

The meta-analytic correlation between two Big Five factors

Something is not quite right in the woodshed

Paul Barrett & Jean-Pierre Rolland*

* Université de Paris Quest -
Nanterre La Défense, France

Aunt Ada Doom is the infamous “mad woman in the attic” of Stella Gibbons’ comedy novel *Cold Comfort Farm* (1932); her mind became unhinged when as a child she saw “something nasty in the woodshed”. The literary phrase may not totally capture the effect of the observations we make below, but something is “not quite right” about the following meta-analytic results reported in a series of studies since 1993.

We do not wish to dwell on the pros and cons of meta-analysis, but rather we find ourselves questioning the implicit understanding that meta-analysis is always capable of revealing the expected population correlation between attributes. The paper by LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997) *Discrepancies between meta-analyses and subsequent large randomized, controlled trials. The New England Journal of Medicine*, 337, 8, 536-542 is perhaps the most famous study showing that meta-analysis does not always produce accurate estimates of population parameters, and the recent study by Schonemann, P.H., & Scargle, J.D. (2008) *A Generalized Publication Bias Model. Chinese Journal of Psychology*, 50, 1, 21-29, helps to explain why.

Of specific interest here though are the various meta-analytic estimates of population correlations between two specific Big Five personality test scales, Emotional Stability and Conscientiousness. These are the two most important broad personality factors associated meta-analytically with job performance.

An aside [Feb, 2010]: a recent paper from De Raad, B., Barelds, D.P.H., Levert, E., Ostendorf, F., Mlacic, B., Di Blas, L., Hrebicková, M., Szirmák, Z., Szarota, P., Perugini, M., Church, A.T., & Katigbak, M.S. (2010) *Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. Journal of Personality and Social Psychology*, 98, 1, 160-173. The three factors they find replicable are:

Extraversion

(+) active, chatty, cheerful, dynamic, energetic, enthusiastic, extraverted, exuberant, lively, open, outgoing, sociable, talkative, vigorous, vivacious
(-) bashful, closed, introverted, lonely, passive, quiet, reserved, shy, silent, solitary, taciturn, timid, unsociable, untalkative, withdrawn

Agreeableness

(+) accommodating, agreeable, benevolent, conciliatory, friendly, gentle, good-natured, kind-hearted, lenient, meek, mild, patient, peaceful, sympathetic, tolerant
(-) aggressive, bossy, brusque, choleric, cold-hearted, despotic, domineering, fierce, hot-tempered, intolerant, irritable, overbearing, quarrelsome, short-tempered, stubborn

Conscientiousness

(+) careful, conscientious, diligent, disciplined, dutiful, hardworking, industrious, methodical, meticulous, orderly, organized, precise, scrupulous, thorough, tidy
(-) absent-minded, careless, chaotic, disorderly, disorganized, frivolous, imprudent, inaccurate, irresponsible, lax, lazy, negligent, rash, undisciplined, untidy

Interestingly they find Emotional Stability/Anxiety is not replicable across cultures.

The Evidence

1 The relevant correlation appears (albeit not referenced in the paper) in Table 6, p. 669, of: Ones, D.S., Viswesvaran, C., & Reiss, A.D. (1996) Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81, 6, 660-679.

Table 6
Estimated Population Correlation Matrix Between the Big Five Dimensions of Personality, Social Desirability, and Job Performance

Personality dimension	1	2	3	4	5	6	7
1. Social desirability	—	.37	.06	.00	.14	.20	.01
2. Emotional stability	.37	—	.19	.16	.25	.26	.07
3. Extraversion	.06	.19	—	.17	.17	.00	.10
4. Openness to experience	.00	.16	.17	—	.11	-.06	-.03
5. Agreeableness	.14	.25	.17	.11	—	.27	.06
6. Conscientiousness	.20	.26	.00	-.06	.27	—	.23
7. Job performance	.01	.07	.10	-.03	.06	.23	—

Note. Presentation format provides a full matrix in order to facilitate computation of partial correlations by readers. The correlations presented are estimated population correlations based on meta-analyses. Correlations between the Big Five and job performance are from Barrick and Mount (1991); correlations among the Big Five are from Ones (1993); correlations between the Big Five and social desirability are from Table 2 in this study; correlations between social desirability and job performance are from Table 5 in this study.

