims about predictive accuracy.

If we just work with what the authors have presented, do their results substantiate a claim: "In sum, we found evidence that personality traits such as Agreeableness and Conscientiousness reliably predicted Interpersonal and Organizational CWBs two decades later"? If we simply use the coefficient of determination to express predictive accuracy, from their reported results we would conclude that two personality attributes each explain **3%** of the variation in CWBs.

But, strangely enough, no analysis attempt was made to predict CWB frequency using personality scores. The most obvious analysis would have been three multiple regression equations, for each CWB variable as the dependent variable, and personality attributes as predictors. Given the author/s seem to assume linearity and normal distributions for all their current analyses, they might have used a standard multiple linear regression. They could otherwise have used a multinomial logistic or even some variant of linear discriminant function analysis or preferably a nonlinear classification and regression tree. All these methods yield direct empirical estimates of predictive accuracy, in contrast to the indirect methods they rely on in their article.

As an aside, I computed the multiple R for the correlation between Agreeableness and Conscientiousness with Global CWB (using their reported bivariate correlations):

$$ry1 := -0.18$$
$$ry2 := -0.18$$
$$r12 := .43$$
$$R := \sqrt{\frac{\left(ry1^2 + ry2^2 - (2 \cdot ry1 \cdot ry2 \cdot r12)\right)}{\left(1 - r12^2\right)}} = 0.2128724626$$

It is not very impressive.

But let's get right down to the data. With the current results they report, what are the 95% and 80% prediction intervals for a frequency of CWB given a particular personality scale score (say Conscientiousness). This is the most direct way of ascertaining the accuracy of prediction given the size of sample the author/s employed, and the linear relationship they observed. The correlation is reported as **-0.18**.

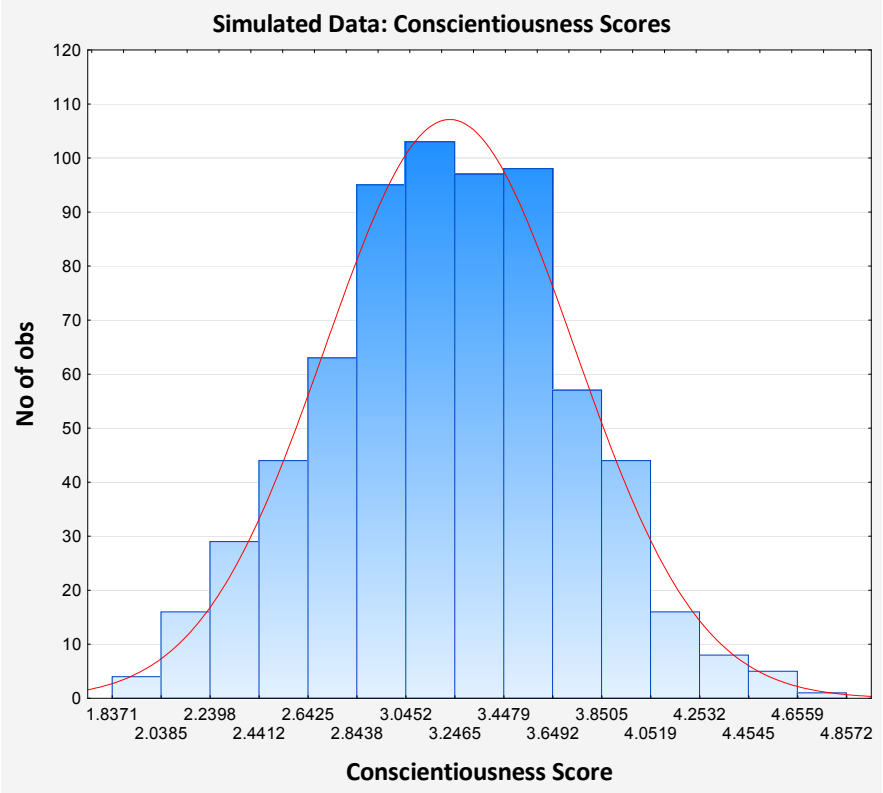
Because the data are most likely skewed badly, I first created a normally distributed dataset from which I selected a subset conforming closely to the distribution parameters provided by the author/s in their table of descriptive statistics.