Ones (1993) is not referenced in the paper, but the citation may refer to Ones, D.S. (1993). *The construct validity of integrity tests*. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA. What's important here is that these are estimated "population correlations".

- 2 Mount, M.K., Barrick, M.R., Scullen, S.M., & Rounds, J. (2005) Higher-order dimensions of the Big Five personality traits and the Big Six vocational interest types. *Personnel Psychology*, 58, 2, 447-478. Table 3, p. 463 of this paper now shows a “true” correlation of 0.52 instead of 0.26.

TABLE 3
True Score Correlations Among RIASEC and FFM Variables

Variable	1	2	3	4	5	6	7	8	9	10	11
1 Realistic	–										
2 Investigative	.45	–									
3 Artistic	.25	.36	–								
4 Social	.18	.26	.39	–							
5 Enterprising	.20	.09	.28	.51	–						
6 Conventional	.27	.17	.01	.29	.53	–					
7 Conscientiousness	.05	.09	–.06	.07	.08	.19	–				
8 Agreeableness	.01	.01	.03	.17	–.06	–.01	.39	–			
9 Openness	.06	.25	.41	.13	.05	–.10	.09	.17	–		
10 Extraversion	.03	.02	.09	.29	.40	.06	.17	.26	.45	–	
11 Emotional Stability	.08	.12	–.02	.04	.08	.03	.52	.42	.19	.24	–

Note. True score intercorrelations among the Big Five personality traits and among the Big Six interests are based on the meta-analytic results presented in Tables 1 and 2. Decimal points omitted.

Looking at Table 2 on page 462, we see ..


TABLE 2
Meta-Analytic Results for Big Five Intercorrelations

Correlate	ρ	SD_{ρ}	90% CV	% Var.
Stability–Extraversion	0.24	0.02	(.22, .27)	72.0
Stability–Openness	0.19	0.12	(–.03, .34)	7.9
Stability–Conscientiousness	0.52	0.19	(.27, .73)	2.9
Stability–Agreeableness	0.42	0.13	(.26, .58)	7.8
Extraversion–Openness	0.45	0.00	(.45, .45)	368.1
Extraversion–Conscientiousness	0.17	0.11	(.04, .31)	10.5
Extraversion–Agreeableness	0.26	0.15	(.07, .44)	6.14
Openness–Conscientiousness	0.09	0.12	(–.06, .23)	8.1
Openness–Agreeableness	0.17	0.11	(.04, .31)	11.4
Conscientiousness–Agreeableness	0.39	0.14	(.21, .56)	7.6

Note. Number of samples in the analysis = 4; Total number of respondents across the samples = 4,000; ρ = estimated true score correlation (corrected for sampling error and unreliability); SD_{ρ} = estimated true standard deviation for the correlation; 90% CV = estimated 90% credibility value for true score correlation; % Var. = percent variance in correlations accounted for by statistical artifacts.

The 0.26 correlation reported in Ones et al (1993) lies just outside the 90% credibility value reported in Mount et al (2005). It would lie within a 95% interval, so a rare result perhaps but not impossible. But, the range of that credibility interval in Mount, et al. is very wide.

It should be noted that this is the estimated true-score correlation between these two attributes used in Table 4, p. 845 in the recent paper by Schmidt, F.L., Shaffer, J.A., & Oh, I-S. (2008) Increased accuracy for range restriction corrections: implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61, 4, 827-868, and not the one reported by Ones (1993).

 Steel, P., Schmidt, J., & Shultz, J. (2008) Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, 134, 1, 138-161.

From Table 6, p. 148 ...

Table 6
Correlations Among NEO, EPQ, and EPI Dimensions

Construct	NEO traits				
	N	E	O	A	C
NEO					
Neuroticism	.83	(57)	(33)	(34)	(37)
Extraversion	-.33	.77	(28)	(28)	(31)
Openness	-.09	.24	.73	(28)	(28)
Agreeableness	-.23	.19	.10	.71	(27)
Conscientiousness	-.33	.28	.01	.18	.79

This correlation is computed over 37 studies, utilizing 17,464 cases (an average of 472 per study as noted on page 14, para 2, column 1).