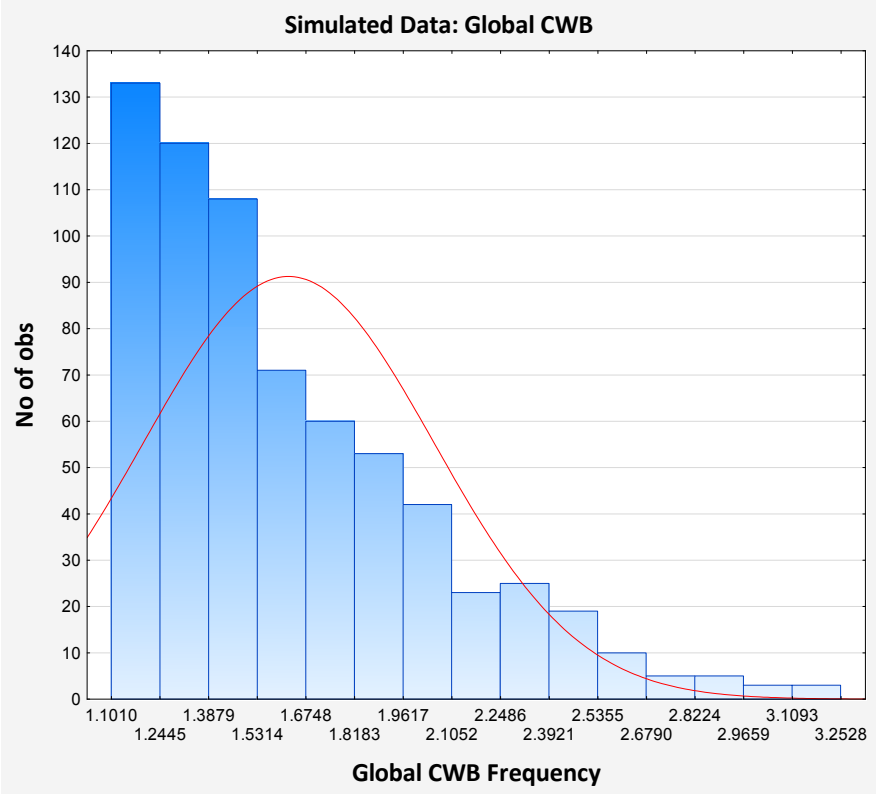
I sampled from continuous-real valued distributions, as the author/s do not report integer-sum scale scores but integer-sum divided by the number response categories.

And I increased the sample size so as to limit the prediction error estimate and provide a 'generous' estimate of prediction error.

	Conscientiousness		Global CWB	
	Actual	Simulated	**Actual**	Simulated
N	~300	680	~300	680
Mean	3.53	3.22	1.63	1.62
SD	.53	0.51	0.59	0.43

The distributions of each simulated variable are:





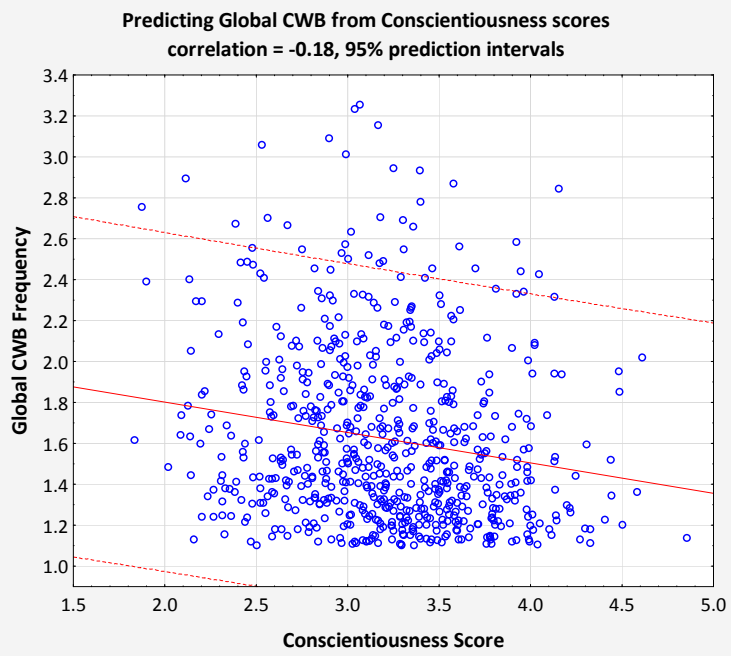
The correlation within the simulated data between these two variables is **-0.18**. The descriptives are:

	Descriptive Statistics (CorViz dataset r = -0_18, n=680.sta)							
Variable	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.	Skewness	Kurtosis
X	680	1.62	1.51	1.10	3.25	0.43	1.111	0.931
Y	680	3.22	3.21	1.84	4.86	0.51	0.037	-0.160

And the Frequency distribution of simulated CWB is:

From To	Frequency table: - Simulated Global CWB (r = -0_18, n=680.sta) Count	Cumulative Count	Percent	Cumulative Percent
1 <x<=1.2	88	88	12.94	12.94
1.2 <x<=1.4	174	262	25.59	38.53
1.4 <x<=1.6	133	395	19.56	58.09
1.6 <x<=1.8	89	484	13.09	71.18
1.8 <x<=2	70	554	10.29	81.47
2 <x<=2.2	52	606	7.65	89.12
2.2 <x<=2.4	30	636	4.41	93.53
2.4 <x<=2.6	24	660	3.53	97.06
2.6 <x<=2.8	9	669	1.32	98.38
2.8 <x<=3	5	674	0.74	99.12
3 <x<=3.2	4	678	0.59	99.71
3.2 <x<=3.4	2	680	0.29	100.00
Missing	0	680	0.00	100.00

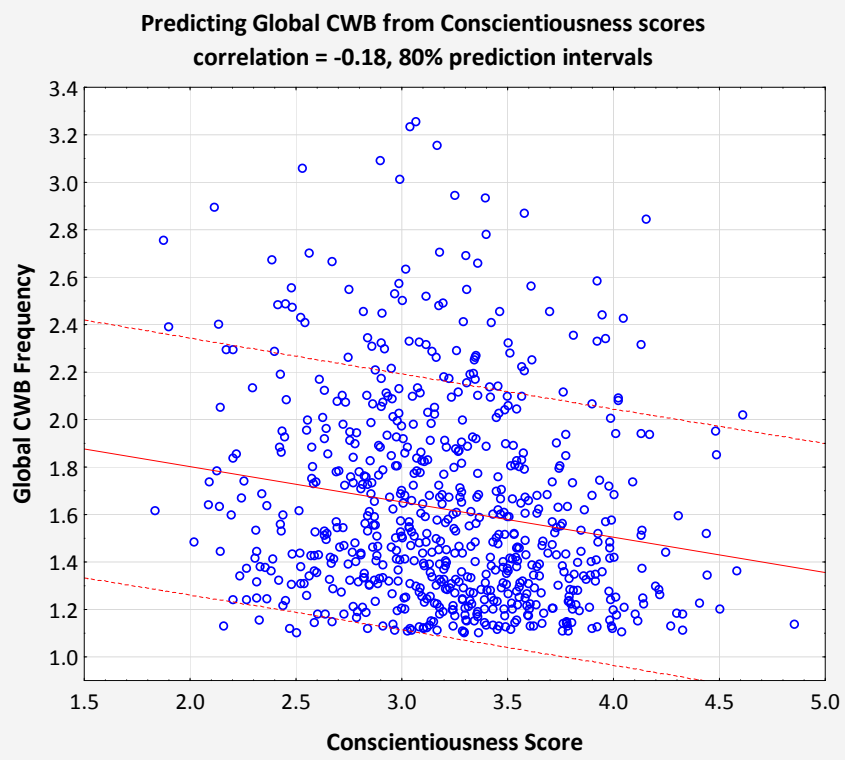
The Scatterplot below shows the linear relationship and the 95% prediction intervals associated with predicting a CWB frequency-score from a Conscientiousness score.



For a Conscientiousness score of 1.5, a predicted CWB score would vary between 1.00 and 2.71
For a Conscientiousness score of 2.0, a predicted CWB score would vary between *0.97 and 2.63
For a Conscientiousness score of 3.0, a predicted CWB score would vary between *0.83 and 2.47
For a Conscientiousness score of 4.5, a predicted CWB score would vary between *0.6 and 2.26
* *the lower bound exceeds the minimum possible observed frequency score, caused by the skew in the CWB data.*

Clearly, predictive accuracy is very low for all practical purposes … exactly what Ferguson (2009) implied with is RMPE recommendation.

The Scatterplot below shows the linear relationship and the **80%** prediction intervals associated with predicting a CWB frequency-score from a Conscientiousness score.



For a Conscientiousness score of 1.5, a predicted CWB score would vary between 1.33 and 2.42
For a Conscientiousness score of 2.0, a predicted CWB score would vary between 1.26 and 2.34
For a Conscientiousness score of 3.0, a predicted CWB score would vary between 1.11 and 2.19
For a Conscientiousness score of 4.5, a predicted CWB score would vary between *0.89 and 1.97
* *the lower bound exceeds the minimum possible observed frequency score, caused by the skew in the CWB data.*

Clearly, predictive accuracy is still very low for all practical purposes.

The reality is you don't need to see the prediction intervals to realize that a **-0.18** correlation is not sufficient to justify statements about 'reliable predictive accuracy'.

We could compute the BESD (Rosenthal, R., & Rubin, D.R. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 2, 166-169), assuming a binary classification and equal marginals, for a **-0.18** correlation This would yield an overall classification accuracy of 59%, where 50% is considered correct classification by chance alone.

We could also re-express this BESD as a Bayesian-oriented estimate of the relative improvement over chance, using the RIOC statistic from Loeber, R. and Dishion, T. (1983). Early Predictors of male delinquency: a review. *Psychological Bulletin*, 94, 68-99, which would indicate an improvement of just 0.21.