Given the Big Five attribute “Emotional Stability” is now called “Neuroticism”, we think it is likely the scoring was reversed (high scores equate to higher Emotional Instability), which would explain the reversal in sign of the correlation between Emotional Stability (Neuroticism) and Conscientiousness attributes.

So far we have correlations of 0.26, 0.52, and now 0.33 ... all “population/true-value estimates”.

4

Meriac, J.P., Hoffman, B.J., Woehr, D.J., & Fleisher, M.S. (2008) Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, 93, 5, 1042-1052.

From Table 3, p. 1047, we see

Table 3
Corrected Meta-Analytic Intercorrelations Among Study Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Cognitive ability	(.90)	.32	.20	.31	.23	.24	.32	.31	.25	-.07
2. Job performance	.34	(.90)	.22	.25	.15	.29	.33	.31	.16	-.10
3. Consideration/awareness of others	.22	.24	(.80)	.37	.39	.43	.30	.36	.52	-.08
4. Communication	.35	.27	.44	(.86)	.41	.45	.42	.40	.42	-.09
5. Drive	.26	.16	.47	.48	(.86)	.58	.44	.41	.46	-.05
6. Influencing others	.28	.31	.51	.52	.67	(.87)	.52	.50	.54	.01
7. Organizing and planning	.36	.35	.36	.49	.49	.60	(.87)	.59	.38	-.07
8. Problem solving	.34	.32	.42	.45	.48	.56	.66	(.91)	.39	-.07
9. Stress tolerance	.28	.18	.63	.50	.56	.63	.44	.45	(.85)	-.08
10. Neuroticism	-.08	-.12	-.10	-.11	-.06	.02	-.09	-.09	-.10	(.79)
11. Extraversion	.08	.29	.10	.16	.29	.21	.13	.11	.17	-.02
12. Openness to Experience	.13	-.02	.09	.17	.08	.11	.12	.14	.15	.00
13. Agreeableness	.15	.12	.07	.13	.12	.11	.03	.09	.09	-.10
14. Conscientiousness	.24	.29	.14	.12	.14	.13	.07	.17	.17	-.24

Note. Values on the diagonal in parentheses represent the artifact distribution for the variables that were used in the meta-analysis. Values on the diagonal are operational validity coefficients. Values below the diagonal are fully corrected coefficients.

The correlation between Neuroticism and Conscientiousness is computed using 1009 cases included in 6 independent studies (from Table 1, p. 1046).

So now we have correlations of 0.24, 0.26, 0.33, and 0.52; all “population/true-value estimates”.

Notice how far the correlation between Extraversion and Neuroticism has dropped in this study from the value shown on the previous page, -0.33 down to -0.02.

5 Hogan, J., Barrett, P., & Hogan, R. (2007) Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92, 5, 1270-1285.

Data are from 5,266 individuals who were administered the Hogan Personality Inventory on two occasions (T1 and T2), whose HPI scale scores were converted into their Big Five equivalents using a formula developed by Smith, B. & Ellingson, J.E. (2002) Substance vs style: a new look at Social Desirability in motivating contexts. *Journal of Applied Psychology*, 87, 2, 211-219. From Table 2, p. 1276

Table 2

Intercorrelations Between Hogan Personality Inventory—Revised (HPI-R) Scales at Time 1 (T1) and Time 2 (T2)

HPI-R scale	1	2	3	4	5	6
1. Emotional Stability T1	.75					
2. Extraversion T1	.06	.77				
3. Openness T1	.31	.24	.75			
4. Agreeableness T1	.17	.33	.16	.53		
5. Conscientiousness T1	.48	.06	.31	.26	.68	
6. Emotional Stability T2	.61	.06	.23	.12	.32	.78
7. Extraversion T2	-.01, <i>ns</i>	.59	.14	.19	-.01, <i>ns</i>	.09
8. Openness T2	.20	.21	.68	.10	.20	.37
9. Agreeableness T2	.08	.24	.10	.46	.12	.21
10. Conscientiousness T2	.34	.07	.22	.16	.58	.52

Note. $n = 5,266$. Coefficient alpha reliabilities for each scale at T1 and T2 are presented in italics on the main diagonal versus T2 same-scale score correlations. Correlations $> |.04|$ are significant at $p < .01$.

i *Note, this is not meta-analytic data, just a very large sample of 5,266 cases.*

If we correct each correlation for the unreliability of each scale, the T1 correlation becomes 0.67, that for T2 becomes 0.70.

Table 1: Summary of results from the various evidence-bases **1** to **5**

Study	Sample Size	Estimated Population Correlation
Ones (1993)	?	0.26
Mount et al (2005)	4,000	0.52
Steel et al (2008)	17,464	0.33
Meriac et al (2008)	1,009	0.24
Hogan et al (2007) – T1	5,266	0.67
Hogan et al (2007) – T2	5,266 (same cases as T1)	0.70

It's possible that the estimated population correlations all lie within a 95% credibility interval extending from 0.20 to 0.80. But is this of any practical value except to indicate that some true correlation is positive, and lies somewhere between 'near zero' and 'very large'?

6 Ultimately, we think we have the definitive answer using the 2007 Hogan Normative Sample dataset of 156, 614 cases. Scoring the data into the Big Five scales using the Smith and Ellingson formula, we can compute what are likely to be the nearest *realistic* population estimates of the attribute correlations for a US population. Further, if we wish to assume each individual possesses a “true score” on an attribute, we can correct these correlations for the unreliability of each scale to create the true score correlations which can be compared directly to those presented in Table 1.

In order to correct each raw-score correlation for scale unreliability, we need the composite alphas for each of the Big Five scales (composite because each Big Five scale is constructed from four HPI HICs). Table 2 presents these correlational results:

Table 2: Raw Score and “True-Score” disattenuated correlations between Emotional Stability and Conscientiousness – using the Big Five rescored HPI US Normative data for the Hogan Personality Inventory

Variable	Correlations (Big 5 Factor scores from Normative HPI - N=156614.sta) Marked correlations are significant at $p < .05000$ N=156614 (Casewise deletion of missing data)				
	Emotional Stability	Extraversion	Openness	Agreeableness	Conscientiousness
Emotional Stability	0.75	0.16	0.28	0.25	0.37
Extraversion	0.20	0.83	0.27	0.34	0.06
Openness	0.36	0.33	0.80	0.22	0.28
Agreeableness	0.38	0.49	0.32	0.58	0.32
Conscientiousness	0.53	0.08	0.39	0.52	0.66

The raw-score correlations are above the diagonal, the *true score* disattenuated values are given in the lower half of the matrix (the shaded cells), the alpha reliability for each scale is given in the main diagonal.

Adding the relevant value to Table 1:

Table 3: Summary of results from the various evidence-bases **1 to **6****

Study	Sample Size	Estimated Population Correlation
Ones (1993)	?	0.26
Mount et al (2005)	4,000	0.52
Steel et al (2008)	17, 464	0.33
Meriac et al (2008)	1,009	0.24
Hogan et al (2007) – T1	5,266	0.67
Hogan et al (2007) – T2	5,266	0.70
Hogan HPI Normative US dataset	156,614	0.53

Of the four meta-analyses (rows 1-4), only one (Mount 2005) came close, using a quarter of the sample size of the one that yielded a population estimate of 0.33.

The Evidence: Two More Studies



Judge, T.A., van Vianen, A. E.M., & De Pater, I.E. (2004) Emotional stability, core self-evaluations, and job outcomes: A review of the evidence and an agenda for future research. *Human Performance*, 17, 3, 325-346.

From Table 2, p. 330) we see ...

TABLE 2
Relationship of Core Traits to Five-Factor Model of Personality

	<i>Neuroticism</i>	<i>Extraversion</i>	<i>Openness</i>	<i>Agreeableness</i>	<i>Conscientiousness</i>
Neuroticism	—	-.30	-.02	-.29	-.49
Self-esteem	-.66	.42	.23	.20	.46
Locus of control	-.51	.36	.03	.16	.47
Generalized self-efficacy	-.59	.54	.25	.20	.46

Note. Correlations are meta-analytic population correlations (corrected for measurement error).