But, whichever way you look at data which correlate at only **-0.18**, predictive accuracy is poor. So why the pretence?

This article reported the results using data acquired from a sample of working adults; regressing age, gender, Big Five personality trait scores, work meaningfulness, and the presence/absence of a workplace policy on inappropriate use of the internet while at work (where the participants self-rated such internet use).

The authors stated in their Introduction: "More relevant to this study, the Big Five have been found to predict the amount of individual Internet use by college students (Landers & Lounsbury, 2006; McElroy, Hendrickson, Townsend, & DeMarie, 2007)."

I was curious about this claim. So I took a look at the primary references…
In Landers & Lounsbury, 2006, Table 3 shows us:

288 *R.N. Landers, J.W. Lounsbury / Computers in Human Behavior 22 (2006) 283–293*

Table 3
Intercorrelations of personality variables with percent time spent on the Internet by category of usage

	Percent time spent on		
	Communication	Leisure	Academic
Agreeableness	0.01	−0.06	−0.02
Conscientiousness	0.01	−0.18*	0.19*
Emotional stability	−0.10	0.06	0.02
Extraversion	0.07	−0.12	0.10
Openness	−0.02	−0.17	0.10

Only two correlations are significant at $p < 0.05$. These data simply do not justify the statement: "the Big Five have been found to predict the amount of individual Internet use by college students". Indeed, no attempt at properly quantifying actual predictive accuracy is made.

From McElroy, Hendrickson, Townsend, & DeMarie, 2007, Table 1, we see …

Table 1. Descriptive Statistics and Correlations (N = 153)

Variable	Mean	Std. Dev.	CA	Self Eff	Gender	Agree	Consc	Extrav	Neuro	Open
Computer Anxiety (CA)	2.32	.36	**.72**							
Self-Efficacy (Self Eff)	4.00	.44	-.23**	**.80**						
Gender	1.55	.58	.02	-.03	**na**					
NEOPIRA (Agree)	116.30	18.65	-.07	.19*	.09	**.89**				
NEOPIRC (Consc)	126.84	18.47	-.11	.59***	.13	.27***	**-.90**			
NEOPIRE (Extrav)	119.68	21.16	-.24**	.36***	.06	.06	.34***	**.91**		
NEOPIRN (Neuro)	85.77	23.67	.15	-.52***	.17*	-.32***	-.38***	-.27***	**.93**	
NEOPIRO (Open)	115.16	20.40	-.26**	.24**	.14	.02	.04	.44***	-.13	**.90**
MBTI T-F	1.99	13.62	-.13	.19*	-.25**	-.23**	.29***	-.12	-.26**	-.14
MBTI E-I	2.28	12.46	.04	.13	-.04	.02	.00	.67***	-.14	.22**
MBTI S-N	-.16	13.68	.21*	-.06	.04	.08	.26**	-.20*	.21*	-.61**
MBTI J-P	5.45	12.83	.15	.10	.16	.09	.40***	-.08	.14	-.31***
Internet Use	3.54	.58	-.29***	.13	-.06	-.08	.03	.25**	.04	.30***

Two out of 5 significant correlations ... Again, these data simply do not justify the statement "the Big Five have been found to predict the amount of individual Internet use by college students". Table 2 in the McElroy article does assist here, although only one beta-weight (Openness) is now statistically significant.

Table 2. Regression Analyses of the Big Five Pers Controlling for Computer Anxiety, Self-Efficacy, a

	Internet Use	
	Step 1	**Step 2**
Computer Anxiety	-.27**	-.22*
Self-Efficacy	.06	.07
Gender	-.05	-.12
Step 2		
Agreeableness		-.07
Conscientiousness		.02
Extraversion		.14
Neuroticism		.19
Openness		.21*
F	4.13**	3.61***
F Change	4.13**	3.09**
Change in R^2	.09	.10
Adjusted R^2	.07	.14

The statements claiming the "Big Five predicts" hardly looks like trustworthy reporting.

I also noted the regression results reported in Table 2 do not adjust the r-squares for shrinkage (in contrast to McElroy, Hendrickson, Townsend, & DeMarie (2007) who do). So, I made the corrections:
For step 1 R^2 was 0.23, it now becomes 0.22
For step 2 R^2 was 0.31, it now becomes 0.28
For step 3 R^2 was 0.36, it now becomes 0.31
The incremental effect for personality is now 6% instead of 8%.

According to Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40, 5, 523-538, the incremental R^2 represents a 0.02 increase over the *recommended minimum effect size* representing a "practically" significant effect for social science data. Does that really justify the author/s statement on page 11 ... lines 11-13 "After controlling for gender and age, the Big Five traits in aggregate explained significant incremental variance, thus demonstrating the usefulness of the trait approach in this area of research."?