These data were computed from the summary (uncorrected) data provided in Judge, T.A., Erez, A., Bono, J.E., & Thoresen, C.J. (2002) Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology*, 83, 3, 693-710, Table 8, p. 704 ...

Table 8
Weighted Average of Relationship Among Measures of the Four Traits and Other Variables Across Studies

Variable	<i>N</i>	Locus of control	Emotional stability	Self-esteem	Generalized self-efficacy
Agreeableness	1,608	.19	.31	.22	.23
Conscientiousness	1,885	.31	.28	.39	.43
Extraversion	1,885	.26	.26	.36	.39
Openness	1,608	.24	.18	.14	.33
Job satisfaction	717	.22	.39	.38	.39
Life satisfaction	1,517	.17	.25	.35	.22
Happiness	826	.52	.52	.51	.68
Stress	1,393	.17	.25	.27	.21
Strain	1,393	.15	.35	.26	.19

Note. *N* = sample size combined across all studies.



Revelle, W., Wilt, J., & Rosenthal, A. (in press) Individual differences in cognition: New methods for examining the personality-cognition link. To appear in Aleksandra Gruszka, Gerald Matthews, and Blazej Szymura (editors): *Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control*. Draft, April 2008, downloaded from: <http://personalityresearch.net/revelle/publications/rwr.08.pdf>

From page 22, Table 12 we see:

Table 12: Correlations between the Big 5 measures, demographics, and ability measures for the Big 5 are shown in the appropriate diagonal.

	Extra	Stability	Cons	Agree	Open
Gender	0.07	-0.20	0.13	0.25	-0.10
Education	0.00	0.05	0.18	0.10	0.16
Age	-0.01	0.09	0.20	0.10	0.13
SAT	-0.11	0.02	-0.08	-0.14	0.25
SATV	-0.07	0.02	-0.08	-0.05	0.33
SATQ	-0.05	0.09	-0.02	-0.08	0.23
ACT	-0.05	0.04	-0.01	-0.06	0.30
Combined	-0.08	0.10	0.00	0.00	0.28
Reasoning	-0.08	0.09	-0.02	-0.03	0.30
Spatial	-0.07	0.09	0.01	0.00	0.20
Verbal	0.04	0.02	0.08	0.12	0.10
Extraversion	0.93	0.28	0.14	0.41	0.30
Stability	0.28	0.93	0.17	0.17	0.17
Conscientiousness	0.14	0.17	0.92	0.25	0.13
Agreeableness	0.41	0.17	0.25	0.90	0.21
Openness	0.30	0.17	0.13	0.21	0.83

The sample size for this correlation was > 50,000 – using the innovative “Synthetic Aperture Personality Assessment” sampling procedure (see:

<http://www.personality-project.org/revelle/publications/sapa.mpa.key.pdf> for more details about this study and the sampling technique; the table data above are similar to the data reported in slide #74).

The Final Picture

Incorporating these studies with those from  to , we can summarize the eight studies' results in Table 4.

Table 4: Summary of results from the various evidence-bases  to 

Study	Sample Size	Estimated Population Correlation
Ones (1993)	?	0.26
Mount et al (2005)	4,000	0.52
Steel et al (2008)	17, 464	0.33
Meriac et al (2008)	1,009	0.24
Hogan et al (2007) – T1	5,266	0.67
Hogan et al (2007) – T2	5,266	0.70
Hogan HPI Normative US dataset	156,614	0.53
Judge et al (2004)	1,885	0.49
Revelle et al	> 50,000	0.17

So, our estimates for the relationship between Emotional Stability and Conscientiousness range from 0.17 through to 0.70. Clearly, this range of “population” estimates indicates that something is not right here.

In Psychology, attributes with the same name are not necessarily the same at all

The clue to what might be causal for this widely varying range of estimates is given in two papers:



Rich, G.A., Bommer, W.H., MacKenzie, S.B., Podsakoff, P.M., & Johnson, J.L. (1999) Apples and apples or apples and oranges? A Meta-Analysis of objective and subjective measures of salesperson performance. *Journal of Personal Selling & Sales Management*, XIX, 4, 41-52.