In what way is a 6% increase in R^2 *significant*, except in the statistical sense of that word. Perhaps if the author/s computed *predicted* inappropriate internet-use self-ratings with and without the use of personality attributes, we might have the kind of information we need to evaluate just how 'significant' are these particular attributes as predictors of this kind of internet usage.

5 The published article:

Connelly, B., & Ones, D. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136, 6, 1092-1122.

Abstract

The bulk of personality research has been built from self-report measures of personality. However, collecting personality ratings from other-raters, such as family, friends, and even strangers, is a dramatically underutilized method that allows better explanation and prediction of personality's role in many domains of psychology. Drawing hypotheses from D. C. Funder's (1995) realistic accuracy model about trait and information moderators of accuracy, we offer 3 meta-analyses to help researchers and applied psychologists understand and interpret both consistencies and unique insights afforded by other-ratings of personality. These meta-analyses integrate findings based on 44,178 target individuals rated across 263 independent samples. Each meta-analysis assessed the accuracy of observer ratings, as indexed by interrater consensus/reliability (Study 1), self– other correlations (Study 2), and predictions of behavior (Study 3). The results show that although increased frequency of interacting with targets does improve accuracy in rating personality, informants' interpersonal intimacy with the target is necessary for substantial increases in other-rating accuracy. Interpersonal intimacy improved accuracy especially for traits low in visibility (e.g., Emotional Stability) but only minimally for traits high in evaluativeness (e.g., Agreeableness). In addition, observer ratings were strong predictors of behaviors. When the criterion was academic achievement or job performance, other-ratings yielded predictive validities substantially greater than and incremental to self-ratings. These findings indicate that extraordinary value can gained by using other-reports to measure personality, and these findings provide guidelines toward enriching personality theory. Various subfields of psychology in which personality variables are systematically assessed and utilized in research and practice can benefit tremendously from use of others' ratings to measure personality variables.

These kinds of technically impressive articles are typical of the I-O 'prestige journal' literature … Tables 10 and 11 present some 'standout' findings … but are they trustworthy? For example, how can we move from a correlation of **0.23** to **0.55** when predicting job performance from other's ratings of an individual's Conscientiousness behaviours (Table 11)? We have more than doubled the size of the relationship we observe using the raw data, to produce 'true validities'.

The most obvious and sensible correction for attenuation of correlation coefficients is restriction of range. That can seriously attenuate any relationship present within a representative random sample from some population.

For example, generating 5,000 cases of data from a bivariate normal distribution (means = 0, SDs of 1) where the full-range correlation is 0.5, and computing a correlation on a subset selected on the predictor who score 1.0 and above, shows substantive attenuation:

	Total Dataset	Above X-cut of 1.0000
Actual No. of Cases	5000	757
Expected No. of Cases	-	773
Actual % of Total Cases	100%	15.14%
Expected % of Cases	-	15.47%
Actual Taylor-Russell SR	-	-
Expected Taylor-Russell SR	-	-
Actual propn. successes	-	-
Expected propn. successes	-	-
Mean of Var X	-0.013831	1.534214
SD of Var X	0.997239	0.436100
Mean of Var Y	-0.016523	0.773201
SD of Var Y	0.995309	0.895943
Correlation	0.499281	0.218495

Table 10
Meta-Analysis of Other-Ratings and Self-Ratings Validities for Predicting Academic Achievement

	Zero-order meta-analytic results									1 other + self (Hough)		1 other + self (Poropat)	
Trait and rating type	k	N	$\bar{r}$	SD_{obs}	SD_{resid}	ρ_{ov}	$SD_{\rho ov}$	$CI_{\rho ov}$	ρ_{xy}	SD_{ρ}	R_{ov}	β	R_{ov}	β
Emotional Stability											**.30**		**.29**	
Other-ratings	6	2,940	.25	.12	.10	**.27**	.11	[.24, .31]	**.46**	.19		.22		.31
Self-ratings—Hough	162	70,588	.20			**.22**		[.21, .23]	**.25**			.14		
Self-ratings—Poropat	104	54,462	.00			**.00**		[−.01, .01]	**.00**					−.10
Extraversion											**.36**		**.39**	
Other-ratings	7	3,081	.32	.29	.29	**.35**	.31	[.31, .38]	**.52**	.47		.38		.43
Self-ratings—Hough	128	63,057	.07			**.08**		[.07, .09]	**.09**			−.08		
Self-ratings—Poropat	103	54,072	−.02			**−.02**		[−.03, −.01]	**−.02**					−.20
Openness											**.20**		**.18**	
Other-ratings	4	1,278	.17	.14	.13	**.18**	.14	[.12, .24]	**.29**	.22		.15		.17
Self-ratings—Hough	8	3,628	.13			**.14**		[.11, .18]	**.17**			.09		
Self-ratings—Poropat	102	54,380	.07			**.08**		[.07, .09]	**.09**					.02
Agreeableness											**.01**		**.05**	
Other-ratings	6	1,460	.01	.08	.05	**.01**	.05	[−.05, .07]	**.02**	.09		.01		.00
Self-ratings—Hough	15	7,330	.01			**.01**		[−.01, .04]	**.01**			.01		
Self-ratings—Poropat	99	53,432	.04			**.05**		[.04, .06]	**.06**					.05
Conscientiousness											**.42**		**.41**	
Other-ratings	9	3,609	.37	.14	.13	**.41**	.14	[.38, .44]	**.69**	.24		.37		.40
Self-ratings—Hough	42	18,661	.23			**.25**		[.24, .27]	**.31**			.11		
Self-ratings—Poropat	127	64,867	.17			**.18**		[.17, .19]	**.22**					.03

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Meta-analytic correlations for self-ratings drawn from Hough (1992) are designated as "Self-ratings—Hough," and meta-analytic correlations for self-ratings drawn from Poropat (2009) are designated as "Self-ratings—Poropat." Corrected correlations (ρ_{ov} and ρ_{xy}) and multiple correlations (R_{ov}) are presented in boldface for emphasis. k = number of independent samples; N = total sample size; $\bar{r}$ = mean observed correlation; SD_{obs} = observed standard deviation; SD_{resid} = standard deviation of correlations after accounting for variability from sampling error and unreliability; ρ_{ov} = operational validity, corrected for unreliability in the criterion only; $SD_{\rho ov}$ = standard deviation of operational validities, corrected for variability due to sampling error and criterion unreliability; $CI_{\rho ov}$ = 95% confidence interval around ρ_{ov} estimates; ρ_{xy} = true score validity, correcting for unreliability in the predictor and criterion; SD_{ρ} = standard deviation of true validities, corrected for variability due to sampling error and predictor and criterion unreliability; R_{ov} = operational multiple correlation from combining self- and one other-rating; β = standardized beta-weight in the multiple regression for other- or self-rating of the trait.

Table 11
Meta-Analysis of Other-Ratings and Self-Ratings Validities for Predicting Job Performance

	Zero-order meta-analytic results										Combined: Self + 1 other	
Trait and rating type	k	N	$\bar{r}$	SD_{obs}	SD_{resid}	ρ_{ov}	$SD_{\rho ov}$	$CI_{\rho ov}$	ρ_{xy}	SD_{ρ}	R_{ov}	β
Emotional Stability											**.19**	
Other-ratings	7	1,190	.14	.06	.00	**.17**	.00	[.10, .25]	**.37**	.00		.16
Self-ratings, Barrick et al. (2001)	224	38,817	.06			**.11**		[.09, .12]	**.12**	.08		.09
Extraversion											**.14**	
Other-ratings	6	1,135	.08	.10	.07	**.11**	.09	[.03, .18]	**.18**	.15		.09
Self-ratings, Barrick et al. (2001)	222	39,432	.06			**.11**		[.09, .12]	**.12**	.12		.09
Openness											**.22**	
Other-ratings	6	1,135	.18	.08	.00	**.22**	.00	[.15, .30]	**.45**	.00		.22
Self-ratings, Barrick et al. (2001)	143	23,225	.03			**.04**		[.02, .06]	**.05**	.11		.00
Agreeableness											**.18**	
Other-ratings	7	1,190	.13	.07	.00	**.17**	.00	[.09, .24]	**.31**	.00		.15
Self-ratings, Barrick et al. (2001)	206	36,210	.06			**.11**		[.09, .13]	**.13**	.09		.07
Conscientiousness											**.31**	
Other-ratings	7	1,190	.23	.07	.00	**.29**	.00	[.22, .36]	**.55**	.00		.25
Self-ratings, Barrick et al. (2001)	239	48,100	.12			**.20**		[.19, .22]	**.23**	.10		.11

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Meta-analytic correlations for self-ratings drawn from Barrick et al. (2001) are designated as "Self-ratings—Barrick et al." Corrected correlations (ρ_{ov} and ρ_{xy}) and multiple correlations (R_{ov}) are presented in boldface for emphasis. k = number of independent samples; N = total sample size; $\bar{r}$ = mean observed correlation; SD_{obs} = observed standard deviation; SD_{resid} = standard deviation of correlations after accounting for variability from sampling error and unreliability; ρ_{ov} = operational validity, corrected for unreliability in the criterion only; $SD_{\rho ov}$ = standard deviation of operational validities, corrected for variability due to sampling error and criterion unreliability; $CI_{\rho ov}$ = 95% confidence interval around ρ_{ov} estimates; ρ_{xy} = true score validity, correcting for unreliability in the predictor and criterion; SD_{ρ} = standard deviation of true validities, corrected for variability due to sampling error and predictor and criterion unreliability; R_{ov} = operational multiple correlation from combining self- and one other-rating; β = standardized beta-weight in the multiple regression for other- or self-rating of the trait.