Abstract

"The goal of this study was to examine the relationship between objective and subjective measures of salesperson performance. The results of a meta-analysis of 21 studies with a total sample size of 4,092 participants indicated an overall mean corrected correlation of **.447**, indicating that the two measures shared only about 20% of variance. Although a moderator subgroup analysis found that the corrected mean correlation was somewhat higher in certain situations, **the findings generally suggest that objective and subjective measures of salesperson performance are not interchangeable**, and that the choice of the most appropriate measure may require a tradeoff between accurately tapping the domain of the performance construct and minimizing measurement error".

This is a clear indication that meta-analyses which utilize studies composed of a "same-name" criterion (here "Sales Performance") are likely to yield quite different "population estimates") if the criterion is not exactly the same.

From pp. 41-42 of this paper:

"In the sales management literature, performance has been measured in a variety of different ways. For example, roughly half of the studies (53.3% of the reported correlations) included in Churchill, Ford, Hartley, and Walker's (1985) meta-analysis of the determinants of salesperson performance measured performance using subjective evaluations obtained from managers, peers, or self-reports. The other half (46.7% of the reported correlations) measured volume, sales commissions, or percent of quota. However, in most instances, no attempt was made to explain why one type of measure was used as opposed to another. The implicit assumption appears to be that objective and subjective measures of performance are interchangeable, and that the domain of sales performance is adequately captured by either subjective ratings or objective results. Thus, practitioners and researchers alike are assuming that these two classes of measures are "apples and apples" and that they may be treated interchangeably. "

The same argument might easily be made for that global criterion, "job performance".



Woods, S.A. (2009). The comparative validities of six personality inventories. *Proceedings of the Division of Occupational Psychology, British Psychological Society, Annual Conference 2009* – reported in detail in Woods, S.A. (2009). *The Structures and Validities of Five Work-related Personality Inventories. Unpublished Working Paper (Jan 30th, 2009)*. The working paper may be requested from S.A. Woods, Aston Business School <http://www.abs.aston.ac.uk/newweb/staff/detail.asp/sfIdStaffID=A0000731>).

Abstract

“This study examined the structures, convergent validities, and criterion validities of five work-related personality inventories (the Hogan Personality Inventory . the Occupational Personality Questionnaire. the Sixteen Personality Factor Questionnaire. the Personality and Preferences Inventory, Profile Match). A sample of 371 individuals from the UK working population completed various combinations of the five inventories, plus a measure of the lexical Big Five and several criterion scales. Overall, the results indicated sensible and interpretable factor structures for the inventories (with the exception of the Personality and Preferences Inventory), theoretically meaningful convergent validities, and similar criterion validity magnitudes and patterns. It was concluded that these data give confidence in the use of these inventories for research and practice, and reveal important similarities and consistencies in their validities. ”

A sub-selection of Table 23 on p. 30 presents a summary of the mean correlations for each particular Big Five scale. The table shows the mean correlation between the 5 different tests’ measures of each of the Big Five attributes. So, given 5 personality tests, each measuring Extraversion, Woods correlates each test’s scores on Extraversion with those from the other tests (pairwise); producing $((5 \times 5) - 5) / 2$ possible correlations. The average is then taken of these correlations.

Table 23. Mean convergence

Domain	All Scales
Extraversion	.38
Agreeableness	.29
Conscientiousness	.31
Emotional Stability	.42
Openness	.29

What these data indicate is that there is only a broad level of agreement between five major work-related personality tests when their scales are re-expressed (where appropriate) as Big Five personality constructs.

Whereas the evidence-bases supporting each questionnaire might well be substantive and excellent, any meta-analysis which incorporated these tests as part of a study looking at the correlation between personality constructs, or even the predictive accuracy of a specified outcome, is going to be influenced by the lack of agreement between test scales.

An aside

Maraun, M.D. (1998) Measurement as a Normative Practice: Implications of Wittgenstein's Philosophy for Measurement in Psychology. *Theory & Psychology*, 8, 4, 435-461 ...