But, in this article, for the data presented in Tables 10 and 11, no correction for restriction of range was made. Instead, corrections were made using criterion reliability estimates, then both criterion and predictor variable reliability estimates.

The 'of interest' predictor variables in each case were others' ratings of an individual. The criteria were academic achievement and Job Performance. Tables 3 and 4 present some of these estimated reliabilities:

Table 3
Meta-Analysis of Single-Rater Interrater Reliabilities, by Information Source

Trait and source	k	N	$\bar{r}_{rr}$	SD_{obs}	SD_{resid}	ρ_{rr}	SD_{ρ}	$Conf_L$	$Conf_U$	FS k
Emotional Stability	72	13,458	.33	.14	.13	**.40**	.15	.37	.41	403
Family	5	774	.37	.16	.14	**.44**	.17	.37	.51	32
Friends	16	3,102	.38	.11	.08	**.45**	.10	.42	.49	106
Cohabitators	4	1,021	.20	.07	.04	**.24**	.04	.17	.31	12
Work colleagues	5	682	.28	.12	.08	**.34**	.10	.25	.41	23
Incidental acquaintances	5	338	.18	.07	.00	**.22**	.00	.09	.33	13
Strangers	41	3,723	.23	.15	.12	**.27**	.14	.24	.31	148
Extraversion	82	12,438	.43	.13	.11	**.51**	.13	.49	.52	623
Family	5	774	.45	.08	.04	**.53**	.05	.46	.59	40
Friends	16	3,111	.46	.08	.05	**.55**	.06	.51	.57	131
Cohabitators	7	1,101	.28	.08	.03	**.34**	.03	.26	.39	32
Work colleagues	6	1,238	.37	.12	.10	**.44**	.12	.38	.49	38
Incidental acquaintances	7	466	.40	.13	.07	**.48**	.08	.38	.56	49
Strangers	49	4,238	.40	.17	.14	**.48**	.16	.44	.50	343
Openness	53	7,990	.32	.13	.11	**.39**	.14	.38	.42	286
Family	2	185	.38	.07	.00	**.47**	.00	.31	.62	13
Friends	9	2,077	.43	.05	.00	**.53**	.00	.49	.58	68
Cohabitators	3	939	.21	.03	.00	**.26**	.00	.19	.34	10
Work colleagues	5	928	.29	.11	.08	**.36**	.10	.29	.43	24
Incidental acquaintances	5	338	.20	.09	.00	**.25**	.00	.12	.38	15
Strangers	31	3,601	.30	.17	.15	**.37**	.18	.34	.41	155
Agreeableness	83	10,689	.32	.14	.12	**.40**	.15	.37	.42	448
Family	5	774	.25	.18	.16	**.31**	.20	.23	.39	20
Friends	20	3,263	.34	.11	.08	**.43**	.09	.38	.46	116
Cohabitators	8	1,172	.33	.06	.00	**.41**	.00	.34	.47	45
Work colleagues	6	1,238	.29	.07	.02	**.37**	.03	.29	.42	29
Incidental acquaintances	5	338	.24	.07	.00	**.30**	.00	.17	.42	19
Strangers	48	4,094	.27	.16	.13	**.33**	.15	.30	.37	211
Conscientiousness	64	11,523	.36	.13	.11	**.44**	.14	.42	.46	397
Family	5	774	.35	.17	.15	**.43**	.19	.35	.51	30
Friends	20	3,394	.37	.08	.04	**.46**	.04	.42	.49	128
Cohabitators	8	1,071	.26	.06	.00	**.32**	.00	.25	.39	34

Table 4
Information Type Moderators Meta-Analysis of Strangers' Single-Rater Interrater Reliabilities