"B. The difficulty faced by psychologists in measuring is not mathematical or empirical in nature, but is instead that the concepts they wish to have enter into their measurement operations are typically of the common-or-garden variety. These concepts have notoriously complicated grammars. In light of (5), this generates serious difficulties.

C. The relative lack of success of measurement in the social sciences as compared to the physical sciences is attributable to their sharply different conceptual foundations. In particular, the physical sciences rest on a bedrock of technical concepts, while psychology rests on a web of common-or-garden psychological concepts. The fact of (B) completes the explanation. "

and on pp 452-453 and onwards ..

"Constraints on Measurement in Psychology

If Wittgenstein's views are correct, then difficulties in psychological measurement may be explained by the fact that psychology has mischaracterized measurement. However, what if one did consider the grammars of psychological concepts. What would grammar say about how to measure psychological concepts? I believe that a grammatical investigation of psychological concepts (e.g. as in Ter Hark, 1990) reveals (a) the obvious fact that, as it stands, common-or-garden psychological concepts are not measurable, and (b) the existence of grammatical constraints on the possibility of measurement involving common-or-garden psychological concepts. These constraints may, in part, explain the enduring difficulty faced by psychology in attempting to measure, and, in particular, as compared to the physical sciences (for related discussions of the latter point, see Campbell, 1921; Schonemann, 1994).

While (a) and (b) may have the tone of overstatement, I believe that it is nevertheless interesting to briefly consider each. To do so requires that two types of concept be distinguished: (1) technical concepts and (2) common-or-garden concepts. A *technical concept* is a concept defined by a specialized or expert community, and employed within a narrow, technical field of application. A *common-or-garden concept*, on the other hand, is a concept with a common employment in everyday life (Baker & Hacker, 1982). Common or- garden concepts are taught, learned and understood by the person on the street, and have meanings that are manifest in broad, normative linguistic practices. They, in addition, are more apt to be the source of confusion in the context of scientific investigation.... "

An Obvious Conclusion?

Meta-Analytic population estimates will be unreliable (unstable) unless identical attributes and predictors/outcomes are used in every contributing study. This makes sense when you think about what meta-analysis was really designed for – correcting sampling errors due to small, individual study attempts at examining construct relations and effect sizes.

As you depart from this rule, independent-study meta-analytic estimates of supposedly the same phenomena will show degrees of variation in rough proportion to the variation in meaning amongst "same name attributes" which comprise constituent studies.

Something is definitely not quite right in the woodshed.

Appendix 1:

The abstract to the paper by LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997) *Discrepancies between meta-analyses and subsequent large randomized, controlled trials. The New England Journal of Medicine*, 337, 8, 536-542.

Background

Meta-analyses are now widely used to provide evidence to support clinical strategies. However, large randomized, controlled trials are considered the gold standard in evaluating the efficacy of clinical interventions.

Methods

We compared the results of large randomized, controlled trials (involving 1000 patients or more) that were published in four journals (the New England Journal of Medicine, the Lancet, the Annals of Internal Medicine, and the Journal of the American Medical Association) with the results of meta-analyses published earlier on the same topics. Regarding the principal and secondary outcomes, we judged whether the findings of the randomized trials agreed with those of the corresponding meta-analyses, and we determined whether the study results were positive (indicating that treatment improved the outcome) or negative (indicating that the outcome with treatment was the same or worse than without it) at the conventional level of statistical significance ($P < 0.05$).

Results

We identified 12 large randomized, controlled trials and 19 meta-analyses addressing the same questions. For a total of 40 primary and secondary outcomes, agreement between the meta-analyses and the large clinical trials was only fair ($\kappa = 0.35$; 95 percent confidence interval, 0.06 to 0.64). The positive predictive value of the meta-analyses was 68 percent, and the negative predictive value 67 percent. However, the difference in point estimates between the randomized trials and the meta-analyses was statistically significant for only 5 of the 40 comparisons (12 percent). Furthermore, in each case of disagreement a statistically significant effect of treatment was found by one method, whereas no statistically significant effect was found by the other.

Conclusions

The outcomes of the 12 large randomized, controlled trials that we studied were not predicted accurately 35 percent of the time by the meta-analyses published previously on the same topics.