Trait and information source	k	N	$\bar{r}_{rr}$	SD_{obs}	SD_{resid}	ρ_{rr}	SD_{ρ}	$Conf_L$	$Conf_U$	FS k
Emotional stability	*41*	*3,723*	*.23*	*.15*	*.12*	**.27**	*14*	.24	.31	148
Visual cues only	18	1,202	.15	.11	.00	**.18**	.00	.11	.24	36
Still visual	8	371	.25	.28	.24	**.29**	.29	.18	.41	32
Silent nonverbal	11	926	.13	.09	.00	**.16**	.00	.08	.23	18
Audio cues only	9	315	.32	.14	.00	**.38**	.00	.26	.50	49
Activity (audio + visual)	17	2,336	.30	.15	.13	**.36**	.15	.31	.40	85
Prescribed behavior	3	267	.22	.06	.00	**.26**	.00	.12	.39	10
Natural behavior	15	2,136	.32	.16	.14	**.38**	.16	.34	.43	81
Personal object	5	411	.15	.05	.00	**.18**	.00	.06	.29	10
Text/electronic communication	4	243	.09	.09	.00	**.11**	.00	−.04	.25	3
All Extraversion	*49*	*4,238*	*.40*	*.17*	*.14*	*.48*	*.16*	.44	.50	343
Visual cues only	20	1,331	.30	.12	.03	**.35**	.04	.29	.41	100
Still visual	9	393	.30	.11	.00	**.35**	.00	.24	.46	45
Silent nonverbal	14	1,187	.30	.12	.07	**.36**	.08	.29	.41	70
Audio cues only	10	393	.45	.25	.21	**.53**	.24	.43	.62	80
Activity (audio + visual)	19	2,388	.48	.11	.09	**.57**	.10	.53	.60	163
Prescribed behavior	3	267	.45	.06	.00	**.53**	.00	.41	.64	24
Natural behavior	16	2,124	.50	.10	.07	**.59**	.09	.55	.62	144
Personal object	5	411	.30	.07	.00	**.35**	.00	.25	.45	25
Text/electronic communication	4	243	.23	.10	.00	**.28**	.00	.13	.41	14
All Openness	*31*	*3,601*	*.30*	*.17*	*.15*	*.37*	*.18*	.34	.41	155
Visual cues only	15	1,270	.19	.16	.13	**.23**	.16	.17	.30	42
Still visual	3	249	.23	.03	.00	**.29**	.00	.14	.43	11

Note that the observed rater reliabilities are themselves corrected for unreliability of each rater (using a test-retest reliability coefficient).

So, correcting a correlation of **0.23** to **0.55** when predicting job performance from other's ratings of an individual's Conscientiousness behaviours (Table 11) looks something like:

$r12 := 0.23$ raw correlation, predicting job performance from others' ratings
$r11 := 0.58$ reliability of criterion (job performance)
$r22 := 0.30$ reliability of predictor (others' ratings) - Conscientiousness

$$\rho_{ov} := \frac{r12}{\sqrt{r11}} = 0.302$$ operational validity

$$\rho_{xy} := \frac{r12}{\sqrt{r11 \cdot r22}} = 0.55$$ true validity

These are 'back of the matchbox estimates' as I'm assuming the authors were correcting individual study estimates prior to averaging. They also describe:
"As before, samples varied in the number of other-raters used to measure personality. To facilitate comparisons with self-ratings, we adjusted all correlations between multi-other composite ratings of traits and criteria to the level of a single other-rater using procedures identical to those in Study 2. That is, validities were individually disattenuated for interrater unreliability of ***r*** raters and then reattenuated for interrater unreliability of a single rater. Similarly, when samples contributed several correlations to a single analysis as a result of using multiple measures of the predictor or criterion, these correlations were composited where possible and otherwise averaged. " p. 1112.

My point of concern is twofold:

❶ there is nothing 'operational' about operational validity. Given there is no range restriction, what any employer or interested party can expect to observe, on average, is the **0.23** relationship. It's no good correcting for unreliability in job performance rating when all you can ever observe are unreliable job ratings (assessed in the manner which produces these unreliable ratings). Change the manner in which you assess job performance and its 'start again' viz a viz establishing the validity of the predictor/s. What you see is all you will get in everyday real-world deployment of these predictors i.e. the raw, uncorrected correlations.

❷ As to 'true validity', what is 'true' about such an estimate? What exactly does it mean to say raters rate so discrepantly from one another that they can only agree with 0.3 or even 0.4 reliability? What this figure tells us is that raters cannot agree between one another with any surety. Simply assuming their average represents some 'true' reliable estimate of an attribute is ridiculous. You no longer have any idea what it is they agree upon; it's just an arithmetic average of their disparate ratings. True scores only exist as hypothetical entities within a statistical theory of test scores. Their 'truth' is confined to a platonic world in which there is no observation or measurement error. They are fairy-tale estimates for a fairy-tale world.

Bottom line: Raters either agree that they see the same *(good enough similar)* event/magnitude of attribute, behaviour or they don't. If they don't you stop right there because it makes no sense to continue not knowing which rater is rating more accurately. **When ratings are this poor, you have to do the due-diligence hard-yards required to understand the degree of rating disparity, and what's causing it.** Not just as the title of the old British comedy film has it: "Carry on Regardless"!

Presenting such estimates as evidence supporting strong claims of effects which will always be unobservable in the real world is yet another instance, in my mind, of what makes some of the I-O literature 'untrustworthy'.